Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

面向联邦基础模型的安全评测与防御方法研究

博士研究生　李佳龙

2025 年 12 月 21 日

问题回溯

- 相关内容
  - 2025.10.27 陈星星《面向数据异构与通信高效的联邦大模型优化与应用研究》
  - 2025.09.16 赵怡清《扩散模型的后门攻击研究》
  - 2023.04.09 杨得山《联邦学习的后门防御方法》
  - 2022.08.30 杨得山《联邦学习的后门攻击方法》
  - 2022.05.15 郝靖伟《联邦学习及其后门攻击方法初探》

内容提要

- 预期收获
- 内涵解析与研究目标
- 研究背景与研究意义
- 研究历史与现状
- 知识基础
- 算法原理
  - SecFFT
  - FL-IDS
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 熟悉联邦基础模型在IoRT/IIoT场景下的系统架构与联邦微调（FFT）范式
  - 掌握针对基础模型（LVLM）的隐蔽后门攻击机制及基于频域分布与长期意图检测的防御新思路
  - 了解基础模型驱动的联邦入侵检测框架及利用CGAN数据增强与知识蒸馏解决数据不平衡与资源受限问题的方法
  - 理解在联邦基础模型中引入安全聚合与差分隐私的基本思路

内涵解析与研究目标

- 研究目的
  - 面向联邦基础模型，研究其在IoRT/IIoT场景下的安全评测与防御
  - 结合频域分析、长期意图检测技术实现对隐蔽后门攻击的精准防御
  - 结合知识蒸馏、CGAN数据增强技术实现在资源受限场景下联邦入侵检测性能与隐私保护能力的平衡
- 内涵解析
  - 联邦微调：利用跨节点分布的数据和算力，对LVLM等基础模型进行特定任务微调，解决通用模型在特定领域中泛化能力不足的问题
  - 隐蔽后门攻击：攻击者通过多轮次、低幅度或频域伪装方式实现恶意触发器植入。能绕过基于浅层统计特征的防御，具有高隐蔽性
  - 基础模型驱动的入侵检测：利用预训练基础模型强大的语义理解能力，通过加速训练过程来缓解计算资源短缺问题

研究背景与研究意义

- 研究背景
  - 集中式微调面临数据孤岛与通信瓶颈的挑战，通用基础模型与联邦微调的兴起引入全新的攻击面与资源挑战问题
  - 现有集中式防御方法难以应对基础模型联邦微调阶段高隐蔽、高复杂的后门攻击
    - 新型攻击方法比如攻击者利用频域伪装模仿正常用户行为，导致基于浅层语义特征或显式梯度异常的检测方法失效
    - 联邦微调范式中攻击者会采用多轮次、低幅度的渐进式注入策略，导致单轮检测机制难以有效应对长期威胁
  - IoRT/IIoT高动态、低资源特性与联邦基础模型高算力、高数据需求间的冲突
    - 无人机网络中攻击样本数据稀缺、Non-IID特性导致传统联邦入侵检测方法泛化能力不足
    - 模型训练过程涉及大量敏感通信数据，导致计算效率与数据隐私难以平衡
- 研究意义
  - 解决联邦基础模型在边缘网络中面临的隐蔽后门攻击以及算力资源受限问题

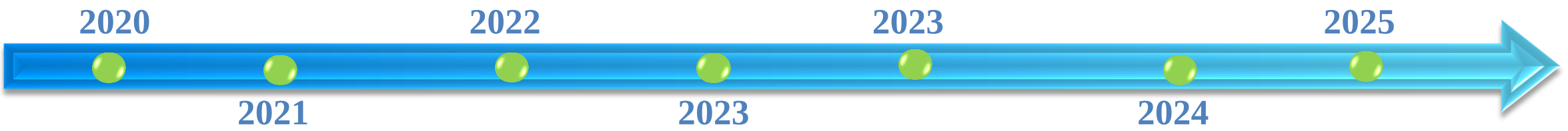
研究历史和现状

Fung等人提出了FoolsGold，一种基于更新相似度的防御机制。该方法通过检测客户端上传梯度的余弦相似性来识别恶意节点，是联邦学习防御Sybil攻击的经典基线方法。

Zhang等人提出了NeuroToxin，一种针对联邦学习的持久性后门攻击。通过对训练过程中变化较小的模型参数进行投毒，提高了后门的持久性和隐蔽性。

Bansal等人提出了CleanCLIP，专门针对多模态对比学习模型的防御方法。通过对视觉和文本编码器进行无监督微调，在保持干净数据性能的同时减轻数据投毒的影响。

Jiao等人提出了一种面向无人机辅助工业物联网的基础模型驱动联邦入侵检测框架。利用条件生成对抗网络（CGAN）解决数据稀缺与不平衡问题，并通过知识蒸馏将基础模型的能力迁移至边缘设备，结合差分隐私实现了资源受限场景下的高精度、隐私保护型入侵检测。



Ramadan等人提出了一种面向无人机网络的分布式入侵检测系统，利用循环神经网络（RNN）增强攻击检测和决策能力，但在应对数据不平衡和隐私保护方面仍有局限。

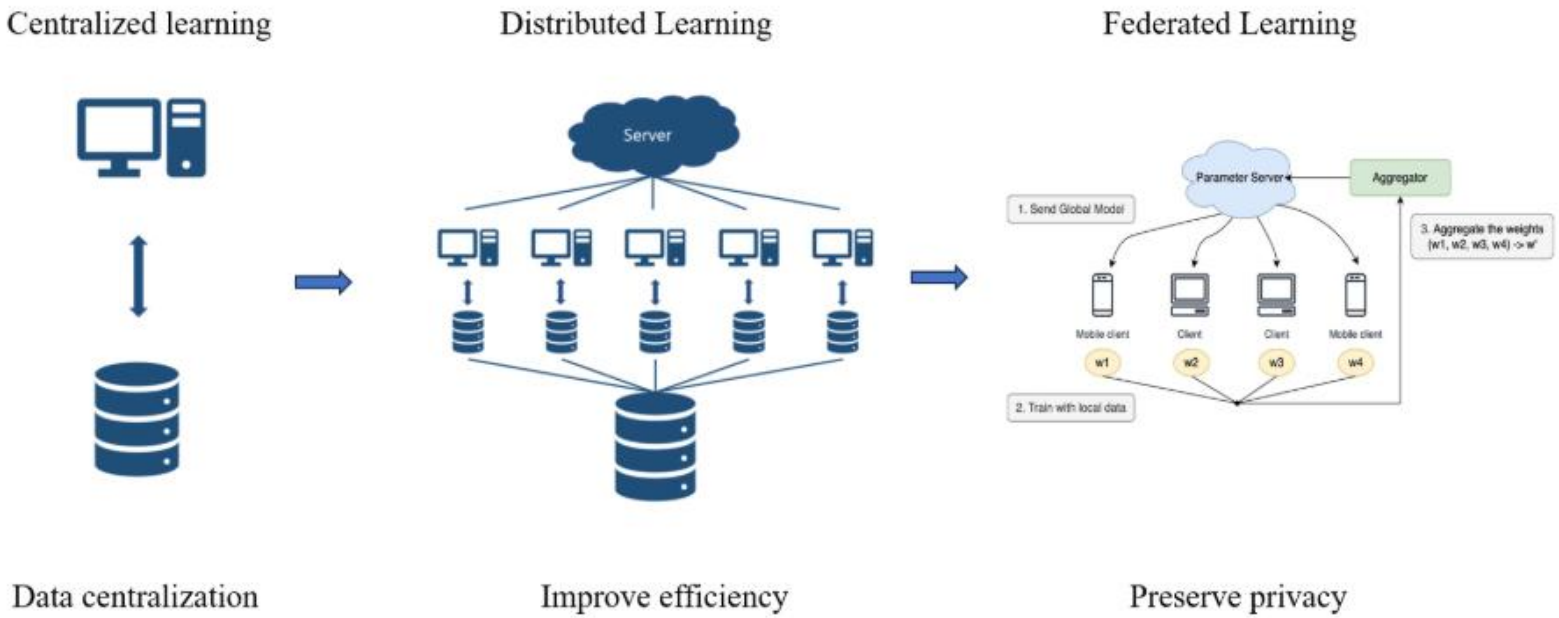
Huang等人提出了Lockdown，一种针对联邦学习的隔离子空间训练防御机制。该方法通过随机分割训练子空间来解耦恶意参数与良性参数的关联，并引入多数共识机制剪除异常神经元，有效抵御了联邦场景下的参数耦合型后门攻击。

Zhou等人提出了一种针对机器人物联网（IoRT）中大视觉语言模型（LVLM）联邦微调的后门防御框架（SecFFT）。该方法通过频域分布一致性检测瞬时攻击行为，并结合长期意图识别机制，有效防御了隐蔽性极强的多轮后门攻击，保障了联邦微调过程的安全性。

知识基础 联邦学习

- 联邦学习（Federated Learning，FL）
  - 一种分布式机器学习框架，2016年谷歌提出，允许多个参与方（如设备、机构或企业）在不共享原始数据的情况下协同训练机器学习模型
  - 数据不动模型动，异构数据兼容
  - 典型聚合方式（FedAvg）：

$$W^t = \sum_{i=1}^{N} \frac{|D_i|}{|D|} W_i^t$$



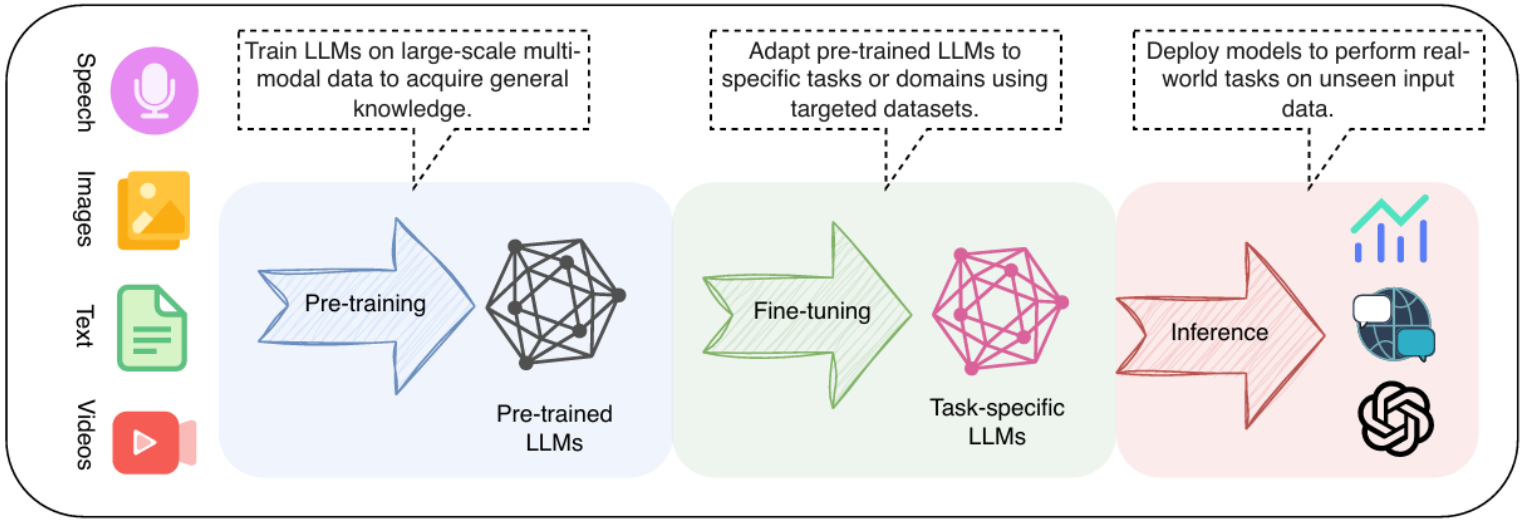
知识基础 联邦基础模型VS联邦大语言模型

对比维度	联邦基础模型（FedFM）	联邦大模型（FedLLM）
概念范畴	涵盖视觉、音频、多模态等广泛领域，不仅局限于文本，LLM是FL的子集	专注于NLP领域
核心能力	具备跨模态的泛化能力，适用于复杂感知任务	具备强大的文本生成、理解与逻辑推理能力
通信与计算开销	较高，参数量通常在数亿级别（CLIP），在边缘设备（机器人）可行性高	极高，参数量通常在百亿级以上，对边缘节点的显存与带宽要求极高
数据异构性	极高	较高
典型模型	CLIP，BERT，Segment Anything (SAM)	GPT系列，LLaMA
LoRA适配性	强，基座模型较小（ViT），边缘设备可轻松加载并进行低秩微调	弱，基座模型庞大，即使使用LoRA，边缘设备显存也难以承载基座权重
网络安全挑战	隐蔽后门与跨模态投毒	隐私泄露与指令攻击

联邦基础模型能够解决边缘落地的难题，但也带来比大模型更隐蔽难防的安全挑战

知识基础 联邦微调范式

- 联邦微调(Federated Fine-Tuning, FFT)
  - 定义
    - 不共享原始数据
    - 利用分布式边缘节点（机器人、无人机）上的私有数据和算力，协同微调预训练基础模型，快速适配下游任务
  - 工作流程
    - 全局模型下发与参数冻结
    - 本地微调与梯度计算
    - 梯度上传与全局安全聚合



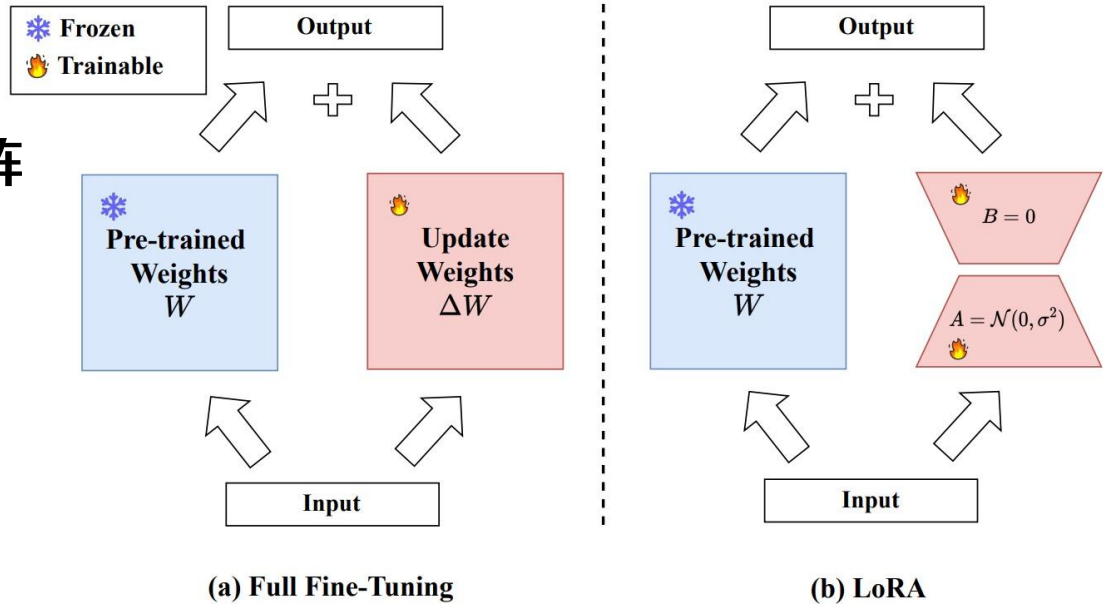
- IoRT/IIoT场景下FFT的挑战
  - 分布式特性导致攻击面扩大与隐蔽性增强
  - 设备异构与数据异构（Non-IID）
  - 边缘设备算力有限，数据稀缺

知识基础 参数高效微调（PEFT）

- 低秩适配（Low-Rank Adaptation，LoRA）
  - 冻结预训练模型权重，仅在Transformer层注入可训练的低秩分解矩阵
  - 数学表达
    - W_0为冻结的原始参数矩阵，ΔW为低秩增量矩阵
    - A是核心矩阵（秩接近W_0），B是低秩矩阵：

      $$W = W_0 + \Delta W \qquad \Delta W = BA$$

    - 只更新低秩矩阵B
  - 在FFT中的应用
    - 全局模型：$\Theta = \Theta_p + \Theta_g$
    - 冻结的全局模型参数：$\Theta_p = \sum W_0$
    - 参与FFT的增量参数：$\Theta_g = \sum W$
  - 意义
    - 大幅降低联邦学习架构的通信开销和边缘节点的计算负担



知识基础 隐蔽后门攻击

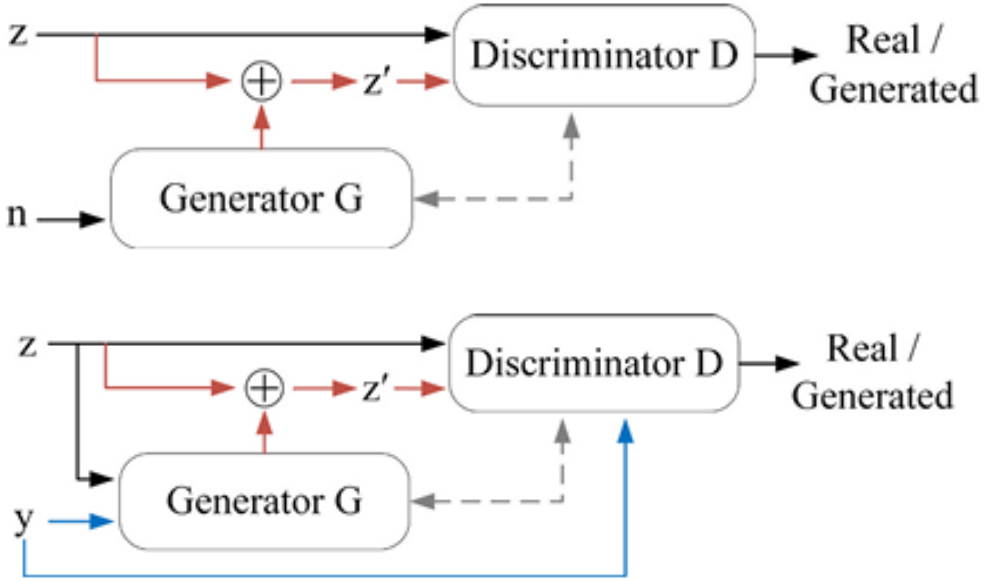
- 攻击目标
  - 在全局模型中植入后门，使模型对带有特定触发器δ的输入产生错误分类（$\Theta(x+\delta)=y_{target}$），同时保持对干净样本的分类精度（$\Theta(x)=y$）
- FFT场景中新型攻击策略
  - 空间隐蔽
    - 通过微调参数或修改特征，使恶意更新在特征空间中模仿良性用户，绕过基于距离（欧氏/余弦）的检测
  - 时间隐蔽
    - 多轮次复合攻击：不在单轮训练中植入后门，而是采用多轮次、低幅度的渐进式注入策略
    - 策略碎片化：限制更新的幅度或者角度，将微小单轮恶意意图尽可能隐藏在长期的良性行为伪装下

知识基础 频域分析与几何检测理论

- 频域变换
  - 利用一维离散余弦变换（1-D DCT-II）将模型梯度的浅层空间特征变换为频域分布
  - 恶意更新与良性更新在低频分量上存在本质的分布一致性差异，可以作为深层鉴别特征，同时显著减少计算开销
- 几何意图建模
  - 最小包围球面（MEHB）
    - 在高维空间中构建覆盖节点历史更新轨迹（射线）的最小超球面
    - 球心代表节点的长期意图点，半径反映意图的置信度
  - 局部异常因子（LOF）
    - 基于密度的离群点检测算法，通过比较某节点及其邻域的密度来判断是否为恶意节点

知识基础 数据增强与知识蒸馏

- 数据增强
  - 解决数据稀缺和数据不平衡问题
  - 过采样&欠采样、生成对抗网络、条件生成对抗网络（CGAN）
    - 包含生成器$G(z,y)$和判别器$D(x,y)$，引入条件向量y（比如攻击类型标签）
    - 目标：$G(z,y) \rightarrow x$，生成器基于噪声向量z和条件向量y合成样本x



$$\min_G \max_D \{\mathbb{E}_{x\sim p_{data}(x)}[log\ D(x,y)] + \mathbb{E}_{z\sim p_z(z)}[log\ (1-D(G(z,y),y))]\}$$

- 知识蒸馏
  - 采用“教师–学生”模式，利用预训练的基础模型（教师）生成软标签
  - 利用软标签指导学生模型训练，在继承泛化能力的同时大幅降低资源需求

SecFFT：Safeguarding Federated Fine-Tuning for Large Vision Language Models Against Covert Backdoor Attacks in IoRT Networks

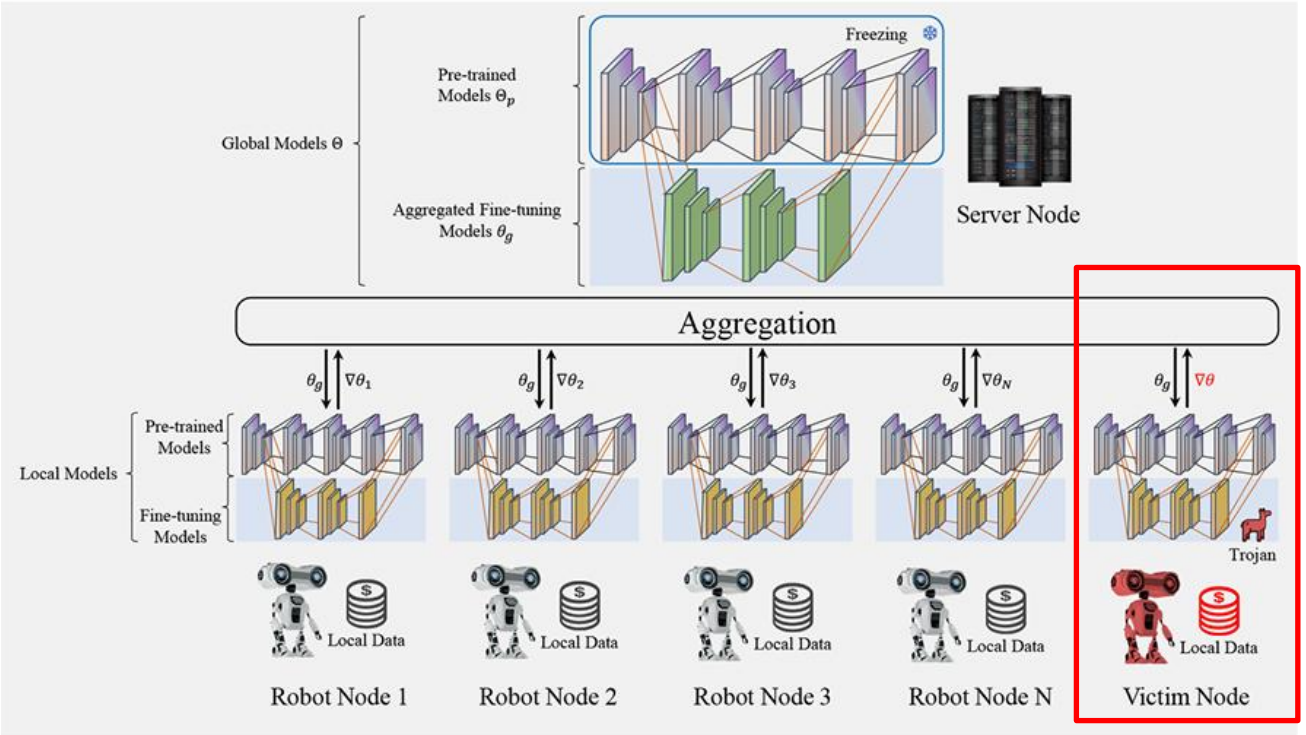
算法原理 TIPO

T	目标	在IoRT网络中，防御针对LVLM联邦微调阶段的隐蔽后门攻击
I	输入	N个机器人节点的本地数据集$D_i = \{x_{i,j}, y_{i,j}\}$（FMNIST图像及标签） 各节点连续$T$轮训练中上传的局部梯度更新序列$\{\nabla\theta_i^1, \nabla\theta_i^2, \ldots, \nabla\theta_i^T\}$
P	处理	1.瞬时检测：基于频域变换（DCT）提取深层特征 2.长期检测：基于最小包围超球（MEHB）构建长期意图，识别多轮攻击策略 3.安全聚合：基于意图置信度的自适应加权聚合
O	输出	1个能够有效防御隐蔽后门攻击的全局模型θ_g^t

P	问题	1.现有防御难以应对LVLM联邦微调中的高隐蔽、多轮次后门攻击 2.频域伪装与碎片化注入策略导致传统的基于浅层特征的单轮检测失效
C	条件	攻击假设：恶意节点占比<50%，良性节点更新方向更一致且聚合度更高
D	难点	SecFFT针对的是相对同质的网络场景，应用到异构性和多模态特性更明显的IoT场景时性能会骤降
L	水平	IEEE Internet of Things Journal 2025（SCI一区）

算法原理 创新点

- **SecFFT**
  - 瞬时行为攻击感知模块，基于频域分布一致性的检测，利用DCT挖掘梯度在频域（主要是低频分量）上的深层语义差异。在降维的同时精准识别微小扰动的单轮隐蔽攻击，但对多轮渐进式攻击效果不佳
  - 长期攻击意图检测模块，构建基于MEHB的长期意图模型，将碎片化的单轮攻击行为还原为完整意图。引入LOF评估节点意图的离群程度，结合置信度实现全局模型安全聚合

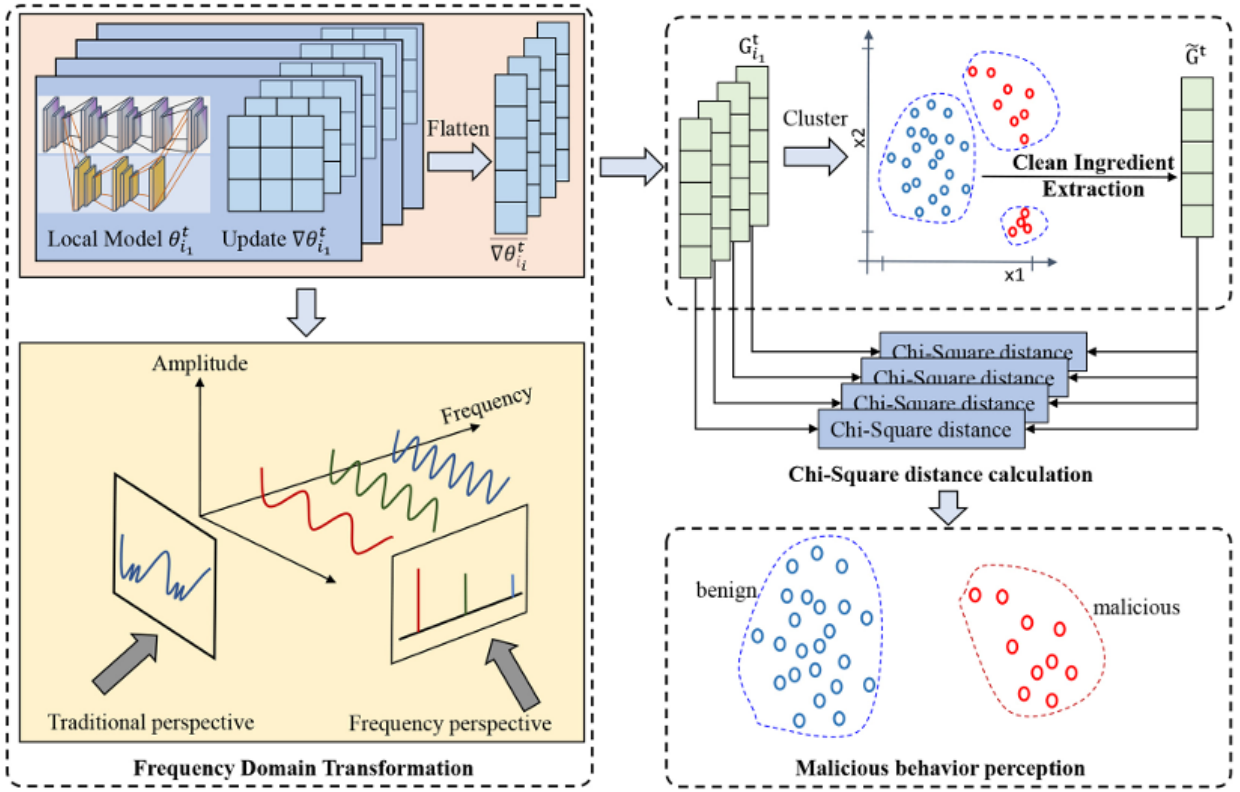


算法原理 瞬时攻击行为感知

- 频域分布一致性检测
  - 用1D DCT-II进行频域变换
    - 对每个节点的梯度更新$\nabla\theta_i^t$进行一维离散余弦变换
    - 提取前m个低频分量作为特征向量G_i^t：

$$G_i^t = Trunc\left(DCT\left(Flatten\left(\nabla\theta_i^t\right)\right), m\right)$$

  - 提取频域中良性节点更新的主成分
    - 对G_i^t使用基于密度的聚类算法HDBSCAN选取包含节点数最多的簇作为良性候选集
    - 对候选簇内所有节点构成的矩阵H^t进行SVD $H^t = \left(G_{i_0}^t, G_{i_1}^t, \ldots, G_{i_n}^t\right) \in \mathbb{R}^{m\times n}$



- SVD分解后的主特征向量$\tilde{G}^t$作为频域上的良性主方向：$\tilde{G}^t = \frac{H^t\xi_{max}^t}{\sqrt{\lambda_{max}^t}}$

算法原理 瞬时攻击行为感知

- 基于卡方距离的异常判别
  - 卡方距离计算
    - 计算每个节点的低频向量G_i^t与主成分$\tilde{G}^t$之间的散度

    得到散度集$S^t = \{Chi_1^t, Chi_2^t, \ldots, Chi_N^t\}$：

    - $Chi_i^t = \sqrt{\sum_{k=0}^{m-1} \frac{\left(G_i^t[k] - \tilde{G}^t[k]\right)^2}{|\tilde{G}^t[k]| + \epsilon}}$

    - 相比欧氏距离，卡方距离更关注分布形态差异，对隐蔽攻击更敏感
  - 单轮恶意行为感知
    - 对散度集S^t进行K-Means二分类
    - 簇中节点数较大的类归为良性节点U_{nor}，另一类U_{mal}

Algorithm 1 Instantaneous Attack Behavior Perception

Input: d, N, $(\nabla\theta_0^t, \nabla\theta_1^t, \ldots, \nabla\theta_N^t) \in \mathbb{R}^{N\times d}$, m ▷ d is the dimension of each update; N is the number of nodes participating during each round; $(\nabla\theta_0^t, \nabla\theta_1^t, \ldots, \nabla\theta_N^t) \in \mathbb{R}^{N\times d}$ is the local updates from robot nodes during the t-th round; m is the length of low-frequency vector
Output: U_{nor}, U_{mal} ▷ benign nodes, malicious nodes
/* Frequency Domain Transformation */
for $R_i \in \{R_1, \ldots, R_N\}$ **do**
  $G_i^t \leftarrow Trunc(DCT(Flatten(\nabla\theta_i^t))), m)$
end for
/* Clean Ingredient Extraction */
$(C_0^t, C_1^t, \ldots, C_\kappa^t) \leftarrow Clustering(G_0^t, G_1^t, \ldots, G_N^t)$ ▷ κ denotes the number of clusters
$C_{\max}^t \leftarrow \arg\max_{C_j^t} |C_j^t|$ ▷ $|C_j^t|$ denotes the number of nodes in cluster C_j^t, $j = 1, 2, \ldots, \kappa$
$H^t \leftarrow (G_{i_0}^t, G_{i_1}^t, \ldots, G_{i_n}^t) \in \mathbb{R}^{m\times n}$ ▷ Stacking to form Matrix H^t, where $R_{i_0}, R_{i_1}, \ldots, R_{i_n} \in C_{\max}^t$.
$\hat{H}^t \leftarrow (H^t)^T H^t$
$\lambda_{\max}^t, \xi_{\max}^t \leftarrow eig(\hat{H}^t)$ ▷ Calculating the maximum singular value and its corresponding eigenvector
$\tilde{G}^t \leftarrow \frac{H^t \xi_{\max}^t}{\sqrt{\lambda_{\max}^t}}$ ▷ The clean ingredient
/* Chi-square distance calculation */
for $R_i \in \{R_1, \ldots, R_N\}$ **do**
  $Chi_i^t \leftarrow \sqrt{\sum_{k=0}^{m-1} \frac{(G_i^t[k] - \tilde{G}^t[k])^2}{|\tilde{G}^t[k]| + \epsilon}}$
end for
$S^t \leftarrow \{Chi_1^t, Chi_2^t, \ldots, Chi_N^t\}$ ▷ The Distance differences calculated by Chi-square distance.
/* Single-round malicious behavior perception */
$\{C_1^{t\prime}, C_2^{t\prime}\} \leftarrow KMeans(S^t, 2)$ ▷ Cluster S into 2 clusters using KMeans.
$U_{nor} \leftarrow C_{\max}^t$, $U_{mal} \leftarrow \{C_i^t \mid i \neq \max\}$

频域分布一致性检测

卡方距离异常判别

算法原理 长期攻击意图检测

- 最小包围超球（MEHB）建模
  - 射线建模
    - 将历史更新当成高维空间的射线：

$$l_i^{t'} = Flatten\left(\theta_g^{t'-1}\right) + \alpha\overline{\nabla\theta_i^{t'}}$$

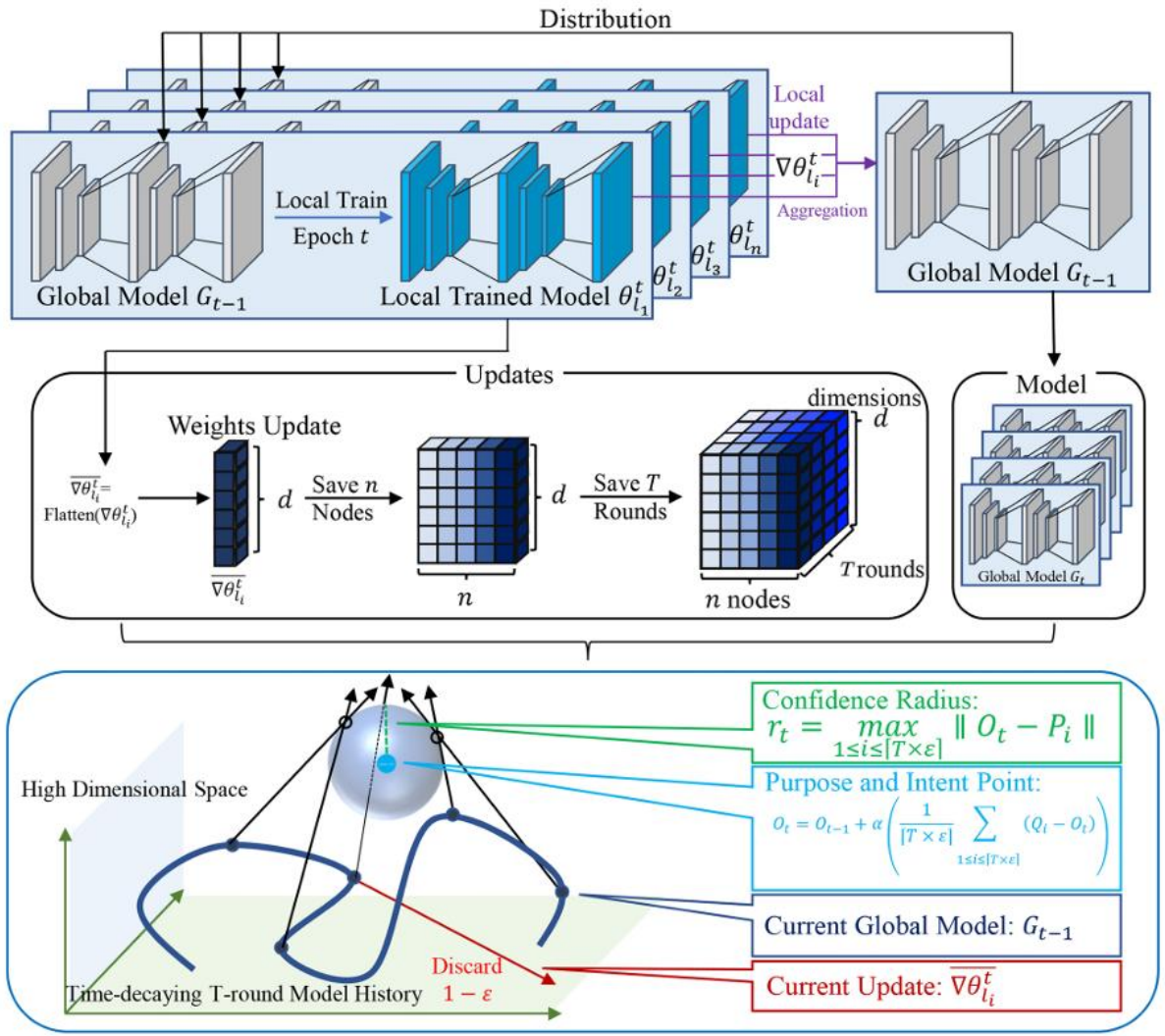
    - 起点：上一轮全局模型参数$\overline{\theta_g^{t'-1}}$

      方向：节点更新梯度方向$\overline{\nabla\theta_i^{t'}}$
  - 提取意图点与意图半径
    - 利用 MEHB算法求解一个最小超球面，使其覆盖该节点的ζT条历史射线：

$$\min_{O_i, r_i} r_i \ \ s.t. \left|\left\{t' : \left|\left|O_i - \tilde{O}_i^{t'}\right|\right| \le r_i\right\}\right| \ge \zeta T$$

    - 球心O_i：某节点的长期意图点

      半径r_i：意图的置信度

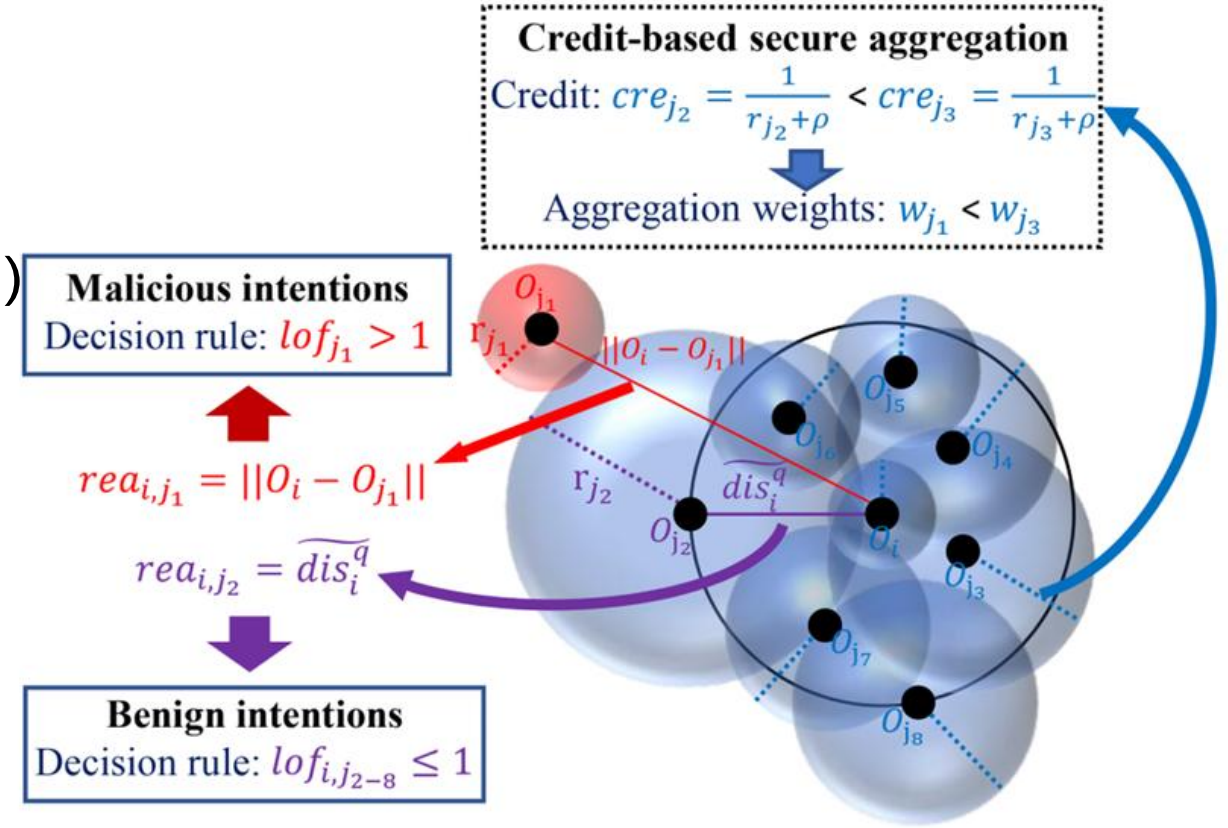


算法原理 长期攻击意图检测

- 基于局部异常因子（LOF）的意图筛选
  - 密度离群检测
    - 计算每个节点意图点O_i 的LOF
  - 恶意节点剔除
    - $LOF \gg 1$（节点周围密度显著低于邻域密度）

      表明这个节点的长期意图显著偏离其他节点，

      判定为恶意节点



- 基于信誉的动态安全聚合
  - 信誉计算
    - $cre_i = \frac{1}{r_i+\rho}$ $\quad w_i = \frac{cre_i}{\sum_{i=1}^{N} cre_i}$
    - r_i越小，历史行为越一致，越可信
  - 全局模型安全聚合
    - $\theta_g^t = \theta_g^{t-1} + \sum_{i \epsilon U_{nor}} w_i \cdot \nabla\theta_i^t$

实验流程 实验设置

- 数据集：FMNIST(FashionMNIST)
  - 60,000张灰度图像，10个类别
  - 模拟IoRT场景下的视觉感知任务
- 基础模型：CLIP-ViT-B/32
  - 12层Transformer编码器
  - 利用LoRA进行高效联邦微调
- 攻击基线：
  - MR（2020年）：强力替换全局模型
  - EDGE CASE（2020年）：利用长尾分布样本植入后门
  - NEUR（2022年）：针对更新频率低的参数投毒，持久性极强
- 防御基线：FedAvg（2016年）、FoolsGold（2020年）FLTrust（2020年）、Flame（2022年）

对比维度	MNIST	FashionMNIST
时间	1998年	2017年
类别数量	手写数字(0~9)	时尚单品/衣物（T恤、鞋、包等）
图片内容	10类	10类
训练集数量	60,000张28×28像素的灰度图	60,000张28×28像素的灰度图
测试集数量	60,000张28×28像素的灰度图	60,000张28×28像素的灰度图
任务难度	低，特征明显	中，纹理复杂
标签含义	0：数字0	0：T恤/上衣

- 评估指标
  - 瞬时攻击行为感知：

ASR（攻击成功率）：指成功使模型误分类为目标类别的中毒样本所占的比例

TSR（测试成功率）：全局模型在测试集上的准确性

- 长期攻击意图检测：

Accuracy：$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

Recall：$Recall = \frac{TP}{TP + FN}$

FPR（假阳性率）：$FPR = \frac{FP}{FP + TN}$

FNR（假阴性率）：$FNR = \frac{FN}{TP + FN} = 1 - Recall$

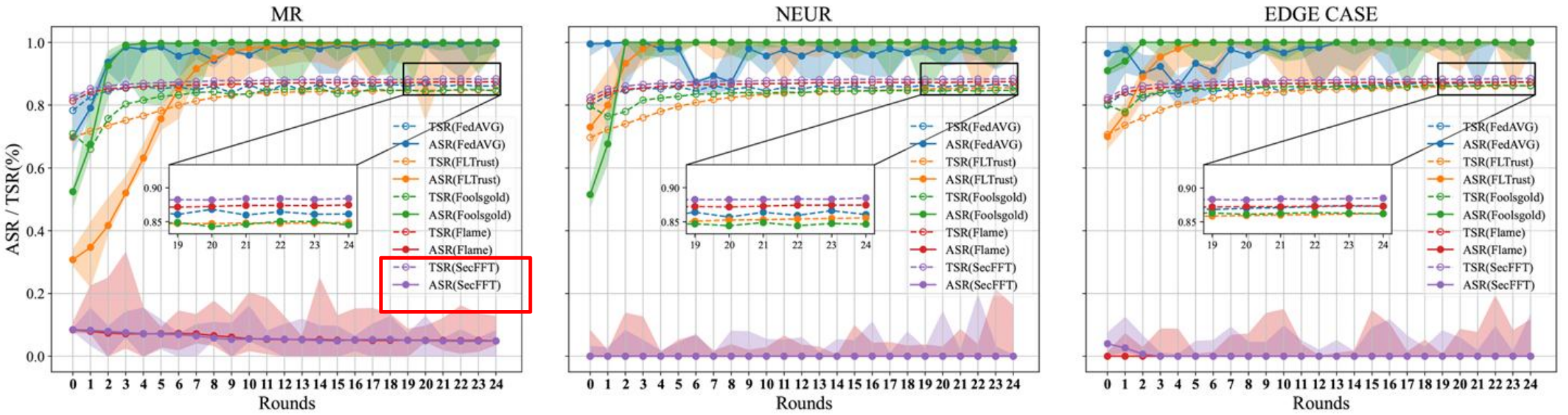
AUC（曲线下面积）：反映模型区分正样本和负样本的整体能力

MCC（马修斯相关系数）：$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

		预测类别	
		0	1
真实类别	0	True Negative(TN)	False Positive(FP)
	1	False Negative(FN)	True Positive(TP)

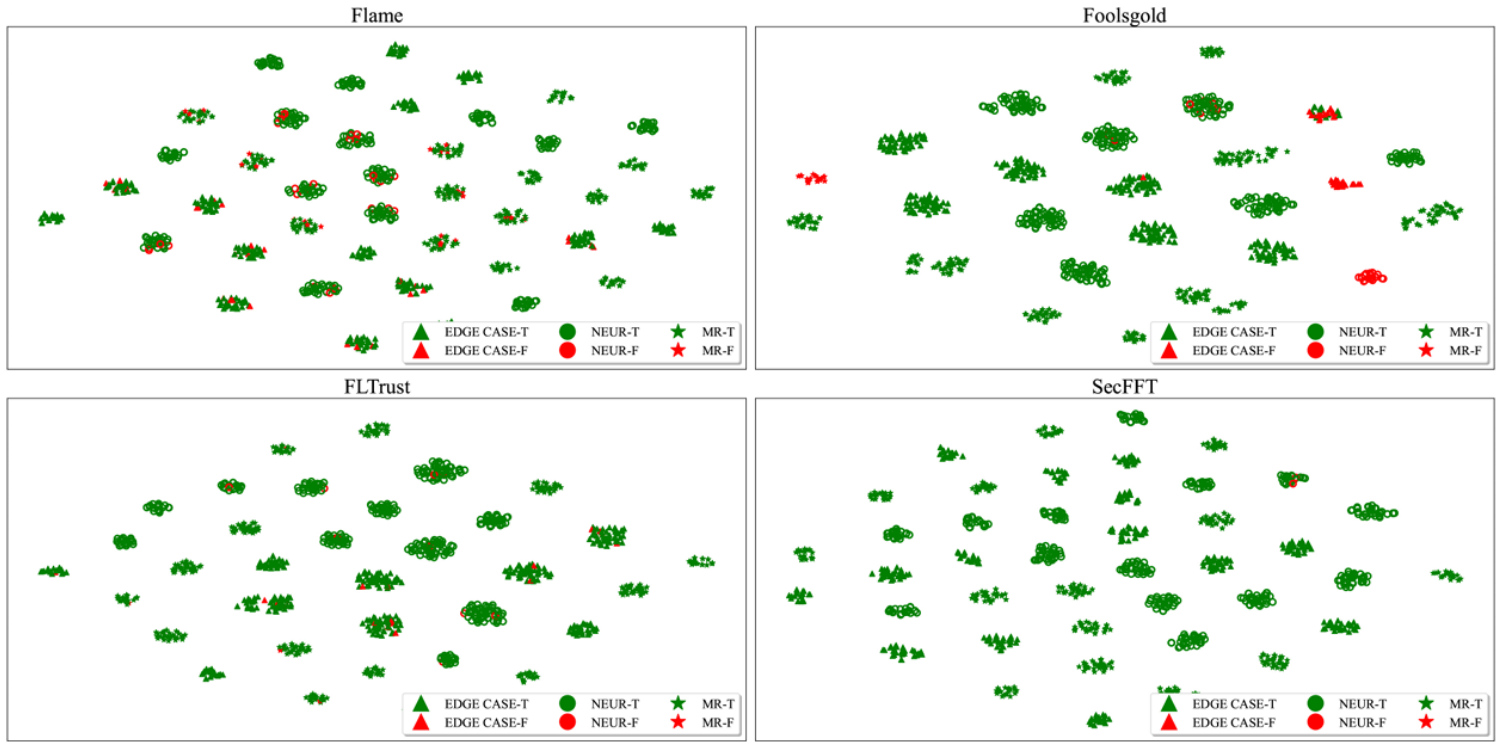
实验结果 整体防御性能

- 模型完整性与抗中毒能力评估



- SecFFT：在所有攻击场景下，ASR始终接近0，同时TSR始终较高
- 对比分析：
  - FLTrust：在MR攻击初期ASR较低，但随迭代轮次增加而增加
  - Flame：能降低ASR，但由于引入了大量噪声，导致TSR显著下降，影响模型可用性
- 结论：SecFFT能够实现防御效果与模型性能的最佳平衡

实验结果 瞬时攻击行为感知性能

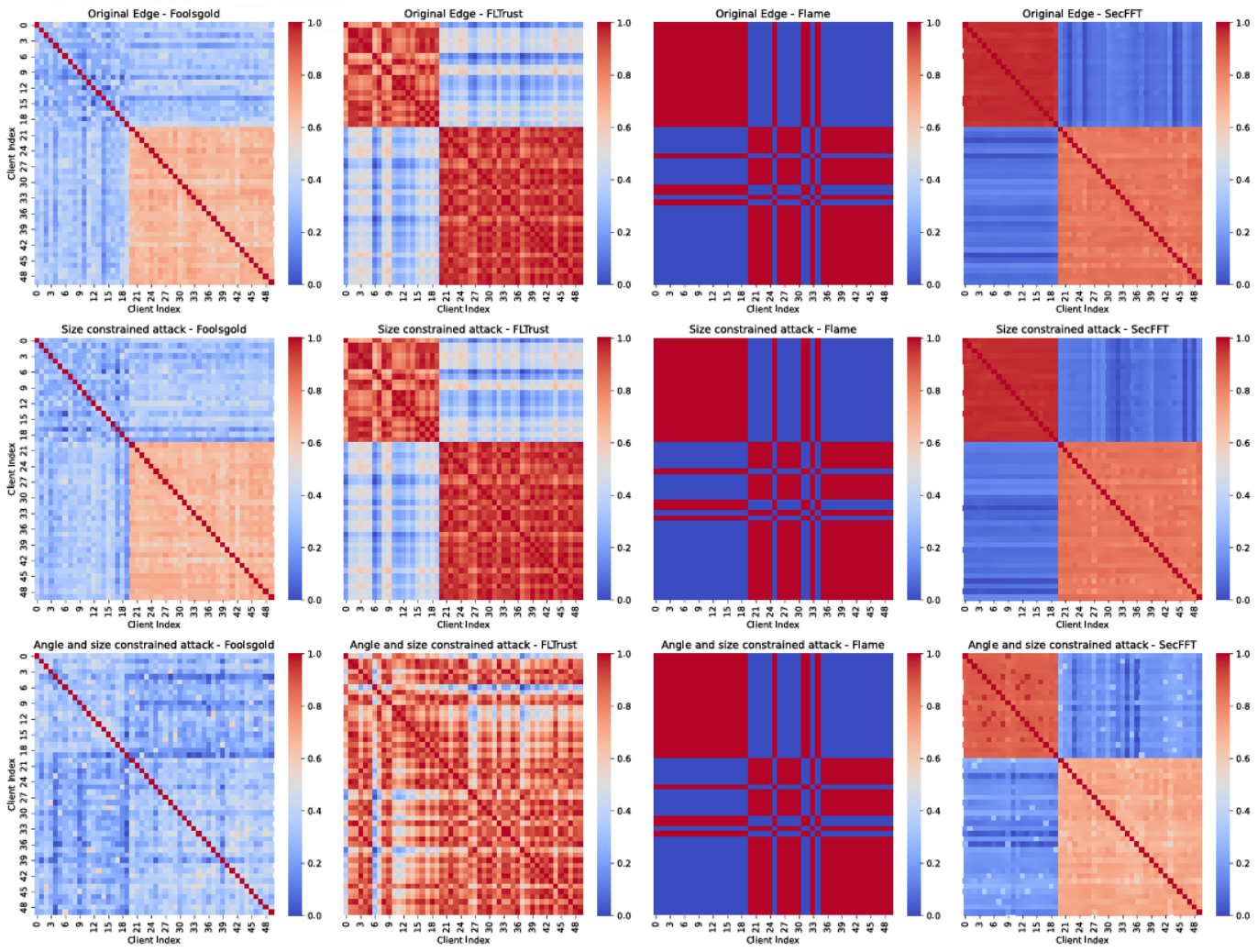


– 可视化验证
  - 利用t-SNE（t-分布随机临近嵌入）算法将各节点在联邦微调阶段的梯度更新映射到二维平面

– 结果分析：
  - SecFFT：红色误判点最少，误报率仅为0.2%
  - FoolsGold：误报率高达9.5%，对于良性异构节点效果尤其差
  - Flame：误报率8%，误判节点分布广泛

– 结论：基于频域分布一致性的特征提取，比传统的空间域特征更能精准捕捉微小的恶意扰动

实验结果 长期攻击意图检测性能



- 攻击场景：攻击者采用多轮复合策略（NERU攻击+限制更新大小+限制更新角度），尽可能隐藏单轮行为特征
- 结果分析：
  - SecFFT：即使单轮更新微小，长期意图轨迹仍保持清晰
  - FoolsGold/ FLtrust：在引入“大小+角度”双重约束后，检测能力显著下降，恶意节点与良性节点混淆严重
- 结论：Flame可以有效识别一些高隐蔽攻击，但引入多轮复杂攻击后，防御机制很快会被绕过

实验结果 量化指标对比

Strategy	Defense	Acc.	Rec.	FPR	FNR	AUC	MCC
None	Foolsgold	0.96	0.90	**0.00**	0.10	0.95	0.92
	FLTrust	**1.00**	**1.00**	**0.00**	**0.00**	**1.00**	**1.00**
	Flame	0.92	**1.00**	0.13	**0.00**	0.93	0.85
	SecFFT	**1.00**	**1.00**	**0.00**	**0.00**	**1.00**	**1.00**
Size	Foolsgold	0.96	0.90	**0.00**	0.10	0.95	0.92
	FLTrust	**1.00**	**1.00**	**0.00**	**0.00**	**1.00**	**1.00**
	Flame	0.92	**1.00**	0.13	**0.00**	0.93	0.85
	SecFFT	**1.00**	**1.00**	**0.00**	**0.00**	**1.00**	**1.00**
Angle	Foolsgold	0.96	0.90	**0.00**	0.10	0.95	0.92
	FLTrust	0.94	0.85	**0.00**	0.15	0.93	0.88
	Flame	0.92	**1.00**	0.13	**0.00**	0.93	0.85
	SecFFT	**1.00**	**1.00**	**0.00**	**0.00**	**1.00**	**1.00**

"Acc." stands for Accuracy, "Rec." for Recall, "FPR" for False Positive Rate, "FNR" for False Negative Rate, "AUC" for Area Under Curve, and "MCC" for Matthews Correlation Coefficient. The strategies are abbreviated as "None" (EdgeCase without other strategies), "Size" (Size-limited strategy), and "Angle" (Angle-limited strategy).

- 在EdgeCase（None）攻击场景下：
  - SecFFT：准确率与召回率均达到1.00，FPR为 0.00
  - baseline：FoolsGold的FNR较高，为0.10，容易漏报恶意节点；Flame的准确率较低，仅为0.92
- 在Size或Angle约束下，对比方法性能下降显著
- 结论：SecFFT在三种攻击场景中的各项分类指标上均表现出绝对优势，验证了其高鲁棒性

Foundation Model-Based Federated Learning for Intrusion Detection in Drone-Aided Industrial IoT

算法原理 TIPO

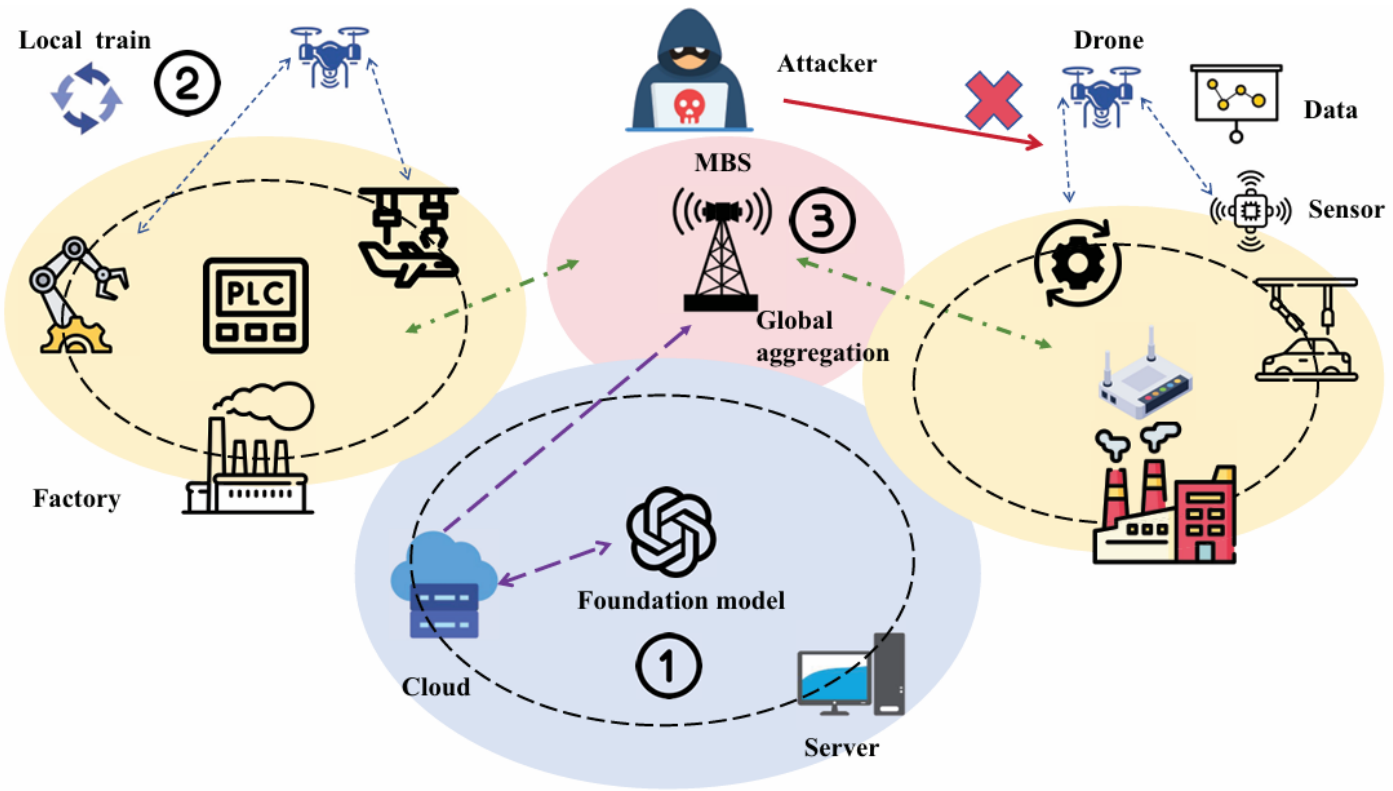
T	目标	在资源受限的无人机辅助IIoT网络中，实现高效、隐私保护的入侵检测
I	输入	无人机网络-物理层数据集（42,263条网络攻击+12,521条物理攻击） 基础模型（RoBERTa）生成的软标签
P	处理	1.数据处理：CGAN数据增强，结合基础模型软标签生成（知识蒸馏） 2.本地更新：利用FedALA自适应初始化本地模型参数，结合FedProx近端正则化项限制本地模型更新幅度 3.全局聚合：采用高斯差分隐私对本地更新梯度加噪，实现全局模型安全聚合
O	输出	1个全局入侵检测模型w^G

P	问题	1.无人机网络样本数据稀缺且存在数据不平衡，传统FL方法泛化差 2.边缘算力有限，对FL方法通信开销要求较高 3.传统FL方法仍存在隐私泄露风险
C	条件	边缘节点（无人机）的计算与存储资源受限
D	难点	Non-IID数据导致的模型偏移；计算效率与隐私保护难以平衡
L	水平	IEEE Internet of Things Journal 2025（SCI一区）

算法原理 创新点

- **FL-IDS**
  - 针对数据稀缺与类不平衡问题，在数据处理阶段利用CGAN生成合成攻击样本。同时利用云端基础模型生成软标签，通过知识蒸馏指导端侧学生模型训练过程，在提高入侵检测性能同时减少算力需求。
  - 在本地训练阶段，结合自适应本地聚合（FedALA）+近端正则化项（FedProx），从而抑制Non-IID导致的模型漂移
  - 在梯度上传阶段引入本地差分隐私，通过高斯噪声扰动在保证入侵检测效率的同时保障数据安全

算法原理 FL-IDS整体架构



- 三层架构
  - 云端（Cloud）
  - 边缘服务器（MBS）
  - 无人机节点（端）
- 三阶段
  - 软标签生成
  - 本地知识蒸馏：本地数据+软标签优化参数
  - 全局安全聚合

算法原理 数据处理

- 数据预处理与数据增强
  - 数据预处理
    - 数据清洗

      平均值填充缺失值：$Missing\ Value = \frac{1}{n}\sum_{i=1}^{n} x_i$

      数据标准化：$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)}$

    - 特征工程

      递归特征消除（RFE）：筛选关键特征

      主成分分析（PCA)：降维

  - CGAN数据增强
    - 扩充少数类，平衡数据集分布
    - $G(z, y)$：输入噪声z和条件变量（标签）y，生成x'；$D(x, y)$: 区分真实样本与合成样本
    - 优化目标：$\min_G \max_D \{\mathbb{E}_{x\sim p_{data}(x)}[log\ D(x, y)] + \mathbb{E}_{z\sim p_z(z)}[log\ (1 - D(G(z, y), y))]\}$

Algorithm 1 Foundation Model-based Dataset Processing

Input: Raw samples X, Post-processing sample number N, Sample ratio r, Normalization weight w_n, RFE feature number n_{rfe}, PCA component number n_{pca}, SMOTE ratio S_r, Generator G, Discriminator D.
Output: Binary-class samples with soft labels X'.
Raw data processing and augmentation:
  $X_1 \leftarrow \text{preprocess}(X)$
  $X_2 \leftarrow \text{Recursive Feature Elimination}(X_1, n_{\text{rfe}})$
  $X_3 \leftarrow \text{SMOTE}(\text{PCA}(X_2, n_{\text{pca}}), S_r)$
  $X_4 \leftarrow \text{Conditional-GAN}(X_3, G, D, r, N)$
Pre-trained foundation model assistance for generating soft labels:
  $Y^s \leftarrow \text{softmax}(\text{RoBERTa}(X_4, \text{prompt}))$
  $\text{avg}_0 \leftarrow \frac{1}{N}\sum_{i=1}^{N} Y^s_{i,0}$
  $\text{avg}_1 \leftarrow \frac{1}{N}\sum_{i=1}^{N} Y^s_{i,1}$
  $\gamma_{\text{smooth},i} \leftarrow (\text{avg}_0 + \text{avg}_1 - Y^s_{i,0} - Y^s_{i,1}), \forall i \in \{1, 2, \ldots, N\}$
  $\Gamma_{\text{smooth}} \leftarrow \{\gamma_{\text{smooth}}\}$
  $y_{\text{soft,i}} \leftarrow y_i - w_n \cdot \text{normalize}(\gamma_{\text{smooth,i}}, \Gamma_{\text{smooth}})$
  return $X' = (X_4, y_{\text{soft}})$

数据预处理与数据增强

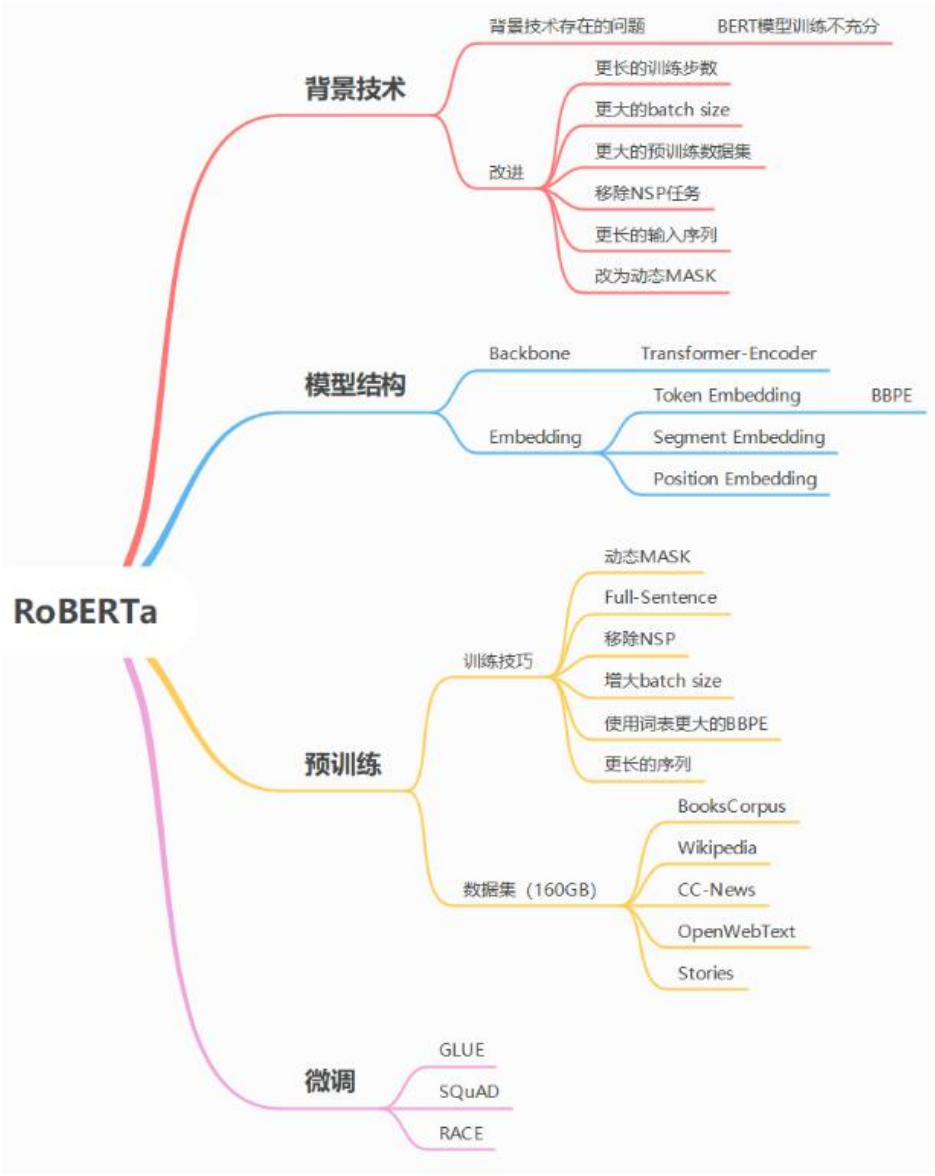
软标签生成与平滑更新

算法原理 数据处理

- 软标签生成
  - 基础模型
    - 利用预训练RoBERTa模型作为“教师”,对处理后的网络流量数据进行无监督特征提取与分类
    - RoBERTa与BERT相比，模型结构不变，训练技巧改变以更好适配下游任务
    - 输入：流量特征文本化描述x，模型参数θ；输出：特征向量$\varphi(x)$
    - $\varphi(x) = RoBERTa(x;\ \theta)$
  - 软标签计算
    - 计算基础模型的预测概率分布：

$$P_r(y|x) = softmax(W \cdot \varphi(x) + b)$$

    - 软标签相比One-hot硬标签，类别间的相关性信息更丰富，能更好地指导学生模型训练



算法原理 联邦入侵检测

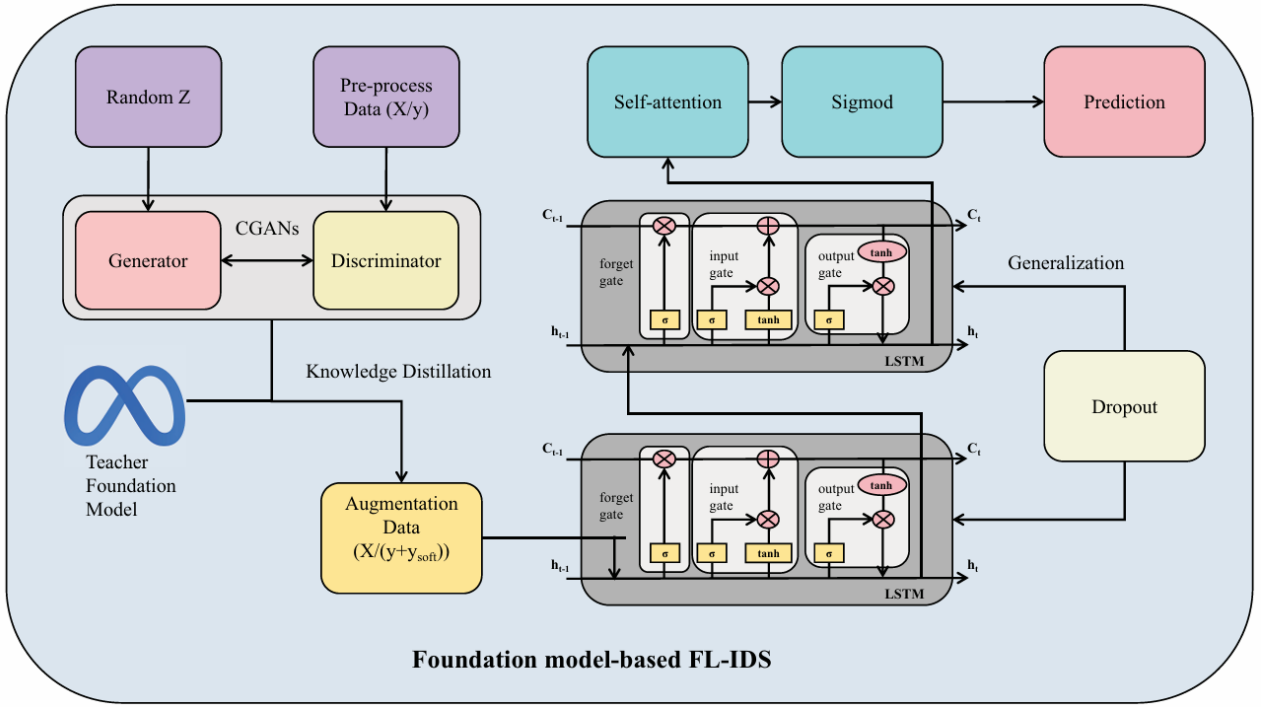
- 本地学生模型与知识蒸馏
  - 学生模型架构
    - 双层LSTM：捕获网络流量数据的时序依赖特征
    - 自注意力：动态加权关键时间步特征
    - 全连接层：分类
  - 软标签更新与自蒸馏
    - 结合当前预测与上一轮软标签更新软标签

$$soft \leftarrow (1-\delta)\cdot soft + \delta \cdot softmax(predict)$$

    - 软标签损失函数：

$$l_s(predict, soft) = \mu_s \cdot softmax(predict, soft)$$

    - 总损失：分类损失+软标签损失+近端项损失



算法原理 联邦入侵检测

- 本地模型更新
  - 自适应本地聚合（ALA）
    - 冻结低层（通用）参数，仅对高层（个性化）参数进行自适应聚合，加快本地模型收敛速度
    - 低层（$i = 1, 2, \ldots, L-p$）：$w_i = w_i^G$
    - 高层（$i = L-p+1, \ldots, L$）：$w_i^T = w_i + (w_i^G - w_i) \odot W_p$

$$W_p \leftarrow W_p - \eta \cdot \nabla_{W_p} \mathcal{L}\left(w_i^T, D_i\right)$$

  - 近端正则化
    - 本地训练过程中，在本地损失函数中引入近端项，限制本地模型更新严重偏离全局模型

$$l_p(w, w^G) = \frac{\mu_p}{2} ||w - w^G||^2$$

Algorithm 2 Federated Learning in Drone Networks

Input: Number of rounds R, Number of clients N, Number of local epochs E, Dataset X, Proximal coefficient μ, weight of soft loss w_{soft},Weight of new soft w_u.
Output: Global model parameters W^G.
Client[i]:
Localtrain$(i, {W_r}^G, X[i])$:
Initialize parameters W_r with FedALA (Freeze higher-level parameters partially and skip the first round)
Extract features $x[i]$, soft labels y_{soft}, and labels y
for each epoch from 1 to E **do**
    Generate predicted labels: $\hat{y} = net(x[i])$
    Compute loss_c: $L_c \leftarrow L(y, \hat{y})$
    Compute loss_s: $L_s \leftarrow L(y, y_{\text{soft}})$
    Compute loss_p: $L_p \leftarrow \frac{1}{2}\sum(w_r - w_r^{\text{G}})^2$
    Compute loss: $L \leftarrow L_{\text{c}} + w_{\text{s}} \cdot L_{\text{s}} + \mu \cdot L_{\text{p}}$
    Backward and compute gradients
    Clip gradients and add Gaussian noise to gradients
    Update model parameters
    Update soft: $y_{\text{soft}} \leftarrow (1 - w_u) \cdot y_{\text{soft}} + w_u \cdot \text{softmax}(\hat{y})$
end for
send ${W_r}^i$ to the server
Server:
Initialize W_0^G
for each round r from 0 to $R-1$ **do**
    $c_r \leftarrow$ Choose clients of N
    for each client $i \in c_r$ **do**
        $W_r^i \leftarrow \text{Localtrain}(i, W_r^G, X[i])$
    end for
    $W_{r+1}^G \leftarrow \frac{1}{N}\sum_{i=1}^{N} W_r^i$
end for

算法原理 联邦入侵检测

- 全局安全聚合
  - 攻击者可能通过分析各个本地模型上传的梯度更新重构原始数据
  - 引入基于高斯噪声的差分隐私
    - 在无人机节点将梯度上传到服务器之前，对本地梯度加噪：

$$\nabla\theta' = \nabla\theta + \mathcal{N}(0, \sigma^2)$$

    - σ满足(ϵ, δ)差分隐私条件：

$$\sigma \geq \frac{\Delta f \cdot \sqrt{2\,ln\left(\frac{1.25}{\delta}\right)}}{\epsilon}$$

Δf是梯度的L2灵敏度，$\Delta f = \max_i(threshold, max(|\nabla\theta_i|))$，防止个别异常大的梯度导致敏感度计算过高，从而被迫添加过大的噪声，破坏模型可用性

    - 通过调节σ中的ϵ和δ，实现隐私保护与入侵检测准确率的平衡

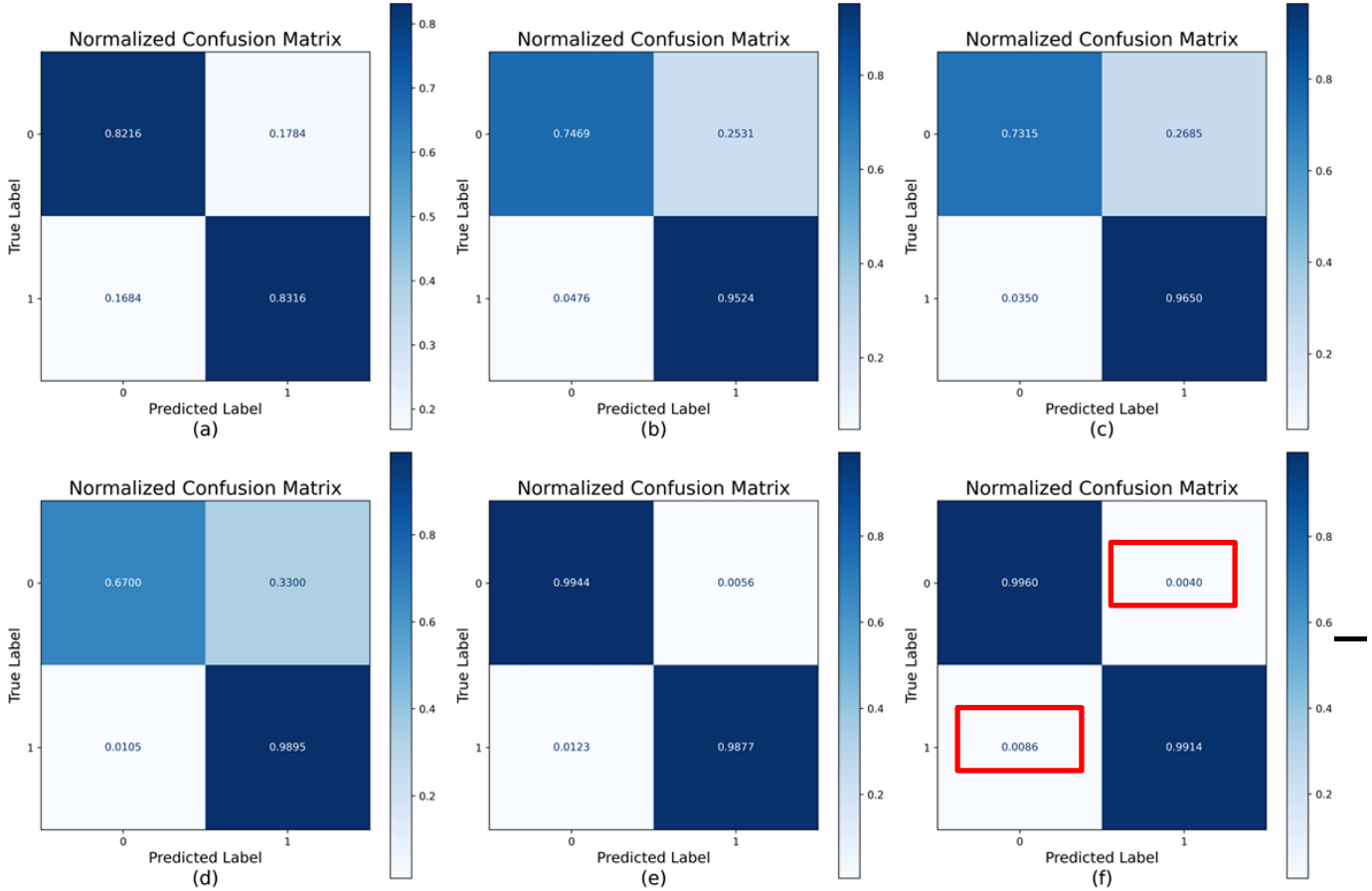
实验流程 实验设置

- 数据集：cyber-physical datasets
  - 源自配备无人机、控制器和数据收集工具的测试台
  - 42,263条网络攻击实例，12,521条物理攻击实例
  - 正常+四种攻击类型
  - 60%训练，20%验证，20%测试
- 评价指标
  - Accuracy $ACC = \frac{TP + TN}{TP + TN + FP + FN}.$
  - Recall：$RECALL = \frac{TP}{TP + FN}.$
  - Precision：$PRECISION = \frac{TP}{TP + FP}.$
  - F1-Score：$F1 = 2 \cdot \frac{PRECISION \cdot RECALL}{PRECISION + RECALL}.$
  - 利用LoRA进行高效联邦微调
- 对比基线：SVM、DT、RF、CNN、FNN

Cyber Features	Cyber Features	Physical Features
timestamp_c	tcp.srcport	timestamp_p
frame.number	tcp.dstport	mid
frame.len	tcp.seq_raw	x,y,z
frame.protocols	tcp.ack_raw	vgx,vgy,vgz
wlan.duration	tcp.hdr_len	templ
wlan.ra	tcp.flags	temph
wlan.ta	tcp.window_size	tof
wlan.da	tcp.options	h
wlan.sa	udp.srcport	bat
wlan.bssid	udp.dstport	baro
wlan.frag	udp.length	time
wlan.seq	data.data	agx,agy,agz
ip.hdr_len	data.len	mpitch
ip.len	wlan.fc.type	mroll
ip.id	wlan.fc.subtype	myaw
ip.flags	time_since_last_packet	est_x,est_y
ip.ttl	ip.src	cntl_x,cntl_y
ip.proto	ip.dst	residual

实验结果 FL-IDS整体性能

Dataset	TN	FP	FN	TP	Accuracy(%)	Recall(%)	Precision(%)	F1 Scores(%)
Cyber-raw	5366	1165	1112	5491	82.66	83.16	82.50	82.83
Cyber-CGANs	2134	723	456	9122	90.52	95.24	**92.65**	93.93
Cyber-CGANs&Foundation model-based	2090	767	335	9243	**91.14**	**96.50**	92.33	**94.37**
Physical-raw	595	293	17	1606	87.65	98.95	84.57	91.20
Physical-CGANs	1249	7	30	2409	98.99	98.77	99.71	99.24
Physical-CGANs&Foundation model-based	1251	5	21	2418	**99.30**	**99.14**	**99.80**	**99.47**

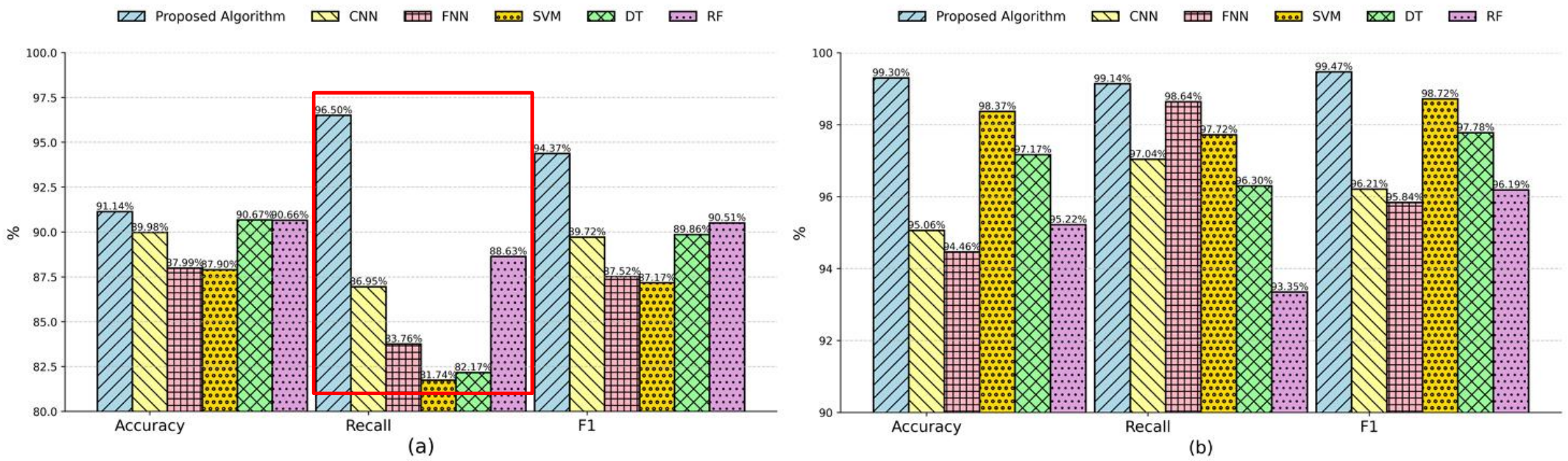


结果分析：

- FL-IDS在Cyber数据集上准确率91.14%，Physical数据集上99.30%
- 与未引入基础模型和CGAN（raw）相比，准确率分别提升了8.48%和11.65%
- 在网络数据集上与仅引入CGAN（Cyber-CGANs）相比，引入基础模型后，准确率提升了0.62%
- 在区分正常和攻击流量时误报率和漏报率很低，尤其是物理攻击

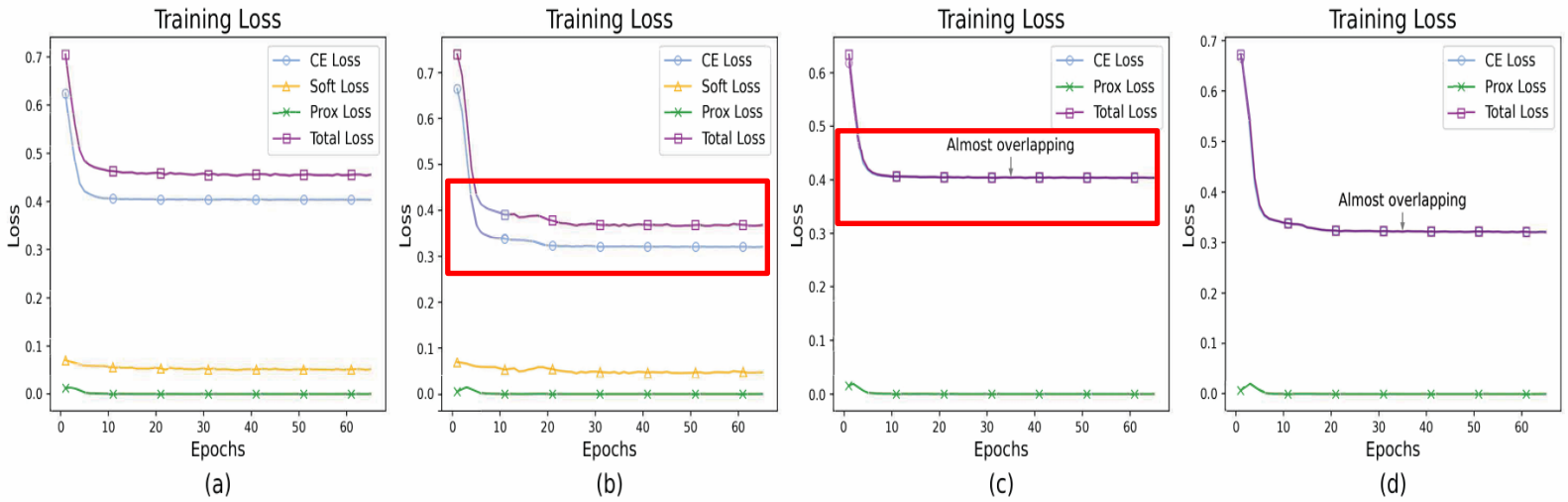
结论：FL-IDS高准确率，性能提升显著

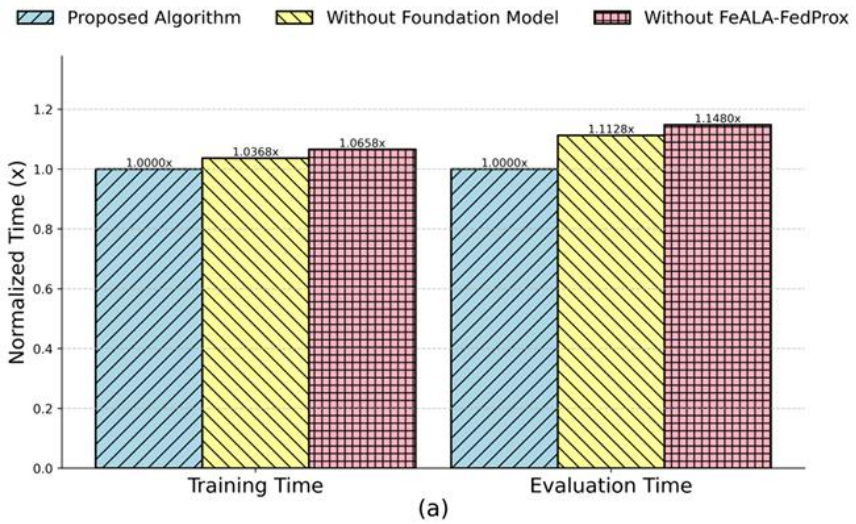
实验结果 对比实验

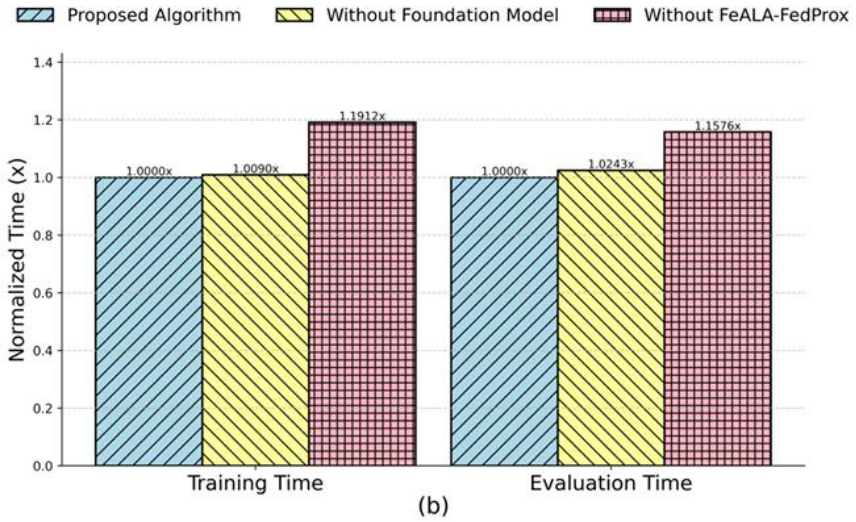


- FL-IDS：准确率、召回率、F1分数在网络和物理数据集上均高于所有对比算法
- 在网络数据集上召回率远高于其他算法，说明极少漏报攻击行为
- LSTM的时序特征捕获能力较强

实验结果 收敛性与时间开销

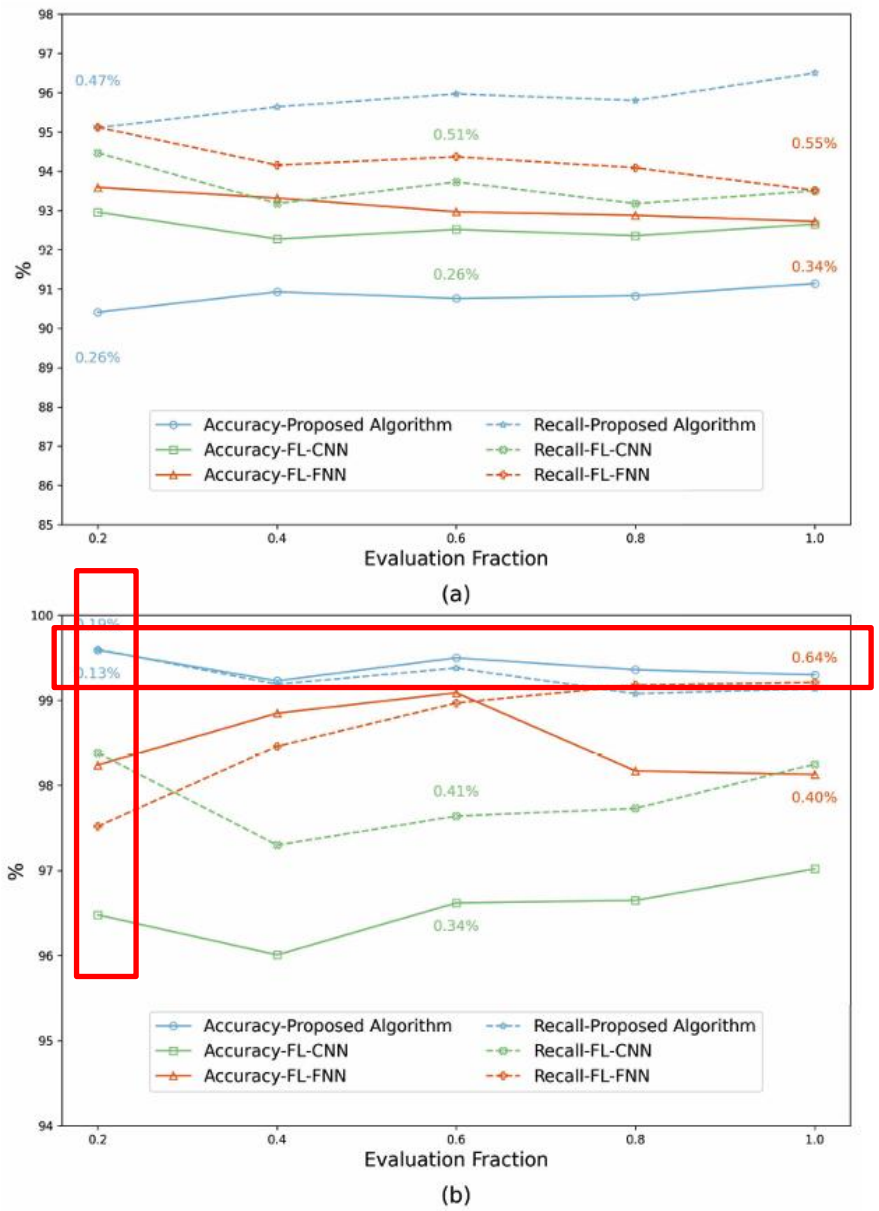






- 收敛性
  - 引入基础模型后，loss下降曲线更陡，收敛速度明显加快
  - 在联邦学习设置下，也能在较少轮次内达到稳定状态
- 时间开销
  - 虽引入了额外的计算模块，但由于加速了收敛，整体训练评估时间没有显著增加，满足无人机网络实时性要求

实验结果 鲁棒性分析



- 不同参与率下的性能稳定性
  - 场景设置
    - 通过改变每轮参与联邦聚合的节点比例模拟无人机网络节点离线或网络波动
  - 结果分析
    - 稳定性强：参与者比例0.2的情况下，FL-IDS的准确率和召回率依旧整体高于对比方法（Physical数据集上更明显）
    - 方差低：不同参与者比例设置下，FL-IDS波动最小，证明了其在应对异构与动态网络环境时的鲁棒性

特点总结与未来展望

特点总结与未来展望

- 算法创新
  - SecFFT：首次提出频域分布与长期意图联合检测机制，有效防御联邦微调中的隐蔽、多轮次后门攻击
  - FL-IDS：构建了基础模型驱动的联邦入侵检测框架，创新性地融合了CGAN数据增强与知识蒸馏技术，解决了边缘网络中数据稀缺与算力受限的矛盾
- 算法优势
  - SecFFT：在多种攻击场景下将ASR降低至接近0，且未损耗全局模型性能
  - FL-IDS：在网络和物理数据集上分别实现了91%和99%的高准确率，且大幅降低了边缘设备的资源消耗
- 未来展望
  - 异构性适应：在极端异构（数据/设备）场景下的鲁棒性防御策略
  - 轻量化部署：探索更高效的模型压缩方法，满足IoRT场景对实时性与低功耗要求

参考文献

[1] Zhou Z, Xu C, Wang B, et al. SecFFT: Safeguarding Federated Fine-Tuning for Large Vision Language Models against Covert Backdoor Attacks in IoRT Networks[J]. IEEE Internet of Things Journal, 2024.

[2] Jiao S, Wang J, Tong Z, et al. Foundation Model-Based Federated Learning for Intrusion Detection in Drone-Aided Industrial IoT[J]. IEEE Internet of Things Journal, 2025.

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！