A Domain Adaptive IoT Intrusion Detection Algorithm Based on AEC-GAT Feature Extraction and Joint Domain Adversary

Qian Wang, Menghui Fan,Zhijuan Wu, Hongnian Yu, Yongqiang Cheng, Bing Zhang

***Abstract*— The high heterogeneity of Internet of Things (IoT) devices causes severe imbalance in network traffic data, and the cost of collecting and labeling sufficient intrusion samples is high or impossible, resulting in data scarcity in IoT security. Therefore, this paper proposes a domain adaptive IoT intrusion detection algorithm based on AEC-GAT feature extraction and joint domain adversary, which leverages abundant data resources from traditional network intrusion detection to improve the detection accuracy in IoT environments. Firstly, a feature extraction method combining a causal embedding autoencoder and a graph attention network (AEC-GAT) is designed. The AEC uses causal inference to uncover deep semantic links between domains, while GAT captures device interaction patterns to enhance semantic relevance and structure awareness in the features. Secondly, to address the pronounced class imbalance in IoT datasets, Focal Loss is introduced to replace the traditional cross entropy loss. This formulation dynamically adjusts the sample weight through the scaling factor to guide the algorithm to focus on the minority samples that are difficult to classify. Meanwhile, a class adaptive independent domain discriminator method is proposed, which incorporates a class-level alignment mechanism within a joint adversarial training method. This method dynamically adjusts both the training intensity and the loss weight of each class specific domain discriminator. The experimental results show that the algorithm in this paper significantly improves the detection performance of IoT intrusion detection by migrating traditional network intrusion detection domain knowledge, and has superior performance in various indicators compared to existing algorithms.**

***Index Terms*— IoT intrusion detection, adversarial domain adaptation, class imbalance, graph attention network.**

I. Introduction

WITH the accelerating advancement of IoT technologies, the number of intelligent devices has increased exponentially. These devices have been extensively deployed across critical sectors such as the Internet of Vehicles, industrial control systems, and Intelligent healthcare [1]. The high connectivity and openness of IoT not only significantly enhance system efficiency and intelligence but also heighten its susceptibility to sophisticated security threats. IoT has become the key target of hacker attacks due to the characteristics of heterogeneous devices, diverse communication protocols, and complex deployment environment [2]. Intrusion Detection Systems (IDS) constitute a fundamental defense mechanism in IoT security, enabling the real-time identification of anomalous traffic and malicious intrusions.

In recent years, with the rapid development of deep learning technology, it has shown strong feature learning and recognition capabilities in IDS. Detection algorithms based on Convolutional Neural Networks (CNN) [3], Recurrent Neural Networks (RNN) [4], and Graph Neural Networks (GNN) [5] have been developed to extract rich semantic features from original traffic data. However, these algorithms highly rely on a large number of complete annotation data, and their acquisition costs are high and the annotation process is tedious. It is particularly challenging for IoT intrusion detection due to the heterogeneity and wide distribution of IoT devices. It is also difficult to build a unified attack sample collection mechanism, resulting in a serious scarce of attack samples.

To address these challenges, Domain Adaptation (DA) has emerged as a promising solution for intrusion detection by enabling knowledge transfer and enhancing the robustness of learning in the target domain. Its core idea is to use the existing models and data in the source domain to improve the generalization ability of the model in the target domain by reducing the distribution difference between the source domain and the target domain in the feature space. Leveraging the richness of traditional network intrusion detection data and their alignment with IoT attack patterns, this study defines them as the source domain, with IoT intrusion data constituting the target domain. By mapping the two to the shared feature subspace, the traditional network intrusion knowledge can be

Manuscript received 14 October 2025; accepted 5 November 2025. This work is supported by the National Natural Science Foundation of China under Grant Nos. 62376240 and 62302428, the S&T Program of Hebei under Grant No. 236Z0304G, the Innovation Capability Improvement Plan Project of Hebei Province under Grant No. 22567626H. (Corresponding author: Qian Wang.)

Qian Wang, Menghui Fan, Zhijuan Wu, and Bing Zhang are with the School of Artificial Intelligence (School of Software) and the Hebei Key Laboratory of Computer Virtual Technology and System Integration, Yanshan University, Qinhuangdao, Hebei, China, 066004 (e-mail: wangqian@ysu.edu.cn; fanmenghui@stumail.ysu.edu.cn; wzj0911@stumail.ysu.edu.cn; bingzhang@ysu.edu.cn).

Hongnian Yu is with the School of Computing Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, EH10 5DT, UK (e-mail: H.Yu@napier.ac.uk).

Yongqiang Cheng is with the School of Computer Science, University of Sunderland, Sunderland, SR1 3SD, UK (e-mail: yongqiang.cheng@sunderland.ac.uk).

effectively migrated to the IoT domain, thus improving the performance of IoT intrusion detection.

In DA algorithms, in order to measure and reduce the distribution difference between the source domain and the target domain, the measurement algorithms commonly used in the research include the Maximum Mean Difference (MMD) [6] and Wasserstein distance [7]. In recent years, with the widespread application of Generative Adversarial Networks (GAN) in cross-domain tasks, domain adaptation algorithms based on adversarial learning have shown significant advantages. These algorithms learn domain-invariant features through adversarial training, aligning feature distributions by minimizing domain-specific discrepancies. The algorithm proposed by Layeghy et al. [8] employs adversarial training to extract domain-invariant features, effectively enhancing model robustness across diverse network environments. Compared to specific network structures, adversarial learning provides a more flexible feature alignment mechanism for domain adaptation, which can be implemented through different algorithms.

Although adversarial learning demonstrates considerable cross-domain transferability, there are still many challenges in existing DA algorithms. Firstly, existing algorithms predominantly rely on statistical distribution alignment while disregarding the deep semantic relationships among attack samples. Secondly, most algorithms adopt the standard cross-entropy loss function. In highly imbalanced class scenarios, this loss tends to underemphasize minority classes resulting in substantially degraded detection performance. Finally, prevailing DA algorithms such as Domain Adversarial Neural Network (DANN) [9] typically employ global distribution alignment, which neglects class-wise discrepancies between source and target domains. Consequently, such confusion often results in feature confusion across different classes (as illustrated in Fig.1).

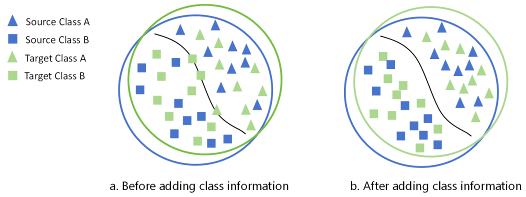

▲ Source Class A
■ Source Class B
▲ Target Class A
■ Target Class B
a. Before adding class information
b. After adding class information

Fig. 1: Classification comparison considering class distribution

To address the shortcomings of these algorithms, a domain adaptive IoT intrusion detection algorithm based on AEC-GAT feature extraction and joint domain adversary is proposed to enable efficient cross-domain intrusion detection.

The main contributions of this paper are as follows:

- An AEC-GAT feature extraction method is proposed to enhance domain invariance. It combines an AEC for causal modeling of semantic relationships with a GAT for attention-based interaction modeling, improving feature robustness and domain generalize ability.
- Focal Loss is introduced to address the class imbalance problem. Focal Loss dynamically adjusts class-specific weights to mitigate majority class dominance during training, prioritizing difficult classified minority samples and thereby enhancing detection performance under class imbalance.
- A class adaptive independent domain discriminator method is proposed to enable class level alignment. Separate domain discriminators are constructed for each class to address the class confusion. Additionally, the intensity of adversarial training is adaptively modulated according to the distributional similarity among classes. This method improves class-level alignment and mitigates negative transfer.

The remainder of this paper is organized as follows. Section 2 introduces the relevant work. Section 3 explains the algorithm proposed in this article. Section 4 is validated through experiments. Finally, Section 5 summarizes this study.

II. Related Work

DA has improved IoT intrusion detection by leveraging knowledge from traditional networks. However, its performance is still hampered by the feature heterogeneity, class imbalance, and coarse alignment common in IoT environments. To address the shortcomings of existing algorithms, this paper conducts research on feature extraction method, class imbalance processing and domain alignment method.

A. Feature extraction method

Feature extraction is central to the performance of IDS, directly affecting the ability of the model to identify and classify network behaviors. Early research methods mainly rely on artificially designed features, such as traffic statistics, protocol field information and behavior pattern extraction [10]. With the development of deep learning, data-driven feature extraction methods gradually replace the traditional methods. CNN and RNN have been widely employed to model the spatial and temporal dependencies of network traffic. For example, literature [11] has constructed a CNN-LSTM fusion algorithm, which can simultaneously capture local structure and sequence dynamics. In recent years, GNN has been introduced into intrusion detection tasks because of its advantages in processing non Euclidean structured data. The feature extraction algorithm proposed in [12] employs Graph Convolutional Networks (GCN) to learn domain-invariant representations, thereby enhancing the cross-scenario adaptability of the algorithm. However, the existing GNN based on intrusion detection algorithms still has some generalization ability problems in cross-domain scenarios, and fails to fully consider the impact of correlation between features on feature extraction.

B. Class imbalance processing

IoT intrusion detection often encounters a serious class imbalance problem. The number of normal traffic samples far exceeds that of abnormal traffic, which limits the model in identifying minority classes. In order to alleviate this problem, traditional methods often use resampling techniques such as undersampling and oversampling [13]. SMOTE [14], as a classic oversampling method, expands the scale of minority classes by generating new samples, and is widely used for network traffic classification. However, it may distort the original data distribution or introduce redundant samples, thereby increasing the risk of overfitting [15]. In contrast, the optimization method based on loss function is more flexible. Saito et al. [16] improved the performance of minority class recognition by alternately optimizing conditional entropy, which enables the algorithm to use a small number of labeled samples in the target domain. The Weighted Cross Entropy loss (WCE) [17] assigns different loss weights to each class to enhance the focus on minority classes. However, the use of static weights limits its ability to adapt to dynamic changes in data distribution. Focal Loss [18] was initially used for target detection and has been introduced into the intrusion detection field in recent years.

C. Domain alignment method

Domain alignment is a common strategy in DA techniques, aiming to reduce distribution discrepancies between source and target domains to improve model generalization. Ganin et al. [19] proposed DANN, which learns domain-invariant features through adversarial training with a gradient inversion layer. However, its performance degrades in intrusion detection due to class imbalance and sparse target labels. To address these challenges, Yao et al. [20] introduced Soft Transmission Network (STN) with soft labels to enhance conditional distribution alignment, while Wang et al. [21] proposed an unsupervised whole graph embedding method called WCGN to achieve graph-level feature alignment. Li et al. [22] proposed Cross-Domain Adaptive Clustering (CDAC), a semi-supervised method that integrates inter- and intra-cluster alignment. In addition, many methods have focused on leveraging sample and structural relationships to achieve finer-grained alignment. Saito et al. [23] proposed a Minimax Entropy method called APE that enhances intra-domain consistency and selective alignment. Similarly, Yao et al. [24] introduced Discriminant Distribution Alignment (DDA), which aligns distributions while enlarging inter-class separability through adaptive classifiers. In contrast, Wu et al. [25] proposed Geometric Graph Alignment (GGA), a graph-based method that captures semantic associations between intrusion domains for domain-invariant representation learning. However, existing methods often overlook the distribution differences between classes, which can lead to confusion of features across different classes, limiting the overall effectiveness and generalization ability of cross-domain detection.

III. The Proposed Algorithm

Firstly, an AEC-GAT feature extraction method is devised to enhance semantic integrity and structural awareness. Secondly, Focal Loss is applied to mitigate class imbalance and heighten sensitivity to underrepresented attack samples. Finally, a class adaptive independent domain discriminator method is proposed to achieve fine-grained, class-level alignment through joint adversarial training. The complete algorithm is illustrated in Fig.2.

A. Feature extraction method based on AEC-GAT

Combining AEC and GAT to construct an adaptive feature extraction method, which provides more robust shared feature representation for subsequent domain alignment. This method combines causal mechanism and graph attention to effectively distinguish normal and abnormal behaviors.

1) Causal embedding autoencoder (AEC): AEC is a structural extension that introduces causal modeling mechanism on the basis of traditional autocoder, aiming to improve the causal consistency and portability of features learned in cross-domain tasks. All class features were transformed using one-hot encoding, and the feature set was standardized through min–max normalization to ensure numerical consistency across dimensions. Features were then ranked according to the Pearson correlation coefficients with the labels, and those with higher correlations were retained for causal graph construction, reducing redundancy and computational complexity. The AEC encoder subsequently maps the high-dimensional features into a lower-dimensional latent space through multiple nonlinear fully connected layers, producing compact representations that preserve critical causal structures while suppressing noise. Subsequently, the constraint based PC algorithm [26] is used to build cause and effect graphs G_S and G_T. The algorithm determines causal directions by leveraging V-structures and the acyclicity constraints inherent in causal graphs. Assuming the input is a feature set $\mathcal{V} = \{v_1, v_2, \ldots, v_d\}$, where each $v_i \in \mathcal{V}$ denotes a feature variable, the procedure consists of two stages: firstly, a fully connected undirected graph is constructed. For any pair of adjacent nodes (v_i, v_j) in the graph, the Fisher Z test is employed as a conditional independence test to assess their dependency. If independence is detected, the corresponding edge is removed, resulting in an undirected graph $G^{'}$. Secondly, the direction of edges is determined based on the d-separation principle, expanding the skeleton into a completed partially directed acyclic graph. The constructed causal graphs G_S and G_T guide the AEC in extracting structurally coherent and transferable feature representations within their respective domains.

Reconstruction loss aims to minimize the error between the reconstructed output and the original input. This objective encourages the encoder to learn compact and shared low-dimensional representations:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{x \sim \mathcal{D}_s \cup \mathcal{D}_t} \left[\|x - \hat{x}\|^2 \right] \tag{1}$$

where x is the input data, $\hat{x}$ is the reconstruction output of the method, and $\| \ . \ \|^2$ denotes the squared L_2 norm. Simply relying on reconstruction loss fails to ensure that the representation aligns with the causal structure of data generation. Therefore, a causal consistency loss $\mathcal{L}_{\text{causal}}$ is introduced:

$$\mathcal{L}_{\text{causal}} = \mathbb{E}_{v_i \in \mathcal{V}} [\|F(v_i) - \Phi(v_i | Pa(v_i))\|^2] \tag{2}$$

where $F(v_i)$ is the prediction or feature representation of the model for the sample, Φ represents the condition generation function based on the causality diagram, and $\text{Pa}(v_i)$ is the parent node set of v_i in the causal graph.

The encoder of the AEC employs a Multi-Layer Perceptron (MLP). ReLU activations enhance its nonlinear modelling

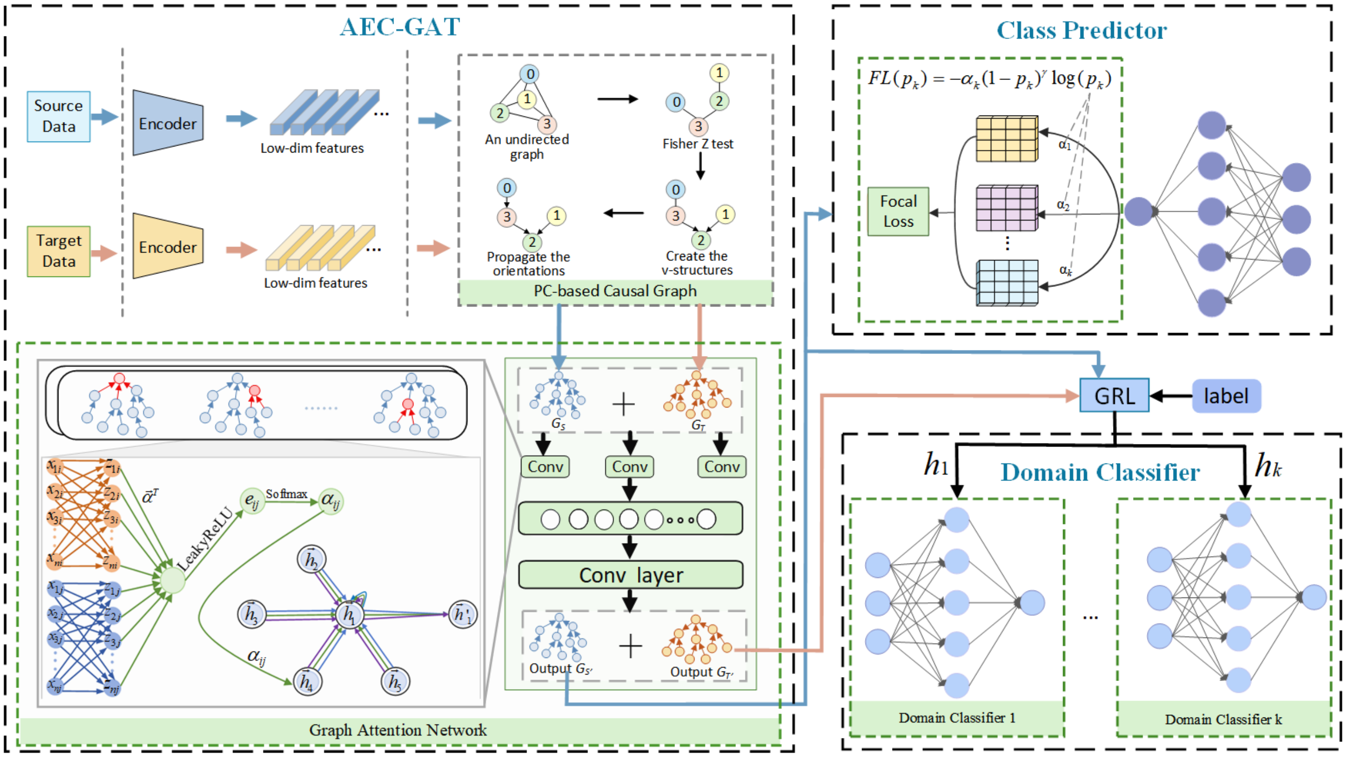


Fig. 2: Overall process of the proposed algorithm

TABLE I: Symbol definition table

Notation	Interpretation
x_s, x_t	Input feature vectors of sample from source and target domains
$\hat{x}$	Reconstructed input vector by the autoencoder
$\mathcal{D}_s, \mathcal{D}_t$	Source domain and target domain datasets
V	Node set in the causal graph (corresponding to the feature set)
$v_i \in V$	The i-th feature node
$\mathbf{h}_i \in \mathbb{R}^m$	Feature vector of node v_i, with dimension m
$\mathbf{h}'_i$	Updated feature representation of node v_i
$Pa(v_i)$	Parent node set of node v_i
$\Phi(v_i \vert Pa(v_i))$	Conditional generation function based on the causal graph
G_S, G_T	Causal graphs of the source and target domains
$G = (V, E)$	Graph structure with node set V and edge set E
$\mathbf{W} \in \mathbb{R}^{m' \times m}$	Weight matrix of the GAT
$\mathbf{a} \in \mathbb{R}^{2m'}$	Learnable attention vector in the attention mechanism
$\sigma(\cdot)$	Activation function (ReLU)
e_{ij}	Unnormalized attention score between node v_i and its neighbor v_j
α_{ij}	Normalized attention weight
K	Total number of classes
$f(x_*)$	Output representation of the feature extractor
y	Sample label (true class)
y_k	One-hot label of class
$P(y_k \vert f(x))$	Classifier-predicted probability that the sample belongs to class
p_k	Model-predicted probability for the true class k
α_k	Weight coefficient of class k (computed from sample distribution)
γ	Focusing parameter of Focal Loss
$D_k(\cdot)$	Independent domain discriminator for class k
$\mathcal{D}_{\mathrm{s}}^k$	Set of samples of class k in the source domain
$\mathcal{D}_t^k$	Set of samples of class k in the target domain
$d_{\mathcal{H},u}(*,*)$	The similarity between the feature distributions
h_k	Dynamic weight of the domain discriminator

capacity, while dimensionality reduction yields compact latent representations. This design preserves key causal dependencies, suppresses noise, and thereby improves the generalization performance of the proposed algorithm.

2) Graph attention network (GAT): The causal graph directs the model to capture true causal relationships between variables. Based on this causal structure, this paper further introduces GAT to model and aggregate node relationships. GCN relies on predefined adjacency matrices for feature propagation, limiting their flexibility in capturing semantic relationships between nodes. In contrast, GAT introduces a learnable attention mechanism that adaptively assigns aggregation weights to neighboring nodes. GAT effectively reduces reliance on fixed graph structures and enhances the capacity of the method to represent complex graph topologies. Simultaneously, its multihead attention mechanism enhances the diversity and robustness of feature representations.

The input data of GAT is graph $G = (V, E)$. The initial feature of each node v_i is $\mathbf{h}_i \in \mathbb{R}^m$, where m represents the dimension of the feature. In order to obtain sufficient expression ability, a shared linear transformation parameterized by the weight matrix $\mathbf{W} \in \mathbb{R}^{m' \times m}$ is applied to each node.

The weight between neighbor nodes is calculated through the attention mechanism. For node v_i and its neighbor $\forall v_j \in \mathcal{N}(i)$, the attention score is calculated as follows:

$$e_{ij} = \mathrm{LeakyReLU}(\mathbf{a}^T \cdot [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j]) \tag{3}$$

where $\mathbf{a} \in \mathbb{R}^{2m'}$ is a trainable weight vector, $\parallel$ is a feature connection operation, and e_{ij} is an unnormalized attention score. Next, use the function softmax to normalize e_{ij} to obtain the attention weight α_{ij} representing the relative importance of neighbor v_i to node v_j:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{4}$$

where α_{ij} represents the relative importance of neighbor v_i to node v_j. Here, attention weight α_{ij} is used to weight and sum the features of neighboring nodes, updating the feature representation of node v_i. Specifically, the new feature representation $\mathbf{h}_i'$ of node v_i is obtained by weighted summation of the features $\mathbf{h}_i$ of all neighboring nodes $v_j \in \mathcal{N}(i)$.

$$\mathbf{h}_i' = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}\mathbf{h}_j\right) \tag{5}$$

where σ is the activation function using ReLU. The multi-head mechanism employs independent attention heads to compute node features, followed by fusion via concatenation or averaging. If there are N attention heads, the update feature of node v_i is shown as:

$$\mathbf{h}_i' = \|_{n=1}^{N}\sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{n}\mathbf{W}^{n}\mathbf{h}_j\right) \tag{6}$$

AEC-GAT combines causal reasoning and graph attention to extract domain-invariant features $f(x_*)$. The adaptive attention mechanism dynamically allocates node weights, effectively capturing the importance of neighborhoods and improving the performance of cross-domain learning.

B. Processing class imbalance through Focal Loss

To alleviate the performance degradation caused by the difference of feature distribution between different domains, DANN is widely applied in unsupervised domain adaptation tasks. Its network architecture consists of three components: feature extractor, label classifier and domain discriminator. Concurrently, DANN embeds a Gradient Reversal Layer (GRL) between the feature extractor and domain discriminator. This layer directly transmits features without any transformation during forward propagation, while in backward propagation, GRL multiplies the gradient returned by the domain discriminator by a negative coefficient, thereby reversing the direction of gradient updates. Through this adversarial mechanism, the feature extractor constantly learns to generate domain invariant features that can confuse the domain discriminator, while the domain discriminator continuously optimizes its domain classification capability. Both of them collectively enhance the domain invariance of the learned features through this adversarial game mechanism.

The training objective of DANN is formulated as a minimax problem, aiming to minimize classification loss $\mathcal{L}_C$ while maximizing domain discriminator loss $\mathcal{L}_D$ across domains.

$$\mathcal{L}_C = -\mathbb{E}_{(x_s,y)\sim\mathcal{D}_s}\sum_{k=1}^{K} y_k \cdot \log P(y_k \mid f(x_s)) \tag{7}$$

$$\mathcal{L}_D = -\mathbb{E}_{x_s\sim\mathcal{D}_s}[\log D(f(x_s))] - \mathbb{E}_{x_t\sim\mathcal{D}_t}\left[\log\left(1 - D(f(x_t))\right)\right] \tag{8}$$

where K is the total number of classes, y_k is the one pot code of the real tag, $P(y_k|f(x_s))$is the prediction probability of classifier C to feature $f(x_s)$, $D(.)$ is the probability that the output sample of the domain discriminator belongs to the source domain.

The label classifier relies on the standard cross-entropy loss as the optimization objective, as shown in Formula (9), without adequately addressing the performance degradation caused by class imbalance. This guides the model toward learning an optimal decision boundary for classification.

$$CrossEntropy(p,y) = \begin{cases} -\log(p), & \text{if } y = 1 \\ -\log(1-p), & \text{otherwise} \end{cases} \tag{9}$$

In practical applications such as IoT intrusion detection, normal traffic samples overwhelmingly outnumber malicious ones. As a result, the training data exhibits severe class imbalance. Under such conditions, cross-entropy loss is dominated by majority class samples during training, causing the model to underemphasize minority classes. Meanwhile, traditional cross-entropy assigns uniform weights across samples, ignoring differences in classification difficulty. This prevents the model from focusing on critical instances near the decision boundary, further hindering its ability to learn from minority classes.

Therefore, this paper introduces the Focal Loss function into the label classifier. The prediction probability p_k of this method is defined as follows.

$$p_k = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{10}$$

where p_k denotes the predicted probability of the true class. The Focal Loss extends the standard cross-entropy by introducing a focusing parameter that adaptively adjusts the contribution of each sample. By reducing the weight of easily classified examples and increasing emphasis on difficult ones during training, the algorithm is effectively guided to focus on more challenging instances. The formulation is as follows.

$$\mathrm{FL}(p_k) = -(1-p_k)^{\gamma}\,\log(p_k) \tag{11}$$

$(1-p_k)^{\lambda}$ can be regarded as a modulation factor that suppresses the contribution of easily classified samples, thereby increasing the relative weight of the difficult to classify samples in the loss function. Specifically, when p_k approaches 1, the modulation factor $(1-p_k)^{\gamma}$ approaches 0, substantially reducing the loss contribution of easily classified samples. Conversely, when the predicted probability p_k approaches 0, the modulation factor $(1-p_k)^{\gamma}$ approaches 1, thereby increasing the relative weight of difficult classified samples in the loss. Through this mechanism, the loss weights are dynamically adjusted, reducing the overemphasis on easily classified samples while placing enhanced emphasis on difficult classified samples. For different values of γ, the loss effect is shown in Fig.3.

As p_k increases—indicating a higher confidence in correct classification—the associated loss correspondingly decreases. To further alleviate the problem of class imbalance in IoT intrusion detection data, this method adds class weight α_k in Focal Loss:

$$\mathrm{FL}(p_k) = -\,\alpha_k\,(1-p_k)^{\gamma}\,\log(p_k) \tag{12}$$

where α_k is the relative weight of different classes. Smaller values of α_k are assigned to majority-class samples, while larger values are given to minority-class samples, thereby increasing the contribution of minority in the overall loss. This

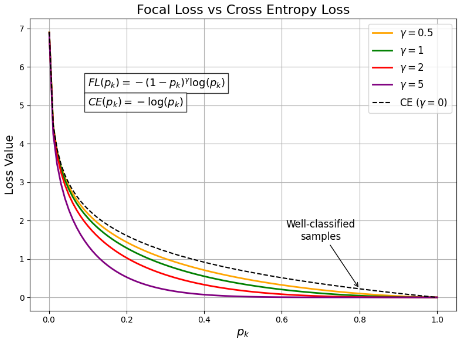


Fig. 3: Influence of the focusing parameter γ on Focal Loss

modification enables the model to maintain overall classification performance while improving its ability to detect critical attack classes in IoT intrusion detection tasks.

The modulation factor α_k for each class k is determined by computing the weighted proportion of class α_k among all source domain samples. This ensures that the weight for each class can be adaptively adjusted according to its prevalence in the training data. The calculation formula is as follows.

$$\alpha_k = \frac{\sum_{(x_j,y_j)\in\mathcal{D}_s} y_k}{\sum_{(x_j,y_j)\in\mathcal{D}_s} x_j \cdot y_k} \tag{13}$$

According to the above inference, in order to enhance the adaptability of the label classifier to the class imbalance data in DANN, Focal Loss is introduced into the loss function of the label classifier. The improved label classifier loss function formula is as follows.

$$\begin{aligned}\mathcal{L}_C^{\text{Focal}} = -\,\mathbb{E}_{(x_s,y)\sim\mathcal{D}_s}\Bigg[\sum_{k=1}^{K}\alpha_k\big(1-P(y_k\mid f(x_s))\big)^{\gamma} \\ \times\, y_k \log P(y_k\mid f(x_s))\Bigg]\end{aligned} \tag{14}$$

where $P(y_k \mid f(x_s))$ is the prediction probability of the label classifier for class k, $\big(1 - P(y_k \mid f(x_s))\big)^\gamma$ is the moderating factor, which gives greater weight to difficult classified samples.

Focal Loss is adopted to address class imbalance in cross-domain scenarios by dynamically adjusting sample weights, enhancing focus on difficult and minority samples. This preserves domain-invariant feature learning while improving classification robustness for imbalanced data.

C. Class adaptive independent domain discriminator

Traditional IoT cross-domain transfer learning aligns global feature distributions but ignores class-level discrepancies, causing class confusion near decision boundaries. To address this issue, the paper proposes a class-adaptive independent domain discriminator (CA-DD) that enhances cross-domain generalization.

Specifically, an independent domain discriminator D_k is designed for each class k. After introducing independent domain discriminators, the DANN training process begins by extracting features from input samples using the feature extractor. These features are initially classified by the label predictor to generate predicted class labels. Then, the samples will be assigned to the corresponding class discriminator according to the predicted class. The domain discriminator D_k then attempts to distinguish whether the class k sample originates from the source domain or the target domain. For each domain discriminator D_k, its loss function is as follows.

$$\mathcal{L}_{D_k} = -\mathbb{E}_{x_s\sim\mathcal{D}_s^k}\big[\log D_k(f(x_s))\big] - \mathbb{E}_{x_t\sim\mathcal{D}_t^k}\big[\log\big(1-D_k(f(x_t))\big)\big] \tag{15}$$

In order to further improve the accuracy of domain alignment, this paper introduces a class adaptability evaluation mechanism. Specifically, the similarity between the source and target domain feature distributions is calculated for each class. This information is used to dynamically adjust the weight of each class-specific domain discriminator. This method addresses the differences in domain alignment effectiveness across classes through fine-grained adjustment of the adversarial training process.

This paper first calculates the similarity $d_{\mathcal{H},\mathcal{U}}\left(\mathcal{D}_s^k, \mathcal{D}_t^k\right)$ between the feature distributions of the source and target domains for each class, using a measurement based on $\mathcal{H}-\text{divergence}$ [27]. For each class k, the distribution discrepancy between the source and target domains is computed and measured using $\mathcal{H}-\text{divergence}.$

$$d_{\mathcal{H},\mathcal{U}}\big(\mathcal{D}_s^k,\mathcal{D}_t^k\big) = \sup_{h\in\mathcal{H}}\Big|\,\mathbb{E}_{x\sim\mathcal{D}_s^k}[h(x)] - \mathbb{E}_{x\sim\mathcal{D}_t^k}[h(x)]\Big| \tag{16}$$

where $\mathcal{D}_s^k$ and $\mathcal{D}_t^k$ respectively denote the sets of source and target domain samples for class k. Dynamic weights for each class-specific domain discriminator are adjusted inversely proportional to the corresponding class values. The dynamic weight h_k of the domain discriminator for class k can be adjusted inversely proportional to the corresponding $\mathcal{H}-\text{divergence}$ value.

$$h_k = \frac{1}{1 + d_{\mathcal{H},\mathcal{U}}(\mathcal{D}_s^k, \mathcal{D}_t^k)} \tag{17}$$

A higher $\mathcal{H}-\text{divergence}$ value indicates a larger distribution gap between the source and target domains, suggesting poorer alignment quality for that class. As a result, a smaller adaptive weight is assigned to its domain discriminator, reducing its training intensity. This helps prevent overtraining on classes that are inherently difficult to align ,and reduces model overfitting on those classes. In contrast, a smaller $\mathcal{H}-\text{divergence}$ indicates greater similarity between the source and target domain distributions, reflecting better alignment quality. Consequently, a higher adaptive weight is assigned to the class, resulting in increased training intensity. This method can increase the training intensity of those classes that have been aligned well, thereby further optimizing their performance in the target domain.

During the optimization process, an adaptive weight h_k is incorporated into the loss function of each class-specific domain discriminator.

$$\mathcal{L}_D^{\mathcal{H}} = \sum_{k=1}^{K} h_k\,\mathcal{L}_{D_k} \tag{18}$$

By jointly optimizing causal loss, label classification loss and domain discrimination loss, the algorithm effectively improves IoT intrusion detection performance. The overall loss function of the algorithm is as follows.

$$\lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{causal}} + \min_{C} \max_{D} \left(\lambda_3 \mathcal{L}_{C}^{\text{Focal}} - \lambda_4 \mathcal{L}_{D}^{\mathcal{H}} \right) \quad (19)$$

IV. Experimental results and analysis

In order to comprehensively evaluate the effectiveness and generalization capability of the proposed algorithm in this paper, multiple experiments are designed to validate its performance across different cross-domain intrusion detection tasks.

A. Dataset introduction and experimental setup

This study selects representative cross-domain intrusion detection datasets to evaluate the proposed algorithm. The dataset are as follows: the source domain is NSL-KDD [28], UNSW-NB15 [29] and CIC-IDS2017 [30]; the target domain is UNSW-BOTIOT [31] and UNSW-TONIOT [32]. The dataset will be described in detail below.

NSL-KDD (recorded as K) includes multiple types of network attacks and has a relatively balanced sample distribution, and it is suitable for intrusion detection in traditional network environments; UNSW-NB15 (recorded as N) is a collection based on the real network environment in the UNSW series, combined with simulated attacks. It contains diverse network traffic characteristics and is widely used to evaluate the performance of intrusion detection systems; CIC-IDS2017 (recorded as C) is a baseline dataset in the field of network intrusion detection. It is based on real network traffic capture, covering multiple network protocols, and can be used for research related to network security in IDS.

UNSW-BOTIOT (recorded as B) simulates the real network behavior of IoT devices, with a small sample size and obvious distribution deviation. It can better reflect the security challenges in the IoT environment; UNSW-TONIOT (recorded as T) is specifically designed for IoT scenarios. It can simulate the network behaviors of IoT devices in real-world operations, and has more complex and representative feature distributions.

These dataset not only reflect the differences between traditional computer networks and IoT environments, but also cover the challenges of significant distribution deviation and sample imbalance.

In terms of experimental implementation, this paper implements the proposed model structure based on PyTorch framework. The classifier and domain discriminator are both constructed as MLP with two hidden layers. After each layer, LeakyReLU activation function and Dropout layer (rejection rate is 0.3) are connected to enhance the nonlinear expression ability of the algorithm and inhibit overfitting. At the same time, batch normalization is introduced to accelerate convergence and improve training stability. This structural design has significantly enhanced the generalization ability and discrimination performance of the algorithm in IoT cross-domain intrusion detection tasks.

B. Comparison experiment of overall performance evaluation

In the overall performance evaluation, this paper uses a variety of evaluation indicators to comprehensively measure the performance of the algorithm, including accuracy, precision, recall, F1-score and AUC.

To fully verify the performance advantages of the proposed algorithm in the cross-domain IoT intrusion detection task. This paper selects eight current mainstream transfer learning algorithms as the comparison baseline, including APE [23], CDAC [22], STAR [16], DDAS [24], STN [20], DDAC [24], WCGN [21] and GGA [25]. The experiment is conducted on three representative cross-domain tasks: K → B, N → T and C → B. The experimental results are shown in TABLE II - IV.

TABLE II: Cross-domain detection of K → B

	Accuracy	**Precision**	**Recall**	**F1**	**AUC**
APE	0.557	0.395	0.516	0.366	0.476
CDAC	0.553	0.251	0.501	0.335	0.435
STAR	0.481	0.250	0.500	0.333	0.542
DDAS	0.674	0.390	0.522	0.376	0.471
STN	0.768	0.525	0.534	0.405	0.484
DDAC	0.769	0.393	0.540	0.407	0.500
WCGN	0.743	0.608	0.617	0.583	0.604
GGA	0.884	0.778	0.773	0.757	0.798
Ours	**0.900**	**0.887**	**0.899**	**0.892**	**0.901**

TABLE III: Cross-domain detection of N → T

	Accuracy	**Precision**	**Recall**	**F1**	**AUC**
APE	0.500	0.495	0.703	0.581	0.580
CDAC	0.500	0.493	0.702	0.579	0.577
STAR	0.491	0.495	0.703	0.581	0.608
DDAS	0.696	0.700	0.718	0.688	0.583
STN	0.693	0.495	0.703	0.581	0.580
DDAC	0.700	0.749	0.712	0.713	0.691
WCGN	0.884	0.763	0.747	0.884	0.882
GGA	**0.927**	0.828	0.800	0.790	0.799
Ours	0.905	**0.898**	**0.885**	**0.891**	**0.902**

TABLE IV: Cross-domain detection of C → B

	Accuracy	**Precision**	**Recall**	**F1**	**AUC**
APE	0.340	0.401	0.363	0.381	0.393
CDAC	0.341	0.352	0.409	0.378	0.481
STAR	0.337	0.465	0.401	0.431	0.462
DDAS	0.351	0.351	0.399	0.373	0.389
STN	0.618	0.627	0.626	0.626	0.680
DDAC	0.352	0.399	0.473	0.433	0.421
WCGN	0.428	0.461	0.447	0.454	0.462
GGA	0.486	0.597	0.530	0.562	0.521
Ours	**0.821**	**0.815**	**0.828**	**0.821**	**0.831**

It can be seen from the results that in the K → B task, the algorithm in this paper achieves the best performance in the five indicators of accuracy, precision, recall, F1 score and AUC. Especially on F1 score and AUC, it reached 0.892 and 0.901

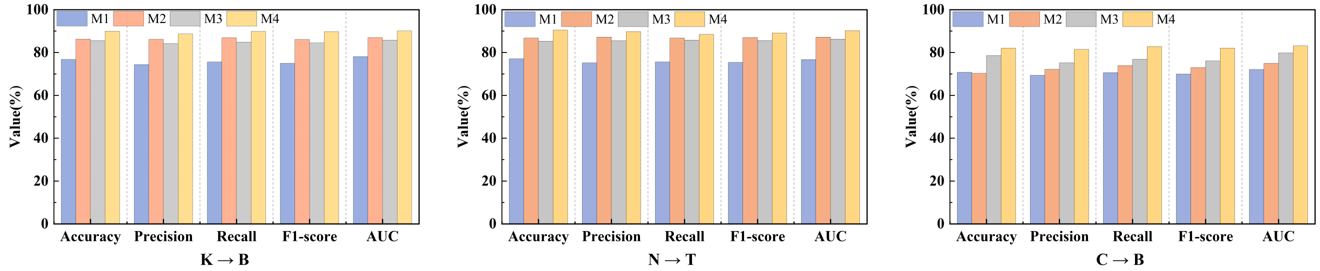
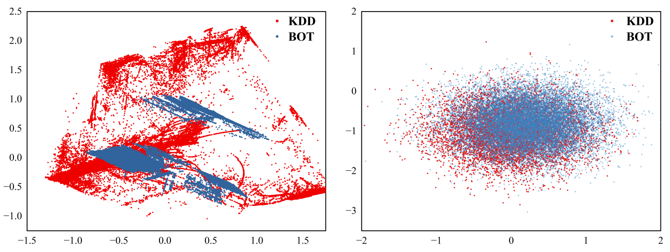


Fig. 4: Ablation experiment



(a) K → B original data

(b) K → B feature alignment data

Fig. 5: Data distribution before(a) and after feature alignment(b)

respectively, indicating that it still maintains excellent overall detection performance in the face of significant cross-domain differences and uneven class distribution. In the N → T task, although the accuracy of the proposed algorithm is slightly lower than that of GGA (0.905 vs. 0.927), it achieves superior performance across four key indicators—precision, recall, F1-score, and AUC. Notably, it attains a precision of 0.898, higher than the precision of GGA (0.828), and improves the F1-score by nearly 10% (0.891 vs. 0.790). It is worth noting that this task presents a severe imbalance between positive and negative classes. Significant improvements in recall and F1-score demonstrate the ability of the algorithm to mitigate class imbalance. In the C → B task, the proposed algorithm achieves the highest performance across all indicators compared to the baseline algorithms. This demonstrates its capability to reliably detect abnormal IoT traffic in cross-domain scenarios. The consistent precision and recall values further highlight the robustness of the algorithm across different IoT environments.

Based on the performance of the three tasks, it can be seen that the algorithm can not only accurately identify most class samples, but also maintain stable performance in detecting minority class attack traffic.

C. Ablation experiment

To comprehensively evaluate the effectiveness of each key component in the proposed algorithm, ablation experiments were conducted to individually analyze the contributions of the AEC-GAT feature extractor, Focal Loss, and class adaptive independent domain discriminator. This is achieved by removing or replacing the corresponding methods and analyzing the resulting changes in detection performance on the target domain. This section carries out ablation tests based on cross-domain tasks to keep data preprocessing, training rounds, optimizer and super parameters consistent. The specific comparison model design is shown in TABLE V.

TABLE V: Experimental design for ablation

	Model Configuration	AEC-GAT	Focal Loss	CA-DD
M1	AEC-GAT method removed	×	✓	✓
M2	Focal Loss replaced with Cross Entropy Loss	✓	×	✓
M3	Global domain discriminator used instead of class adaptive domain discriminator	✓	✓	×
M4	The proposed algorithm (ours)	✓	✓	✓

To verify the actual effectiveness of each key method in cross-domain intrusion detection, this paper conducted ablation experiments on representative cross-domain tasks, including K → B, N → T, and C → B. The experimental results are shown in Fig.4. From the comparison between the model M1 and the complete model M4, it can be seen that removing the AEC-GAT reduces accuracy to 76.8%, 77.1%, and 70.8% on the three tasks, indicating that the AEC-GAT method plays a crucial role in maintaining overall performance. After Focal Loss is replaced by the traditional CE loss function in model M2, F1-score in the three cross-domain tasks is lower than the complete model M4. This shows that Focal Loss can effectively promote overall performance improvement. By comparing M3 and M4, it can be observed that replacing the class-adaptive domain discriminator globally leads to accuracy drops of 4.0%, 5.2%, and 4.5% on the three cross-domain tasks, demonstrating the importance of fine-grained class-level alignment. Therefore, each method proposed in this paper has played an important role in cross-domain intrusion detection. The complete model M4 has achieved optimal performance in both tasks, verifying its wide adaptability and robustness in IoT scenarios.

D. Visual analysis of feature distribution

To further validate the effectiveness of the proposed algorithm in cross-domain feature alignment, Principal Component Analysis (PCA) was employed to reduce the dimensionality of high-dimensional features. Both source and target domain samples were projected into a two-dimensional space for visualization. The cross-domain intrusion detection performance was evaluated using the transfer task K → B. For this task, the feature distributions were systematically analysed, and the visualisation results confirm the effectiveness of the domain alignment, as shown in Fig. 5.

Fig.5a illustrates the original feature distribution in the K → B task, where a clear separation between source and target domains reveals pronounced domain shift. After applying the

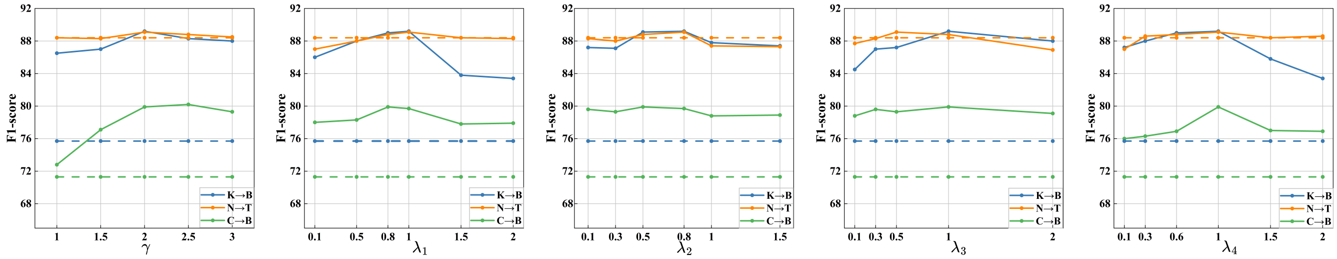


Fig. 6: Parameter analysis of loss weight

TABLE VI: The overhead analysis

	K → B			N → T			C → B		
	AEC-GAT	CA-DD	Ours	AEC-GAT	CA-DD	Ours	AEC-GAT	CA-DD	Ours
Memory usage (GB)	3.42	2.81	6.80	3.17	2.68	6.52	3.61	2.51	6.92
GPU usage (%)	55.67	52.34	75.09	51.55	49.30	72.24	57.45	54.63	72.28
Training time (s)	50.62	7.22	68.18	48.51	7.46	62.77	50.62	7.22	68.18
Throughput (samples/s)	8196	19607	7407	8849	19730	7936	7961	18323	7450

proposed algorithm (Fig.5b), the two distributions exhibit substantial overlap, indicating that the feature difference between the two domains is significantly reduced, and the domain alignment effect is obvious.

E. Real intrusion dataset detection experiment

In order to further verify the adaptability and effectiveness of the proposed algorithm in real-world IoT environments, this paper introduces two additional real industrial IoT monitoring datasets: Gas Pipeline (recorded as G) and Water Storage Tank (recorded as W). Gas Pipeline dataset is collected from the natural gas pipeline monitoring system, covering a variety of sensor readings and attack injection scenarios. Water Storage Tank dataset originates from an operational water treatment control system and includes multi-dimensional monitoring indicators under both normal and malicious conditions. Compared to traditional datasets, these industrial datasets exhibit higher noise levels, greater complexity, and more pronounced distributional differences, thereby providing a more realistic evaluation setting for IoT intrusion detection. In the cross-domain evaluation, two tasks were considered: K → W and N → G. The corresponding experimental results are presented in Table VII.

TABLE VII: Cross-domain detection of K → W and N → G

	Accuracy	Precision	Recall	F1-score
K → W	0.797	0.781	0.791	0.796
N → G	0.821	0.828	0.825	0.826

In the cross-domain evaluation on real industrial IoT datasets, the proposed algorithm maintains high detection performance, demonstrating its effectiveness in diverse operational scenarios. For the K → W task, the algorithm achieves an accuracy of 79.7%, demonstrating its ability to distinguish between normal and attack samples when traditional network traffic is applied to a real water treatment system. Similarly, in the N → G task, the algorithm achieves consistently high accuracy, indicating stable cross-domain detection even for gas pipeline monitoring data with pronounced distributional differences. Analysis of precision and recall indicates that the algorithm delivers consistently reliable detection, evidencing both robustness and generalizability in real-world IoT environments. Overall, these results substantiate that the proposed algorithm is not only effective on traditional IoT datasets but also practical and reliable for deployment in real-world industrial IoT environments.

F. Parameter analysis

In the overall loss function in this paper, the weight of each task is controlled by the parameters γ, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and the overall optimization goal is shown in Formula (19). In order to verify the rationality of parameter selection, this paper designs an analysis to examine the effects of γ and $\lambda_1 - \lambda_4$ on algorithm performance. Furthermore, transfer tasks K → B, N → T and C → B were selected for detailed analysis, with corresponding results visualized in Fig.6. Experiments show that the rationality of parameter values directly affects the cross-domain detection effect. The parameter configuration finally adopted in this paper is: $\gamma = 2$, $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 1.0$, $\lambda_4 = 1.0$. The best-performing baseline algorithm in each task is also indicated by a dashed line in the corresponding color. It can be observed that when the parameters change, the performance of the algorithm in this paper remains relatively stable without sharp fluctuations. In addition, in almost all parameter ranges, the solid line is always higher than the corresponding dotted line elevation, which means that the algorithm in this paper is superior to its corresponding best baseline algorithm in most parameter ranges.

G. The overhead analysis

To further evaluate the practicality of the proposed algorithm in IoT scenarios, the computational overhead is systematically analyzed in terms of memory usage, GPU utilization, training time, and throughput. The evaluation was conducted on three cross-domain tasks (K → B, N → T and C → B), and the results are shown in Table VI.

The results indicate that the complete algorithm (ours) maintains memory usage between 6.5–6.9 GB, confirming that the resource requirements are compatible with practical IoT deployment. The throughput of 7,400–7,950 samples/s across three cross-domain tasks further demonstrates the efficiency of the data processing capability. Both AEC-GAT and CA-DD exhibit overhead within desirable ranges, supporting feasibility in IoT deployment.

V. Conclusion

To address the challenge of degraded intrusion detection performance caused by data scarcity in IoT scenarios, this paper proposes a domain adaptive IoT intrusion detection algorithm based on AEC-GAT feature extraction and joint domain adversary. The algorithm enhances feature representation through causal reasoning and graph attention, mitigates class imbalance via Focal Loss, and achieves fine-grained class-level alignment using a class-adaptive independent domain discriminator. Extensive experiments including cross-domain comparative evaluations, ablation studies, parameter analyses, and real-world industrial IoT validations demonstrate that the proposed algorithm consistently outperforms existing baselines across multiple performance indicators. The results substantiate that each constituent method contributes significantly to enhancing algorithm robustness, generalization capability and deployment feasibility. Moreover, the overhead analysis shows that the algorithm maintains desirable computational efficiency, supporting feasibility in real-world IoT deployment. Future work will further explore causal structure enhancement and continual learning mechanisms to better meet practical IoT deployment needs.

VI. Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62376240 and 62302428, the S&T Program of Hebei under Grant No. 236Z0304G, the Innovation Capability Improvement Plan Project of Hebei Province under Grant No. 22567626H. The authors are grateful to valuable comments and suggestions of the reviewers.

References

[1] M. A. Khatun, S. F. Memon, C. Eising and L. L. Dhirani, ”Machine Learning for Healthcare-IoT Security: A Review and Risk Mitigation,” in IEEE Access, vol. 11, pp. 145869-145896, 2023, doi: 10.1109/ACCESS.2023.3346320.

[2] P. Baniya, A. Agrawal, K. Abid, J. Nath, B. K. Chaudhary and B. Kunwar, ”The Internet of Things: Security Challenges and Opportunities,” 2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 2024, pp. 153-158, doi: 10.1109/PARC59193.2024.10486356.

[3] Deshmukh, A.; Ravulakollu, K. An Efficient CNN-Based Intrusion Detection System for IoT: Use Case Towards Cybersecurity. Technologies 2024, 12, 203. https://doi.org/10.3390/technologies12100203

[4] C. Yin, Y. Zhu, J. Fei and X. He, ”A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,” in IEEE Access, vol. 5, pp. 21954-21961, 2017, doi: 10.1109/ACCESS.2017.2762418.

[5] Sun Z, Teixeira A M H, Toor S. GNN-IDS: Graph Neural Network based Intrusion Detection System[C]//Proceedings of the 19th International Conference on Availability, Reliability and Security. 2024: 1-12.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scho¨lkopf, and A. Smola, “A kernel two-sample test,” The Journal of Machine Learning Research, vol. 13, no. 1, pp. 723–773, 2012.

[7] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, “Optimal transport for domain adaptation,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 9, pp. 1853–1865, 2016.

[8] Layeghy S, Baktashmotlagh M, Portmann M. DI-NIDS: Domain invariant network intrusion detection system[J]. Knowledge-Based Systems, 2023, 273: 110626.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 1 (January 2016), 2096–2030.

[10] W. Wang et al., ”HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection,” in IEEE Access, vol. 6, pp. 1792-1806, 2018, doi: 10.1109/ACCESS.2017.2780250.

[11] C. Yin, Y. Zhu, J. Fei and X. He, ”A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,” in IEEE Access, vol. 5, pp. 21954-21961, 2017, doi: 10.1109/ACCESS.2017.2762418.

[12] Q. Wang, X. Wang, H. Liu, Y. Wang, J. Ren and B. Zhang, ”A Domain Adaptive IoT Intrusion Detection Algorithm Based on GWR-GCN Feature Extraction and Conditional Domain Adversary,” in IEEE Internet of Things Journal, vol. 11, no. 24, pp. 41223-41234, 15 Dec.15, 2024, doi: 10.1109/JIOT.2024.3457894.

[13] H. He and E. A. Garcia, ”Learning from Imbalanced Data,” in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[14] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[15] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem[C]//Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13. Springer Berlin Heidelberg, 2009: 475-482.

[16] A. Singh, N. Doraiswamy, S. Takamuku, M. Bhalerao, T. Dutta, S. Biswas, A. Chepuri, B. Vengatesan, and N. Natori, “Improving semi-supervised domain adaptation using effective target selection and semantics,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2709–2718.

[17] Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9268-9277.

[18] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[19] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation [C]//International conference on machine learning. PMLR, 2015: 1180-1189.

[20] Y. Yao, Y. Zhang, X. Li, and Y. Ye, “Heterogeneous domain adaptation via soft transfer network,” in Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 1578–1586.

[21] L. Wang, C. Huang, W. Ma, X. Cao, and S. Vosoughi, “Graph embedding via diffusion-wavelets-based node feature distribution characterization,” in Proceedings of the 30th ACM International Conference on Information and Knowledge Management, ser. CIKM ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 3478–3482.

[22] J. Li, G. Li, Y. Shi, and Y. Yu, “Cross-domain adaptive clustering for semi-supervised domain adaptation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2505–2514.

[23] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, “Semisupervised domain adaptation via minimax entropy,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8050–8058.

[24] Y. Yao, Y. Zhang, X. Li, and Y. Ye, “Discriminative distribution alignment: A unified framework for heterogeneous domain adaptation,” Pattern Recognition, vol. 101, p. 107165, 2020.

[25] J. Wu, H. Dai, Y. Wang, K. Ye and C. Xu, ”Heterogeneous Domain Adaptation for IoT Intrusion Detection: A Geometric Graph Alignment Approach,” in IEEE Internet of Things Journal, vol. 10, no. 12, pp. 10764-10777, 15 June15, 2023, doi: 10.1109/JIOT.2023.3239872.

[26] M. Kalisch and P. Bühlman, “Estimating high-dimensional directed acyclic graphs with the PC-algorithm,” J. Mach. Learn. Res., vol. 8, no. 3, pp. 613–636, 2007.

[27] Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains[J]. Machine learning, 2010, 79: 151-175.

[28] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in 2009 IEEE symposium on computational intelligence for security and defense applications. Ieee, 2009, pp. 1–6.

[29] N. Moustafa and J. Slay, “Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set),” in 2015 military communications and information systems conference (MilCIS). IEEE, 2015, pp. 1–6.

[30] Stiawan D, Idris M Y B, Bamhdi A M, et al. CICIDS-2017 dataset feature analysis with information gain for anomaly detection[J]. IEEE Access, 2020, 8: 132911-132921.

[31] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset,” Future Generation Computer Systems, vol. 100, pp. 779–796, 2019.

[32] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. den Hartog, “Ton iot: The role of heterogeneity and the need for standardization of features and attack types in iot network intrusion data sets,” IEEE Internet of Things Journal, vol. 9, no. 1, pp. 485–496, 2021.

Qian Wang received the B.S. degree, M.S. degree and Ph.D. degree in the School of Information Science and Engineering, Yanshan University, China, in 2009, 2012 and 2016, respectively. She has been in University of Hull from 2015-2016 as a visiting scholar. From 2016, she has become a lecture in the School of Information Science and Engineering, Yanshan University, China. And she has been an associate professor from 2019. She currently serves as a supervisor for both master and doctoral students. She is now at Edinburgh Napier University as a visiting scholar. Her research interests include machine learning, network security and healthcare analysis.

Menghui Fan is currently pursuing a master degree in the School of Artificial Intelligence (School of Software), Yanshan University, China. She received her bachelor degree in Computer Science and Technology at Hebei Normal University. Her research interests primarily focus on network security, with an emphasis on intrusion detection, domain adaptation, and intelligent data analysis in IoT environments. She has been actively engaged in developing adaptive and generalized intrusion detection models based on deep learning and causal reasoning techniques to enhance cross-domain detection performance. Her current work explores adversarial learning and graph-based representation methods to improve the robustness and interpretability of IoT security systems.

Zhijuan Wu is a master student at the School of Artificial Intelligence (School of Software), Yanshan University in Qinhuangdao, Hebei, China. She obtained her bachelor degree in Computer Science and Technology from Hebei University. Her research interests include data mining, multi-label classification, and intelligent information processing. She focuses on algorithm optimization and its applications to complex, real-world data problems. She has developed strong skills in model development, feature selection, and performance evaluation. Her current work investigates effective methods for handling high-dimensional and imbalanced data in multi-label learning. She aims to contribute to machine learning through both theoretical study and practical innovation in real applications.

Hongnian Yu received the B.S. degree, M.S. degree, and Ph.D. degree from Harbin Institute of Technology, China, Northeast Heavy Machinery Institute, China, and King’s College London, UK, in 1982, 1985, and 1994, respectively. He has successfully supervised 20 Ph.D. theses and 18 Master by Research theses, and has trained 12 post-doctoral research fellows. He has published more than 200 journal and conference papers. He has held several research grants totaling approximately eight million pounds from the UK EPSRC, the Royal Society, the European Union, AWM, and industry. His research interests include robotics and nonlinear control, complex network modeling, wireless communications, and applied artificial intelligence.

Yongqiang Cheng received the B.S. and M.S. degrees from Tongji University, Shanghai, in 2001 and 2004, respectively, and completed his Ph.D. at the University of Bradford in 2010. Between 2004 and 2007, he worked full-time at the ZTE R&D Centre in Shanghai. He is now a professor of Artificial Intelligence aligned with Digital Healthcare at the University of Sunderland. He serves as the REF lead for the School of Computer Science and Technology, deputy director of the Research Centre for Emerging Technologies and Innovation, and leads the Data Science and AI Workstream at the John Dawson Drug Discovery and Development Institute. His research interests include artificial intelligence, trustworthy AI, machine learning, wireless sensor networks, smart systems, and digital technologies.

Bing Zhang received the B.S. degree from the College of Computer and Information Technology, Three Gorges University, China, in 2012, and the Ph.D. degree from the School of Information Science and Engineering, Yanshan University, China, in 2018. Since 2018, he has been a lecturer with the School of Information Science and Engineering, Yanshan University, China, and has been an associate professor since 2020. He has been in the Norwegian University of Science and Technology as a visiting scholar. He currently supervises both master’s and doctoral students. His research interests include data mining, network security, machine learning, and software security.