西安电子科技大学
XIDIAN UNIVERSITY

博士学位论文答辩

面向深度神经网络的对抗鲁棒性关键技术研究

Research on Key Techniques for Adversarial Robustness of Deep Neural Networks

答 辩 人： 黄梦蝶

导　　师： 陈晓峰 教授

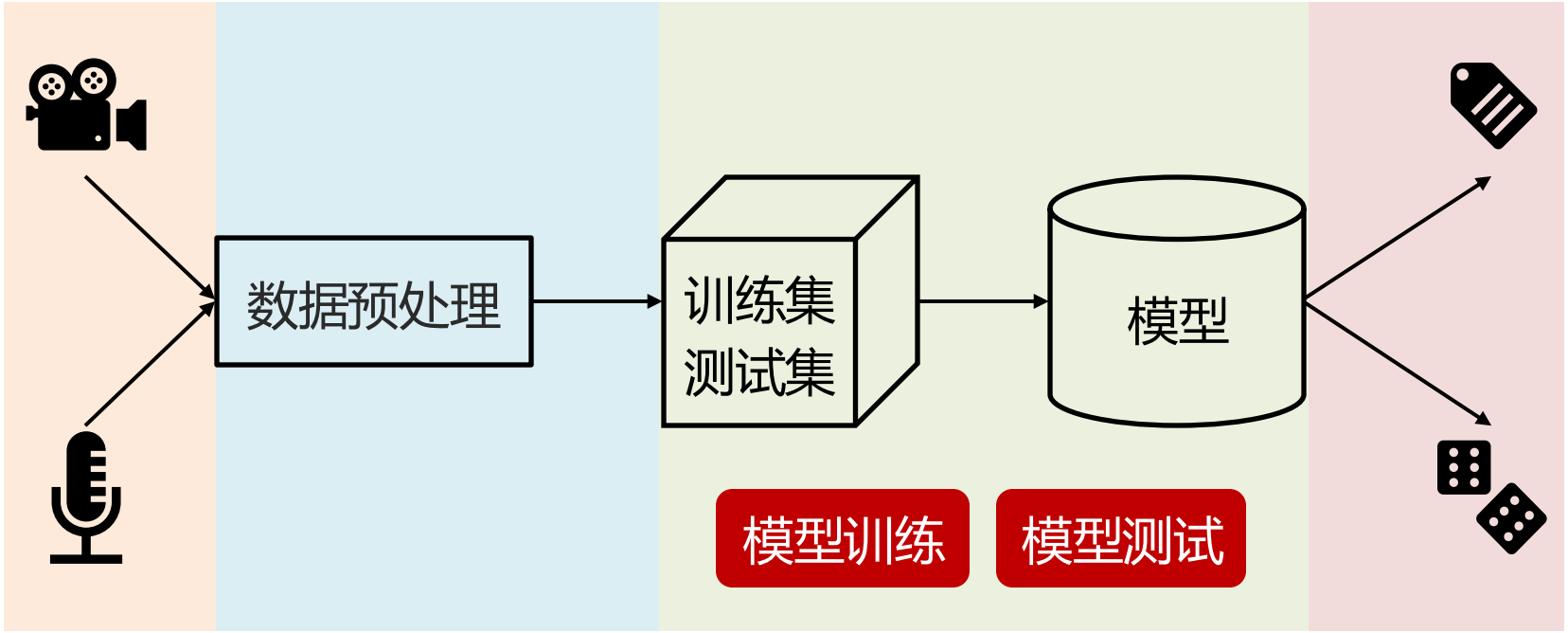
答辩时间： 2025年 5月15日

目 录

目 录

人工智能系统工作流程

人工智能系统=数据+模型



绪　论

《新一代人工智能治理原则》

—发展负责任的人工智能

- 第五项 安全可控原则
- 人工智能系统需求

属性	功能
透明性	可审核
可解释性	可监督
可靠性	可追溯
可控性	可信赖

- 人工智能技术突破
  - ✓ 提高人工智能鲁棒性及抗干扰性
  - ✓ 形成人工智能安全评估管控能力

中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

人工智能安全风险

- 模型可靠性威胁
  - ➢ 训练阶段 数据集污染
  - ➢ 推断阶段 对抗样本干扰

投毒攻击

逃逸攻击

对抗攻击实际危害

- 计算机视觉系统
  - ✓ 路标识别系统 →自动驾驶事故
  - ✓ 人脸识别系统 →人员非法进入
- 语音识别系统
- 自然语言处理系统

研究面向深度神经网络的对抗鲁棒性关键技术具有重要的理论价值和实际意义

《新一代人工智能治理原则》

—发展负责任的人工智能

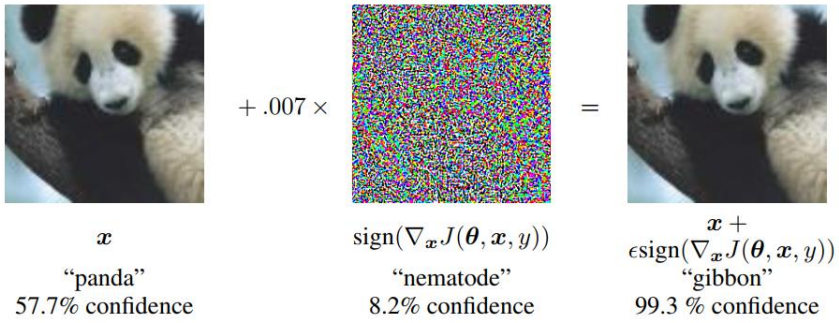
- 第五项 安全可控原则
- 人工智能系统需求

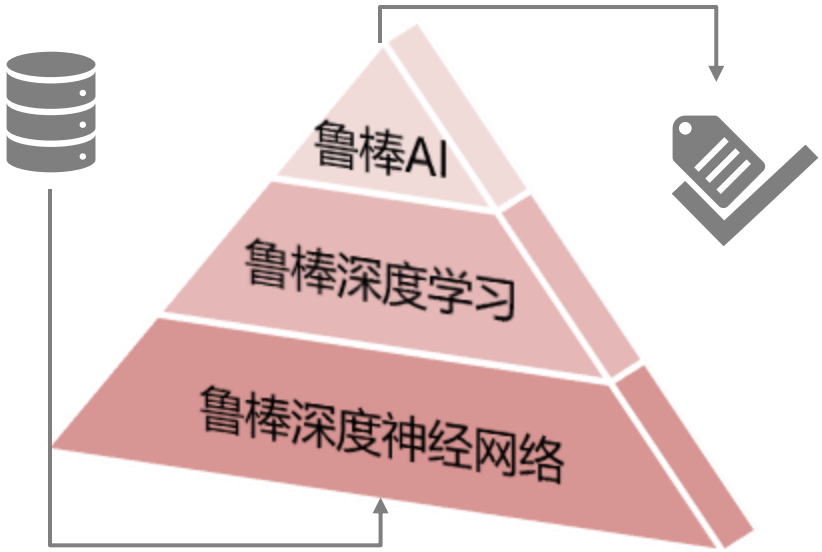
属性	功能
透明性	可审核
可解释性	可监督
可靠性	可追溯
可控性	可信赖

- 人工智能技术突破
- ✓ 提高人工智能鲁棒性及抗干扰性
- ✓ 形成人工智能安全评估管控能力

中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

核心威胁对抗样本

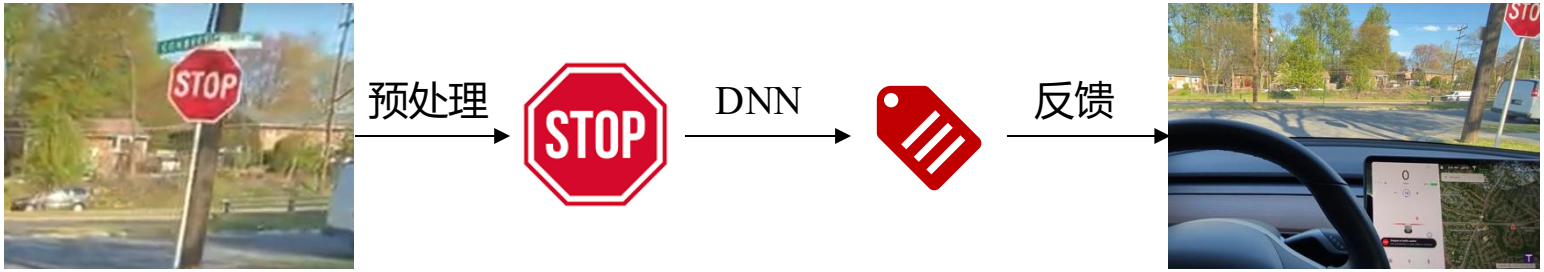




研究面向深度神经网络的对抗鲁棒性关键技术具有重要的理论价值和实际意义

□ 对抗攻击 (Adversarial Attack)

- 攻击对象：深度学习模型的推理阶段
- 攻击目标：使深度学习模型对输入的对抗样本做出错误预测。
- 现实威胁：图像分类模型、网络流量分类模型等。

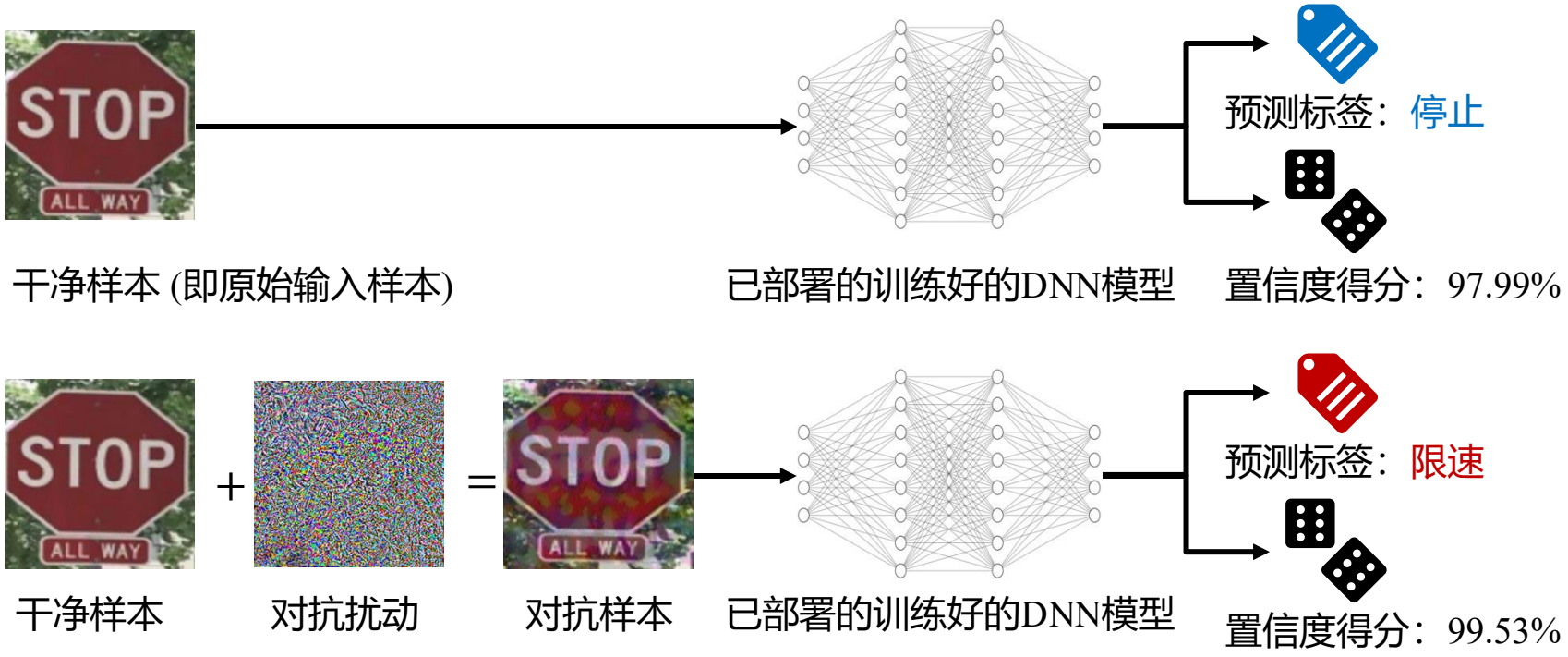


自动驾驶仪动作：停止

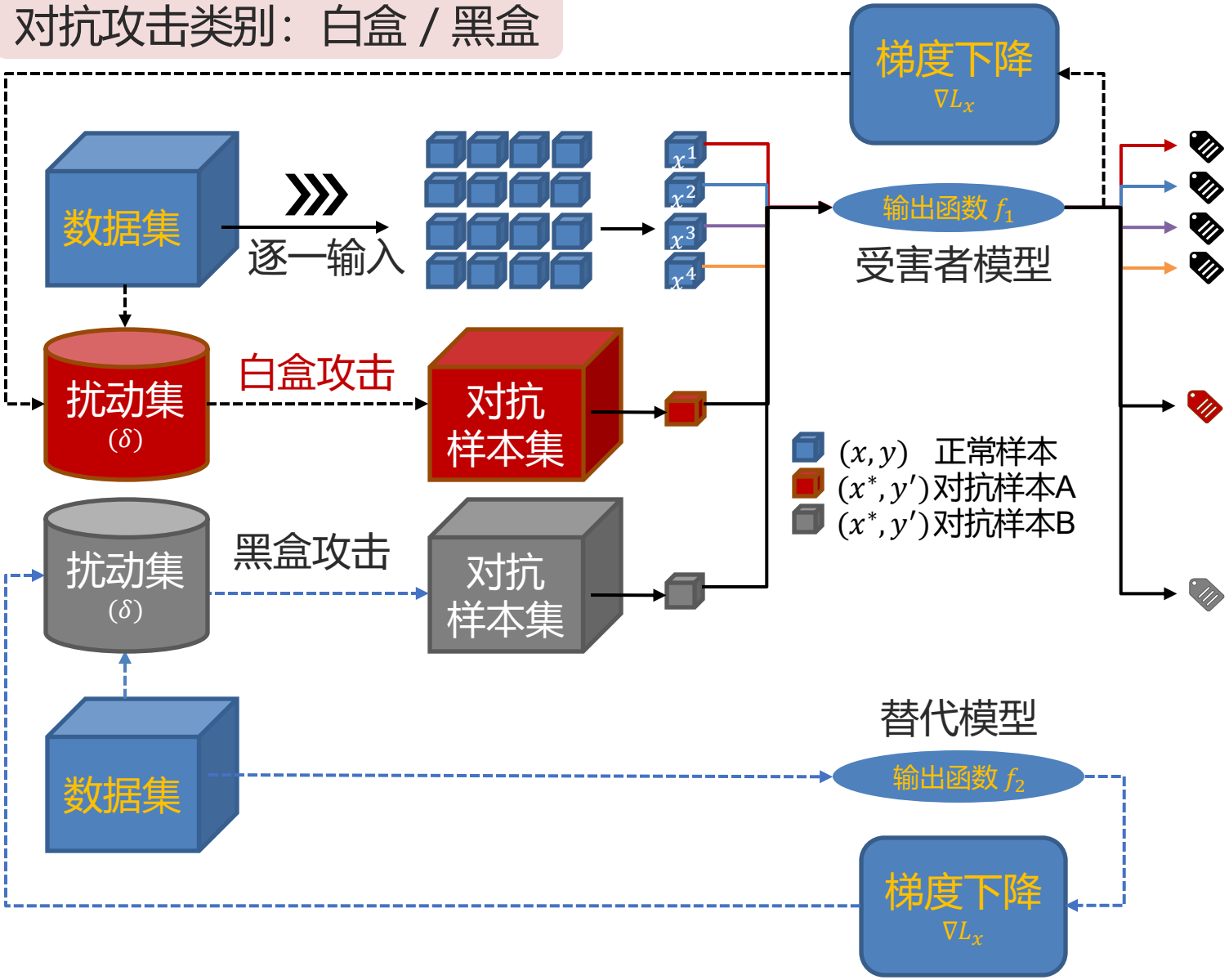
自动驾驶仪动作：限速

对抗攻击 (Adversarial Attack)

- 攻击对象：深度学习模型的推理阶段
- 攻击目标：使深度学习模型对输入的对抗样本做出错误预测。
- 核心思路：通过添加精心制作且难以察觉的扰动改变输入，使分类错误。

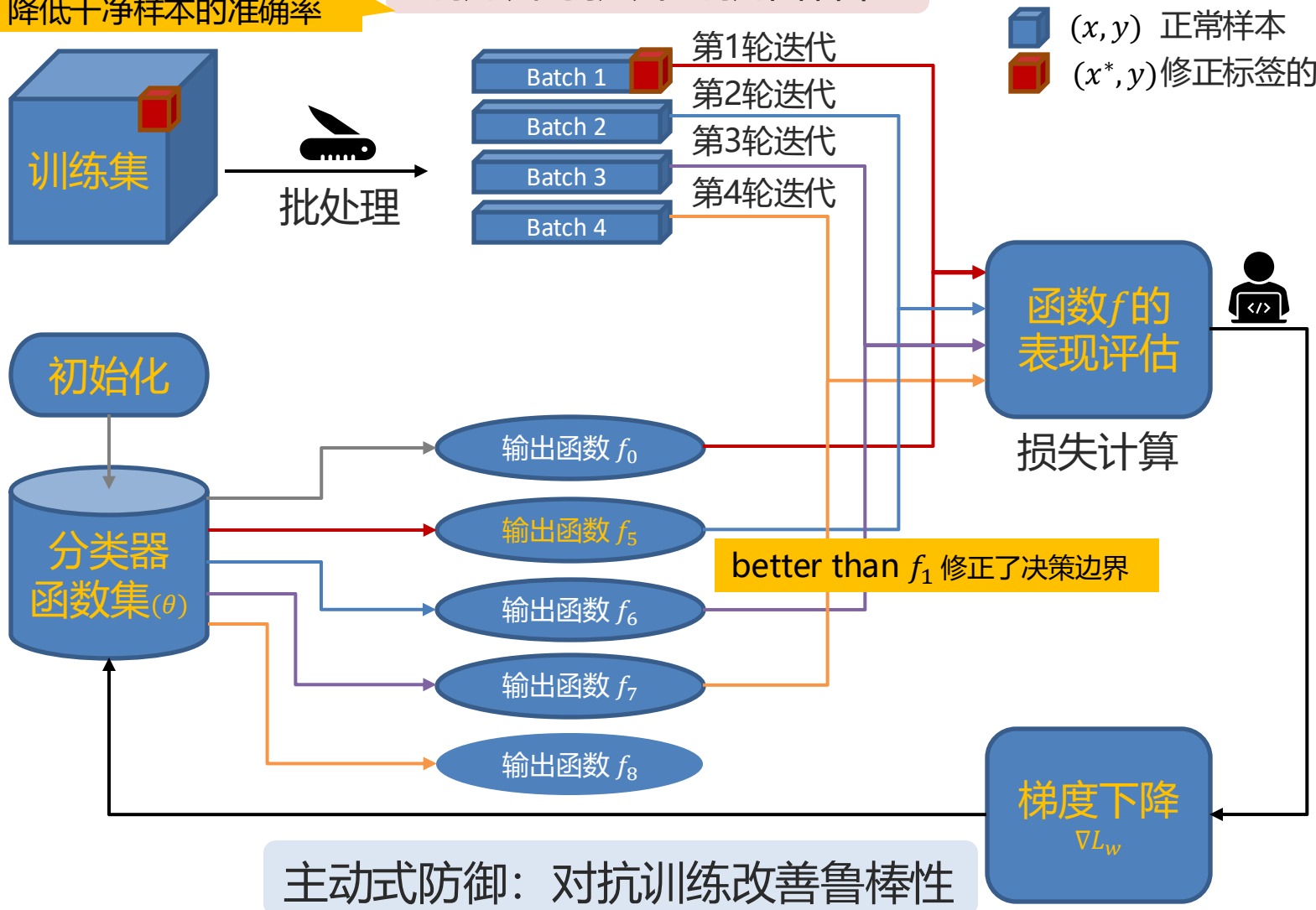


对抗攻击类别：白盒 / 黑盒


梯度下降
∇L_x
数据集
逐一输入
x^1
x^2
x^3
x^4
输出函数 f_1
受害者模型
扰动集
(δ)
白盒攻击
对抗
样本集
(x, y) 正常样本
(x^*, y')对抗样本A
(x^*, y')对抗样本B
扰动集
(δ)
黑盒攻击
对抗
样本集
替代模型
数据集
输出函数 f_2
梯度下降
∇L_x

缺点1: 过拟合对抗扰动特征
缺点2: 降低干净样本的准确率

对抗训练提升对抗鲁棒性



主动式防御：对抗训练改善鲁棒性

绪　论

对抗样本多样性	鲁棒性形式化保证	鲁棒训练成本
不可预见对抗攻击的鲁棒防御问题	模型鲁棒评估结果的可靠性问题	模型鲁棒增强训练的昂贵开销问题

面向深度神经网络的对抗鲁棒性关键技术研究关键挑战

模型鲁棒性泛化技术	模型鲁棒性验证技术	模型鲁棒性迁移技术
泛化范围	验证界限	迁移场景
不同敌手能力假设	l_1范数半径	跨数据域迁移
不同深度神经网络	l_2范数半径	跨模型结构迁移
不同对抗攻击类型	l_∞范数半径	跨数据域兼模型迁移

绪　论

潜在表征学习
潜在空间插值
混合样本

多元线性组合
多元遮罩组合
未知样本

同构输入特征
异构输入特征
定义界限

零阶预测信息
一阶梯度信息
鲁棒证明

模型表征蒸馏
多视对比学习
萃取模型

面向深度神经网络的对抗鲁棒性关键技术研究主要贡献

基于潜在表征混合的模型鲁棒性泛化技术

基于多阶随机平滑的模型鲁棒性验证技术

基于对比表征蒸馏的模型鲁棒性迁移技术

泛化范围	验证界限	迁移场景
不同敌手能力假设	l_1范数半径	跨数据域迁移
不同深度神经网络	l_2范数半径	跨模型结构迁移
不同对抗攻击类型	l_∞范数半径	跨数据域兼模型迁移

绪　论

- 本文主要贡献

创新点 1

基于潜在表征混合的对抗鲁棒性泛化技术

- 提出基于多模式流形插值的数据扩增技术
- 提出基于语义混合样本的多目标训练技术

创新点 2

基于多阶随机平滑的对抗鲁棒性验证技术

- 提出基于多阶信息的自适应随机平滑技术
- 提出基于特征敏感性的维度鲁棒半径度量技术

创新点 3

基于对比表征蒸馏的对抗鲁棒性迁移技术

- 提出基于自适应维度对齐的跨域蒸馏技术
- 提出基于双重鲁棒感知的对比迁移学习技术

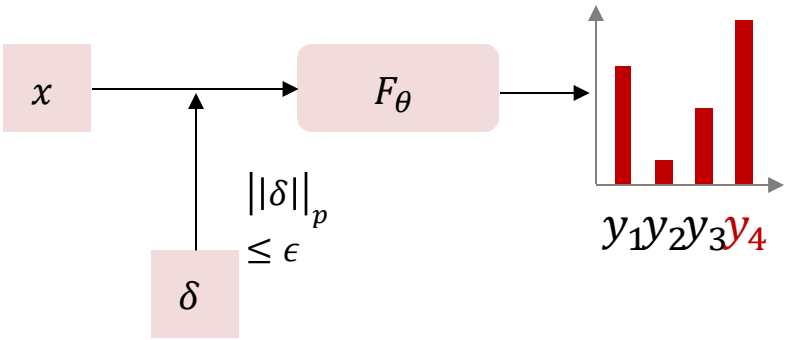
目 录

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

对抗攻击威胁

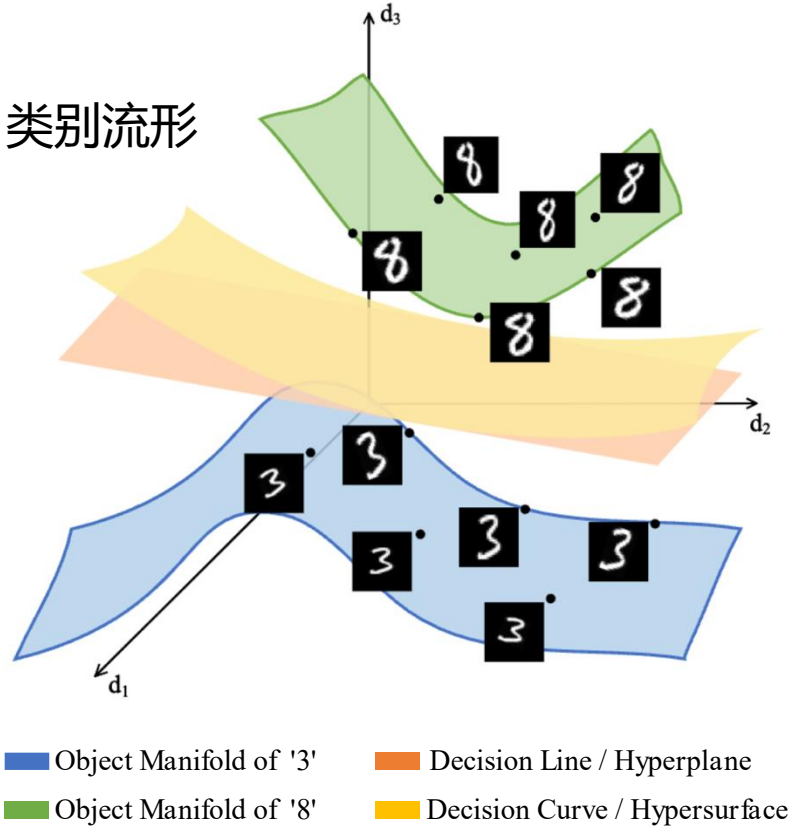
- 流形外(Off-Manifold)对抗攻击
  - 即输入空间对抗攻击
  - 对抗扰动优化目标

$$\max_{||\delta||_p \leq \epsilon} \mathcal{L}(F_\theta(x+\delta), y_{true})$$



  - 代表性对抗攻击算法：FGSM, PGD, JSMA, DeepFool, CW, AutoAttack

类别流形



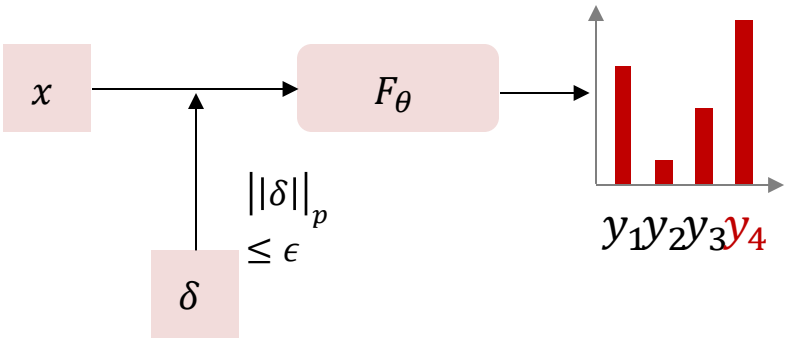
输入空间:
28x28 pixels→728 dimensions

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

对抗攻击威胁

- 流形外(Off-Manifold)对抗攻击
  - 即输入空间对抗攻击
  - 对抗扰动优化目标

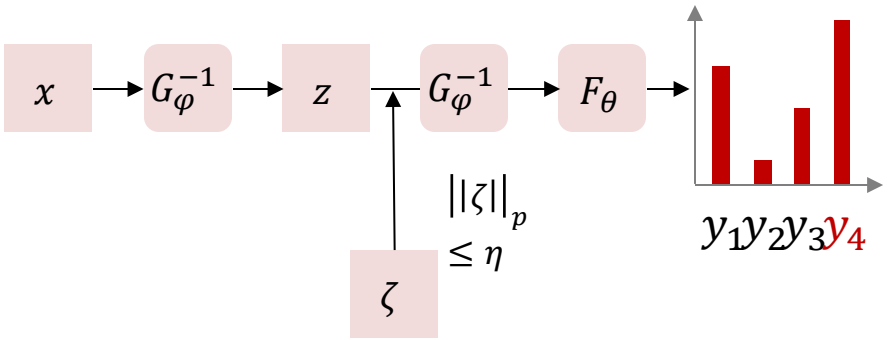
$$\max_{||\delta||_p \le \epsilon} \mathcal{L}(F_\theta(x+\delta), y_{true})$$



  - 代表性对抗攻击算法：FGSM, PGD, JSMA, DeepFool, CW, AutoAttack

- 流形上(On-Manifold)对抗攻击
  - 即潜在空间对抗攻击
  - 对抗扰动优化目标

$$\max_{||\zeta||_p \le \eta} \mathcal{L}(F_\theta(G_\varphi(z+\zeta)), y_{true})$$



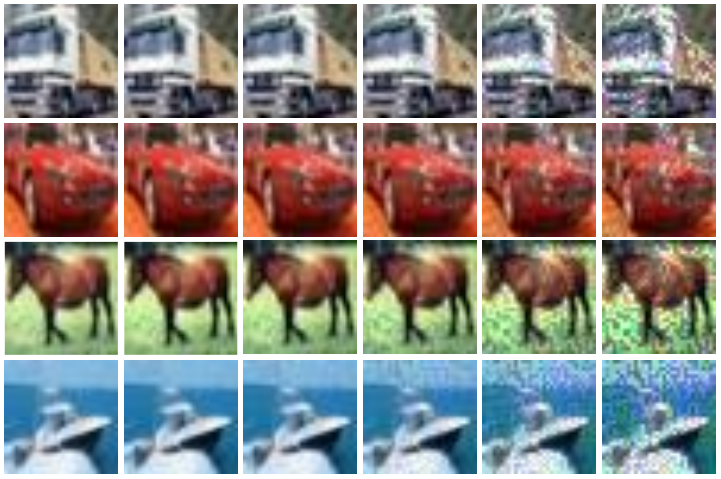
  - 代表性对抗攻击算法：OM-FGSM, OM-PGD

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

对抗攻击威胁

- 流形外(Off-Manifold)对抗攻击
  - 即输入空间对抗攻击
  - 对抗扰动优化目标

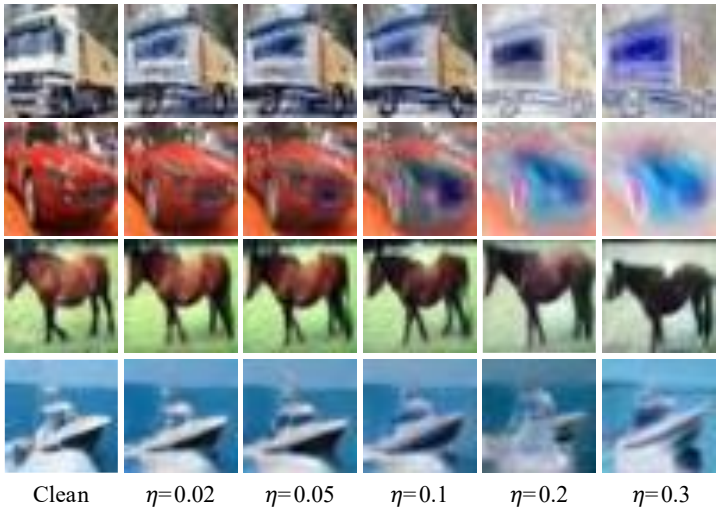
$$\max_{||\delta||_p \leq \epsilon} \mathcal{L}(F_\theta(x+\delta), y_{true})$$



PGD CIFAR-10

- 流形上(On-Manifold)对抗攻击
  - 即潜在空间对抗攻击
  - 对抗扰动优化目标

$$\max_{||\zeta||_p \leq \eta} \mathcal{L}(F_\theta(G_\varphi(z+\zeta)), y_{true})$$



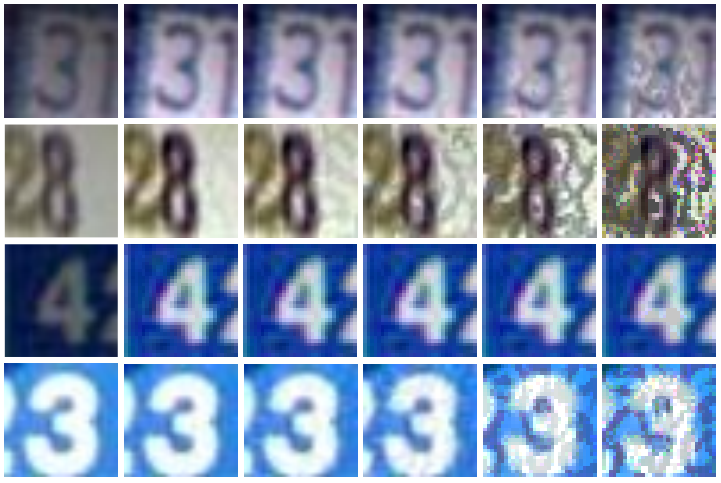
OM-PGD CIFAR-10

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

对抗攻击威胁

- 流形外(Off-Manifold)对抗攻击
  - 即输入空间对抗攻击
  - 对抗扰动优化目标

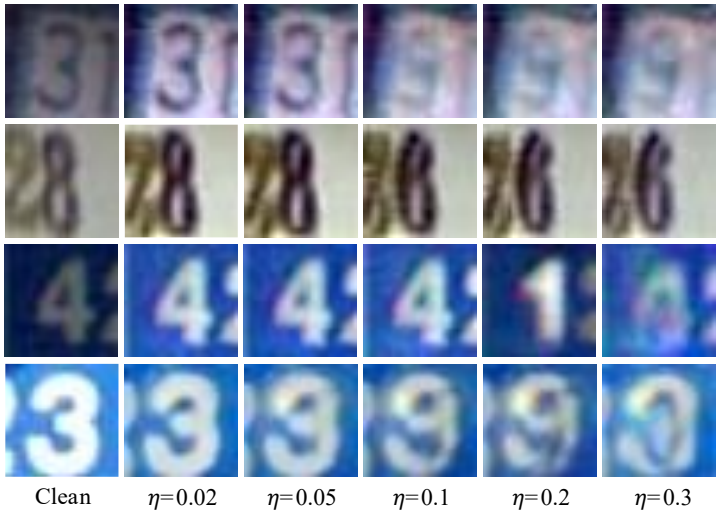
$$\max_{||\delta||_p \le \epsilon} \mathcal{L}(F_\theta(x+\delta), y_{true})$$



PGD SVHN

- 流形上(On-Manifold)对抗攻击
  - 即潜在空间对抗攻击
  - 对抗扰动优化目标

$$\max_{||\zeta||_p \le \eta} \mathcal{L}(F_\theta(G_\varphi(z+\zeta)), y_{true})$$



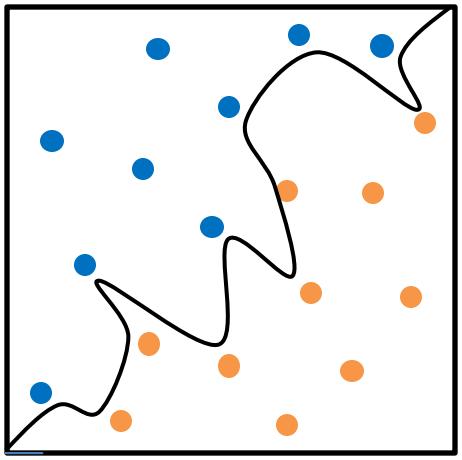
OM-PGD SVHN

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

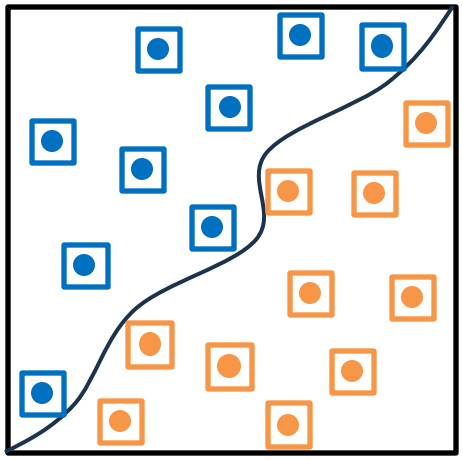
鲁棒训练方法

- 针对流形外(Off-Manifold)对抗攻击
  - 输入空间对抗训练
  - 混合训练
    - 输入空间混合训练
    - 隐层空间混合训练
- 针对流形上(On-Manifold)对抗攻击
  - 潜在空间对抗训练

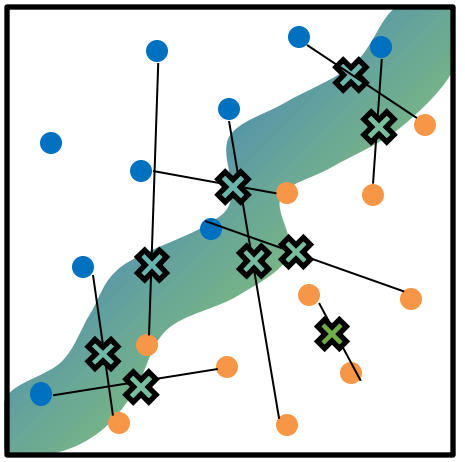
标准训练



对抗训练

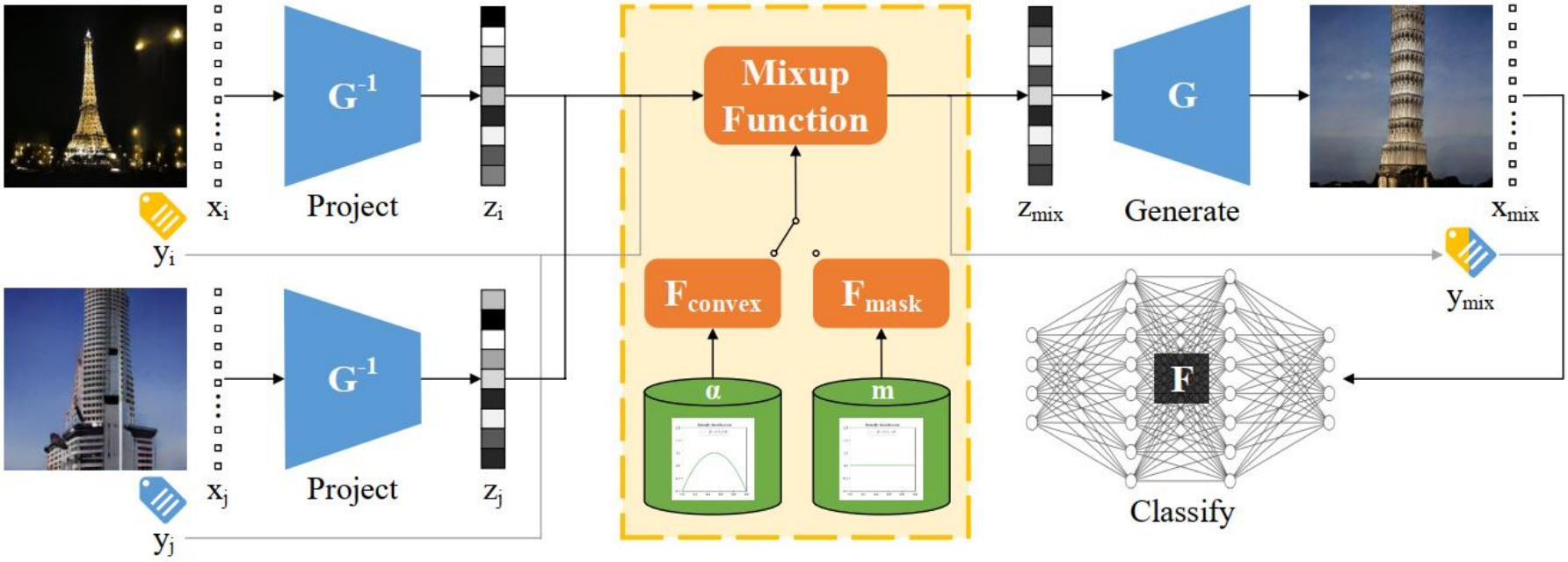


混合训练



基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

□ Latent Representation Mixup (LarepMixup) 框架



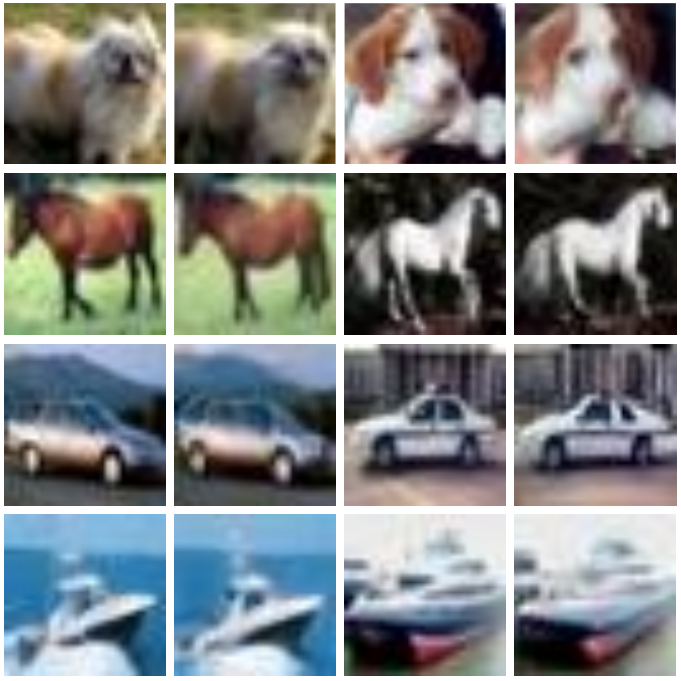
(a) 低维流形嵌入模块　(b) 潜在空间表征混合模块　(c) 语义混合样本多目标训练模块

研究不依赖任何敌手知识的深度神经网络的对抗鲁棒性泛化技术具有重要意义

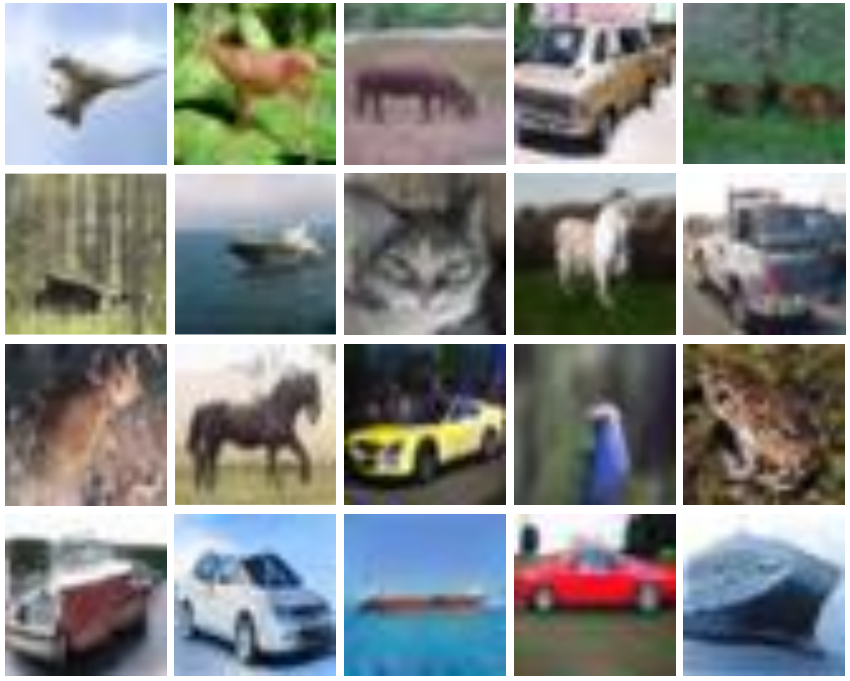
基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

阶段一：低维流形嵌入

- 基于StyleGAN的潜在表征学习
  - 生成模型 $G^{-1}(x, y_{true}) \rightarrow (z, y_{true})$
- 流形上未知数据生成
  - 随机采样潜在表征，合成未知样本



x in D_{train} G(z) x in D_{test} G(z)

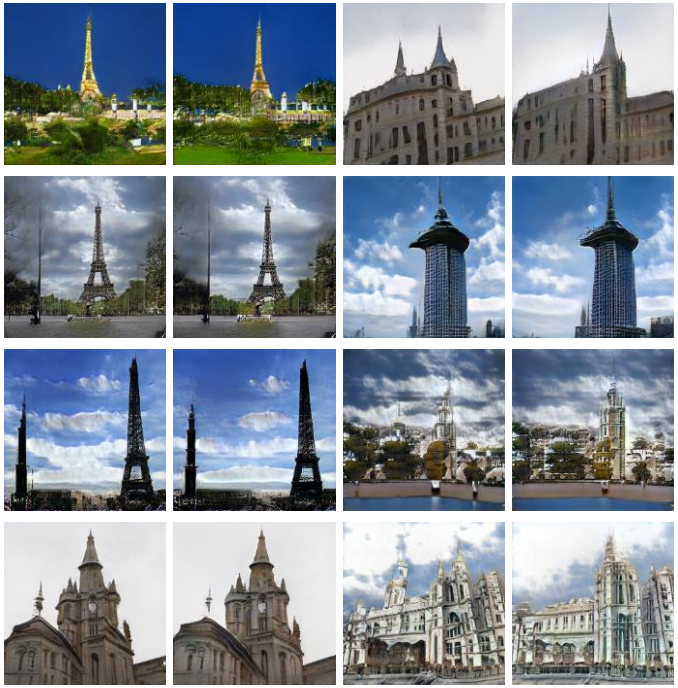


G(z)

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

阶段一：低维流形嵌入

- 基于StyleGAN的潜在表征学习
  - 生成模型 $G^{-1}(x, y_{true}) \rightarrow (z, y_{true})$
- 流形上未知数据生成
  - 随机采样潜在表征，合成未知样本

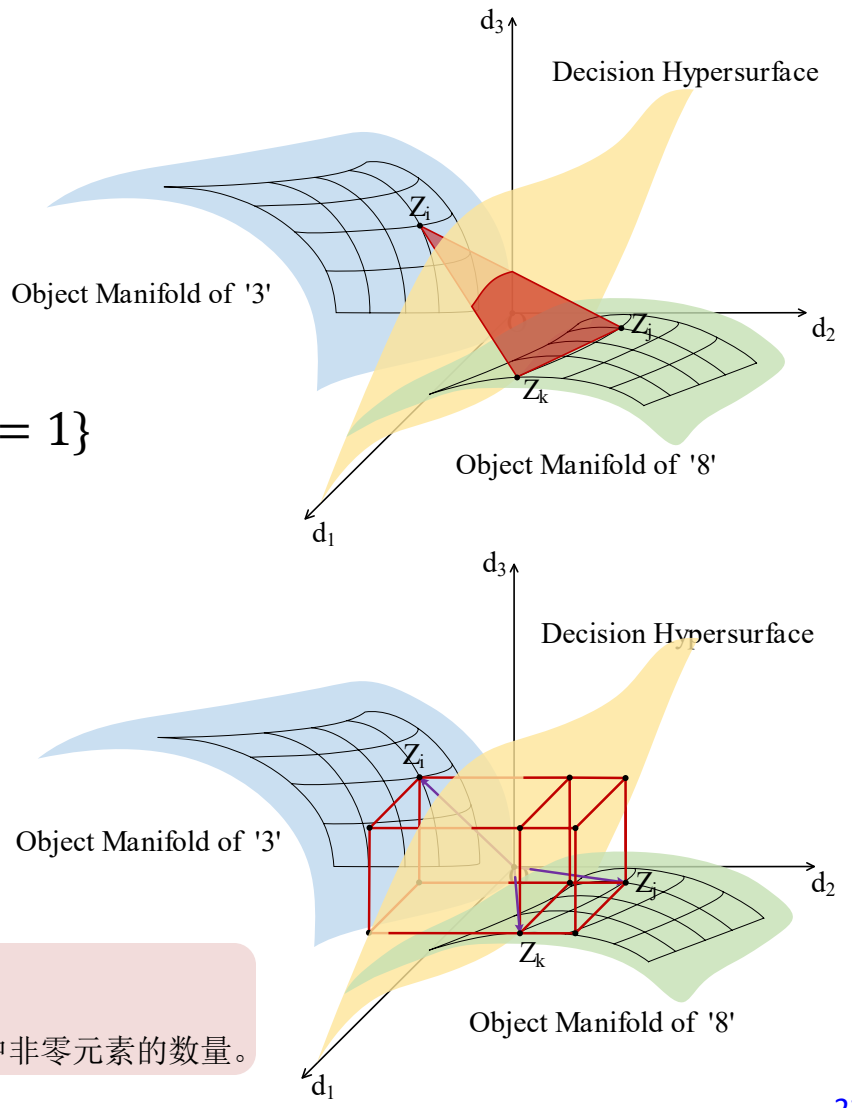


基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

阶段二：多模式流形插值

- 基于线性组合的插值
  - 混合样本 $z_{mix} = \alpha_1 z_1 + \cdots + \alpha_k z_k$
  - 混合标签 $y_{mix} = \alpha_1 y_1 + \cdots + \alpha_k y_k$
  - 插值系数 $\alpha \in A := \{R^k, \alpha_i \in [0,1], \sum_{i=0}^{k} \alpha_i = 1\}$
- 基于二值遮罩组合的插值
  - 混合样本 $z_{mix} = m_1 z_1 \odot \ldots \odot m_k z_k$
  - 混合标签 $y_{mix} = \lambda_1 y_1 + \cdots + \lambda_k y_k$
  - 插值系数 $m_i \in B := \{0,1\}^n, \sum_{i=0}^{k} m_i = 1_B$

$$\lambda_i = \frac{Num_{m_i=1}}{n}$$



- 从 $Uniform(0,1)$ 中抽取 p。
- 如果 k=2，从 n重 $Bernoulli(p)$ 中抽取 m_1，n 是 z 的维度。
- 如果 k>2，从 q重 $Bernoulli(p)$ 中抽取 m_2，q 是向量 $1_B - m_1$ 中非零元素的数量。

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

阶段二：多模式流形插值

- 基于线性组合(Convex Combination)的二元/三元插值
- 基于二值遮罩组合(Binary Mask combination)的二元/三元插值

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

阶段三：多目标混合训练

- 常规训练
  - 在原始干净的训练集上训练 DNN

$$D_{ori\_tra} = \{(x, y_{true})\}$$

  - 基于独热编码的交叉熵损失

$$L(f(x), y_{true}) = \sum_{i=1}^{C} y_i \log(p_i)$$

  - 优化目标

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D_{ori\_tra}} L(f_\theta(x, y))$$

- 微调训练
  - 在扩增数据集上重新训练原始 DNN

$$D_{fin\_tun} = D_{mix} \cup D_{ori\_tra}$$

  - 基于软标签的交叉熵损失

$$\begin{aligned} & L_{soft}(f(x), y_{mix}) \\ & = L_{soft}(f(x), \alpha_1 y_1 + \cdots + \alpha_k y_k) \\ & = \alpha_1 L(f(x), y_1) + \cdots + \alpha_k L(f(x), y_k) \end{aligned}$$

  - 优化目标

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D_{fin\_tun}} L_{soft}(f_\theta(x, y))$$

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

口 实验评估 – 实验设置

- 环境
  - PyTorch 1.8.1、CUDA V11.1.74
  - NVIDIA GV102 GPU
- 数据集
  - CIFAR-10、SVHN
  - ImageNet-Mixed10（10 个类别）
- 模型
  - 基于卷积块：Alexnet 和 VGG
  - 基于残差块：ResNet、DenseNet、PreActResNet 和 WideResNet
  - 基于 Inception 块：GoogLeNet
- 攻击方法
  - 流形外攻击 (5)：FGSM、PGD、AutoAttack、DeepFool、CW
  - 流形内攻击 (2)：OM-FGSM、OM-PGD
  - 感知攻击 (4)：Fog, Snow, Elastic, JPEG
- 防御方法对比
  - 标准训练
  - 混合训练方法 (5)
  - 对抗训练方法 (2)

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

口 实验评估 – 实验设置

- 对比防御方法的敌手能力与防御者能力假设比较

Table 3.4 Comparison of defense methods attributes

Method	Faced Attack Surfaces	Knowledge of Attacker	Augmentation Space
PGD-AT[12]	Off-manifold	Known	Input Space
PGD-DMAT[38]	Off-manifold & On-manifold	Known	Input & Latent Space
InputMixup[39]	Off-manifold	Unknown	Input Space
CutMix[42]	Off-manifold	Unknown	Input Space
PuzzleMixup[43]	Off-manifold	Unknown	Input Space
ManifoldMixup[45]	Off-manifold	Unknown	Latent Space
PatchUp[46]	Off-manifold	Unknown	Latent Space
`LarepMixup`(Ours)	Off-manifold & On-manifold	Unknown	Latent Space

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 横向对比实验

- 与SOTA混合训练的性能比较
- 流形外扰动 δ 预算 ϵ=0.05，单步预算为0.02。
- 流形上扰动 ζ 预算 η=0.05，单步预算为0.005。

Table 2: Accuracy (%) of CIFAR-10 classification models on off/on-manifold adversarial examples

PreActResNet18

Method	Clean	FGSM	PGD	AutoAttack	DeepFool	CW	Known Attacker	Modify Network
Vanilla	**87.37±0.00**	32.07±0.00	28.93±0.00	7.59±0.00	10.36±0.00	2.60±0.00		
InputMixup[56]	84.48±1.45	63.58±3.36	68.12±3.46	56.63±10.20	37.97±2.58	41.11±2.10	✗	✗
CutMix[54]	82.14±3.00	65.51±1.03	69.67±1.34	64.41±5.55	36.79±2.60	39.74±3.10	✗	✗
PuzzleMixup[29]	83.11±1.64	65.73±2.46	70.35±2.60	64.03±6.06	38.86±1.53	41.83±1.74	✗	✗
ManifoldMixup[52]	71.10±4.17	49.26±1.34	52.49±1.91	44.08±1.60	25.33±2.76	27.19±2.53	✗	✓
PatchUp[14]	72.02±4.10	51.35±2.13	55.91±2.29	44.61±2.56	28.81±3.35	30.94±3.13	✗	✓
Ours-Convex	84.02±1.77	**68.86±2.88**	**72.65±3.59**	**66.98±5.93**	39.03±2.16	42.03±2.31	✗	✗
Ours-Mask	84.60±1.27	66.56±1.50	71.22±1.93	63.69±4.61	**39.27±2.97**	**42.54±2.74**	✗	✗

PreActResNet34

Method	Clean	FGSM	PGD	AutoAttack	DeepFool	CW	Known Attacker	Modify Network
Vanilla	**83.57±0.00**	31.37±0.00	25.71±0.00	5.27±0.00	12.27±0.00	1.89±0.00		
InputMixup[56]	68.42±7.38	62.19±4.22	63.84±4.98	63.79±4.99	26.36±4.07	29.77±4.16	✗	✗
CutMix[54]	71.21±6.16	62.45±2.71	64.61±3.50	64.30±3.16	28.88±2.07	32.12±2.58	✗	✗
PuzzleMixup[29]	67.06±7.62	60.89±4.99	62.55±5.76	62.66±5.84	25.89±2.98	28.96±3.37	✗	✗
ManifoldMixup[52]	73.69±1.78	49.65±1.94	52.24±2.08	43.75±2.04	31.09±3.13	32.81±3.18	✗	✓
PatchUp[14]	72.71±2.96	49.53±1.44	52.76±2.80	42.31±1.80	32.35±3.66	34.10±3.45	✗	✓
Ours-Convex	78.44±1.60	**67.81±1.04**	**71.12±1.08**	**70.60±1.30**	**33.98±1.04**	**37.42±1.03**	✗	✗
Ours-Mask	77.13±3.17	66.16±1.58	68.96±1.62	68.40±2.16	32.95±2.26	36.38±2.23	✗	✗

线性插值

遮罩插值

粗体：冠军

下划线：亚军

基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 横向对比实验

- 与SOTA混合训练的性能比较
  - 流形外扰动 δ 预算 ϵ=0.05，单步预算为0.02。
  - 流形上扰动 ζ 预算 η=0.05，单步预算为0.005。

线性插值

遮罩插值

Table 2: Accuracy (%) of CIFAR-10 classification models on off/on-manifold adversarial examples

PreActResNet18

Method	Clean	FGSM	PGD	AutoAttack	DeepFool	CW	OM-FGSM	OM-PGD	Known Attacker	Modify Network
Vanilla	**87.37±0.00**	32.07±0.00	28.93±0.00	7.59±0.00	10.36±0.00	2.60±0.00	51.02±0.00	21.68±0.00		
InputMixup[56]	84.48±1.45	63.58±3.36	68.12±3.46	56.63±10.20	37.97±2.58	41.11±2.10	58.53±0.43	44.11±1.34	✗	✗
CutMix[54]	82.14±3.00	65.51±1.03	69.67±1.34	64.41±3.55	36.79±2.60	39.74±3.10	57.59±0.31	43.50±1.71	✗	✗
PuzzleMixup[29]	83.11±1.64	65.73±2.46	70.35±2.60	64.03±6.06	38.86±1.53	41.83±1.74	57.80±0.77	43.68±2.19	✗	✗
ManifoldMixup[52]	71.10±4.17	49.26±1.34	52.49±1.91	44.08±1.60	25.33±2.76	27.19±2.53	50.16±1.66	38.64±0.80	✗	✓
PatchUp[14]	72.02±4.10	51.35±2.13	55.91±2.29	44.61±2.56	28.81±3.35	30.94±3.13	52.22±2.32	41.33±1.24	✗	✓
Ours-Convex	84.02±1.77	**68.86±2.88**	**72.65±3.59**	**66.98±5.93**	39.03±2.16	42.03±2.31	**60.02±0.91**	**46.72±1.52**	✗	✗
Ours-Mask	84.60±1.27	66.56±1.50	71.22±1.93	63.69±4.61	**39.27±2.97**	**42.54±2.74**	58.36±0.60	44.80±0.73	✗	✗

PreActResNet34

Method	Clean	FGSM	PGD	AutoAttack	DeepFool	CW	OM-FGSM	OM-PGD	Known Attacker	Modify Network
Vanilla	**83.57±0.00**	31.37±0.00	25.71±0.00	5.27±0.00	12.27±0.00	1.89±0.00	49.23±0.00	17.05±0.00		
InputMixup[56]	68.42±7.38	62.19±4.22	63.84±4.98	63.79±4.99	26.36±4.07	29.77±4.16	54.68±3.84	47.18±2.29	✗	✗
CutMix[54]	71.21±6.16	62.45±2.71	64.61±3.50	64.30±3.16	28.88±2.07	32.12±2.38	55.65±2.56	46.40±0.99	✗	✗
PuzzleMixup[29]	67.06±7.62	60.89±4.99	62.55±5.76	62.66±5.84	25.89±2.98	28.96±3.37	54.04±3.87	46.31±2.05	✗	✗
ManifoldMixup[52]	73.69±1.78	49.65±1.94	52.24±2.08	43.75±2.04	31.09±3.13	32.81±3.18	52.99±0.24	39.47±1.34	✗	✓
PatchUp[14]	72.71±2.96	49.53±1.44	52.76±2.80	42.31±1.80	32.35±3.66	34.10±3.45	53.03±2.37	39.38±1.63	✗	✓
Ours-Convex	78.44±1.60	**67.81±1.04**	**71.12±1.08**	**70.60±1.30**	**33.98±1.04**	**37.42±1.03**	**58.96±0.67**	**47.99±1.16**	✗	✗
Ours-Mask	77.13±3.17	66.16±1.58	68.90±1.62	68.40±2.16	32.95±2.26	36.38±2.23	58.31±0.96	47.30±1.06	✗	✗

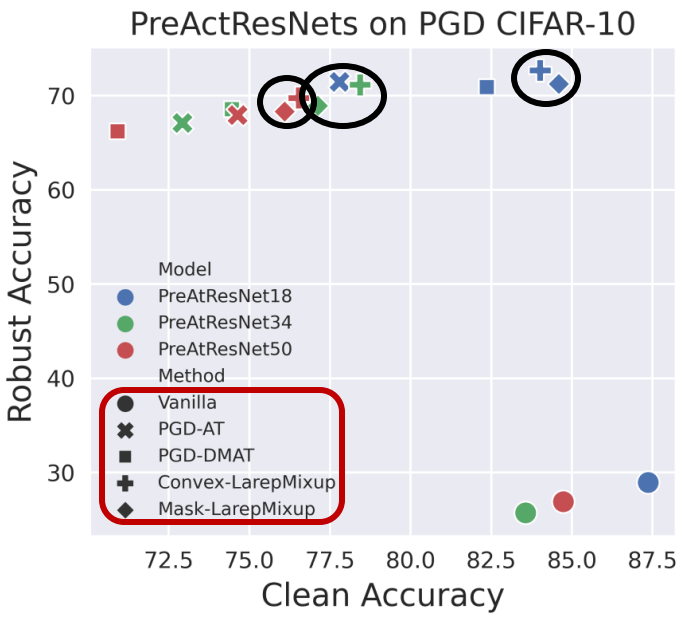
基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 横向对比实验

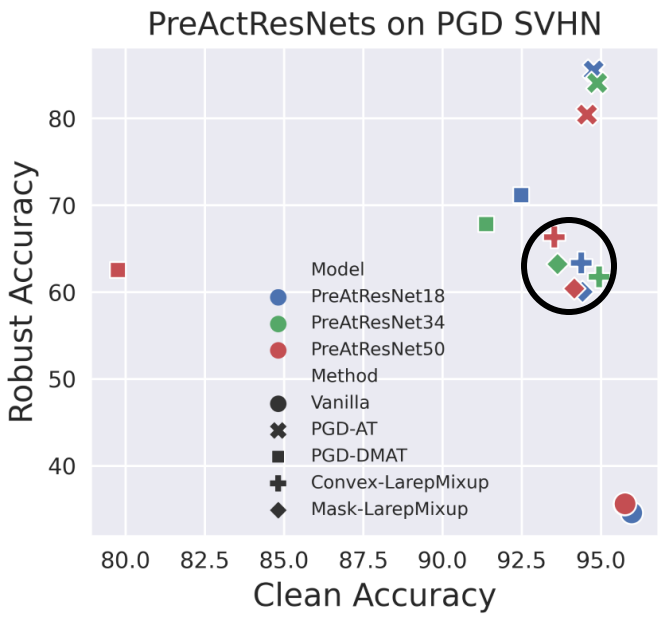
- 与SOTA对抗训练的性能比较
  - 流形外扰动 δ 预算 ϵ=0.05，单步预算为0.02。
  - 流形上扰动 ζ 预算 η=0.05，单步预算为0.005。

- ✓ 位置越左的得分表示在干净样本上准确率越高。
- ✓ 越上的分数表示在对抗样本上准确率越高。
- ✓ 相同颜色表示一组比较结果。

CIFAR-10



SVHN



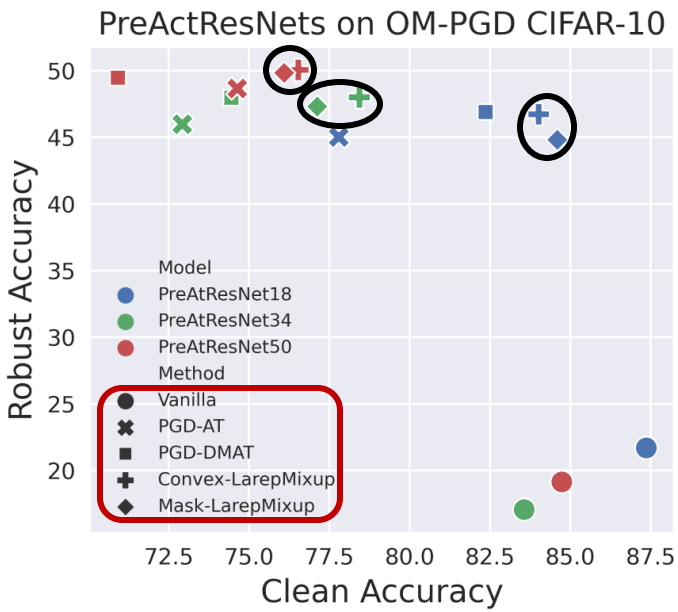
基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 横向对比实验

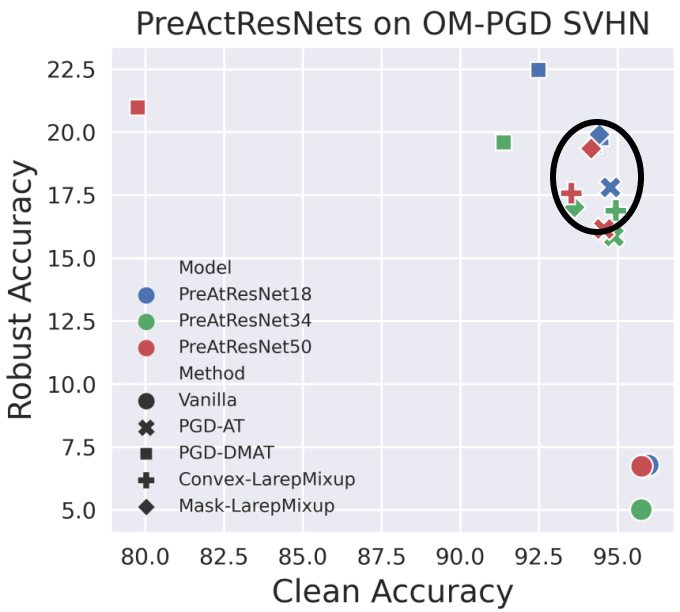
- 与SOTA对抗训练的性能比较
  - 流形外扰动 δ 预算 ϵ=0.05，单步预算为0.02。
  - 流形上扰动 ζ 预算 η=0.05，单步预算为0.005。

- ✓ 位置越左的得分表示在干净样本上准确率越高。
- ✓ 越上的分数表示在对抗样本上准确率越高。
- ✓ 相同颜色表示一组比较结果。

CIFAR-10



SVHN

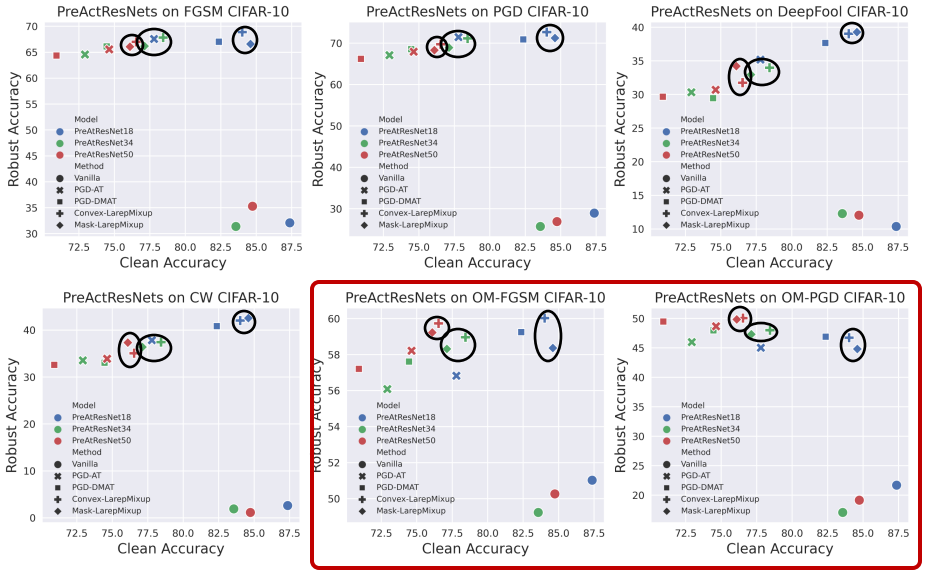


基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

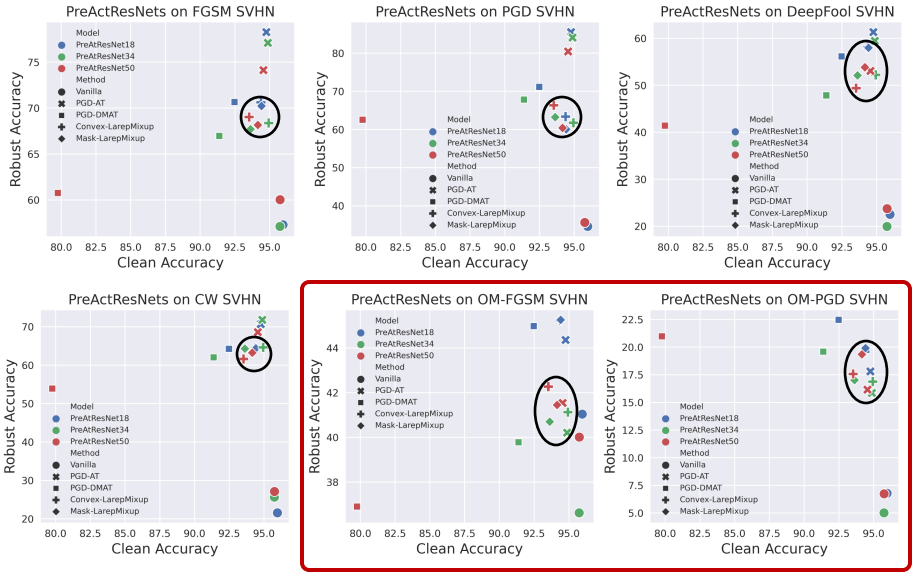
□ 实验评估 – 横向对比实验

- ➢ 与SOTA对抗训练的性能比较
  - ○ 流形外扰动 δ 预算 ϵ=0.05，单步预算为0.02。
  - ○ 流形上扰动 ζ 预算 η=0.05，单步预算为0.005。

CIFAR-10



SVHN

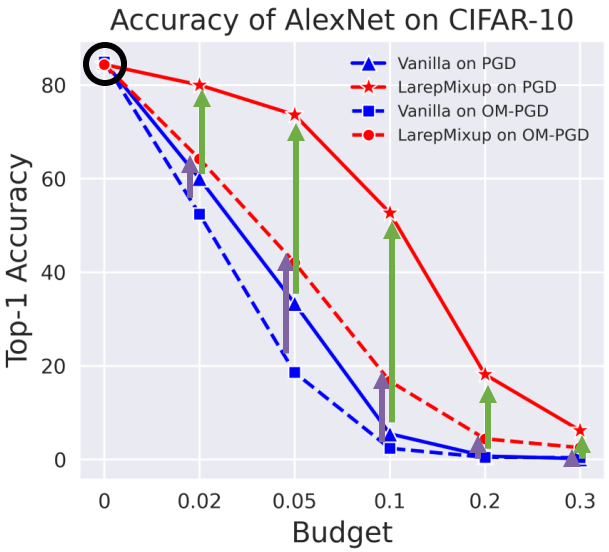


基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

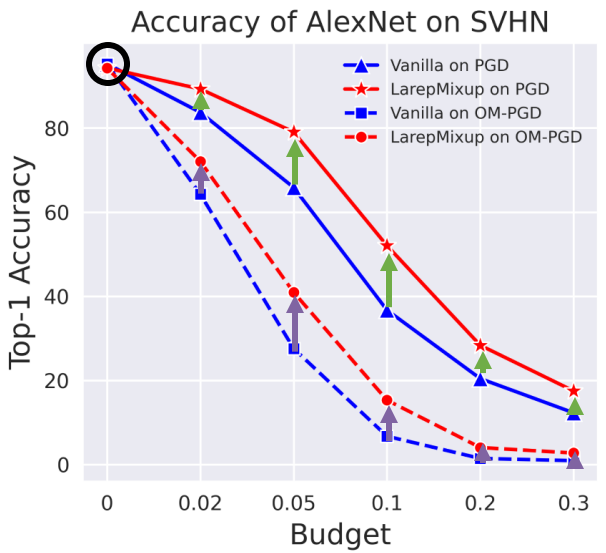
实验评估 – 纵向对比实验

- 针对不同对抗 l_p 攻击预算的鲁棒性： $||\delta||_p \leq \epsilon$ 和 $||\zeta||_p \leq \eta$
  - 流形外扰动 δ 预算 $\epsilon \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$，单步预算为0.02。
  - 流形上扰动 ζ 预算 $\eta \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$，单步预算为0.005。

CIFAR-10



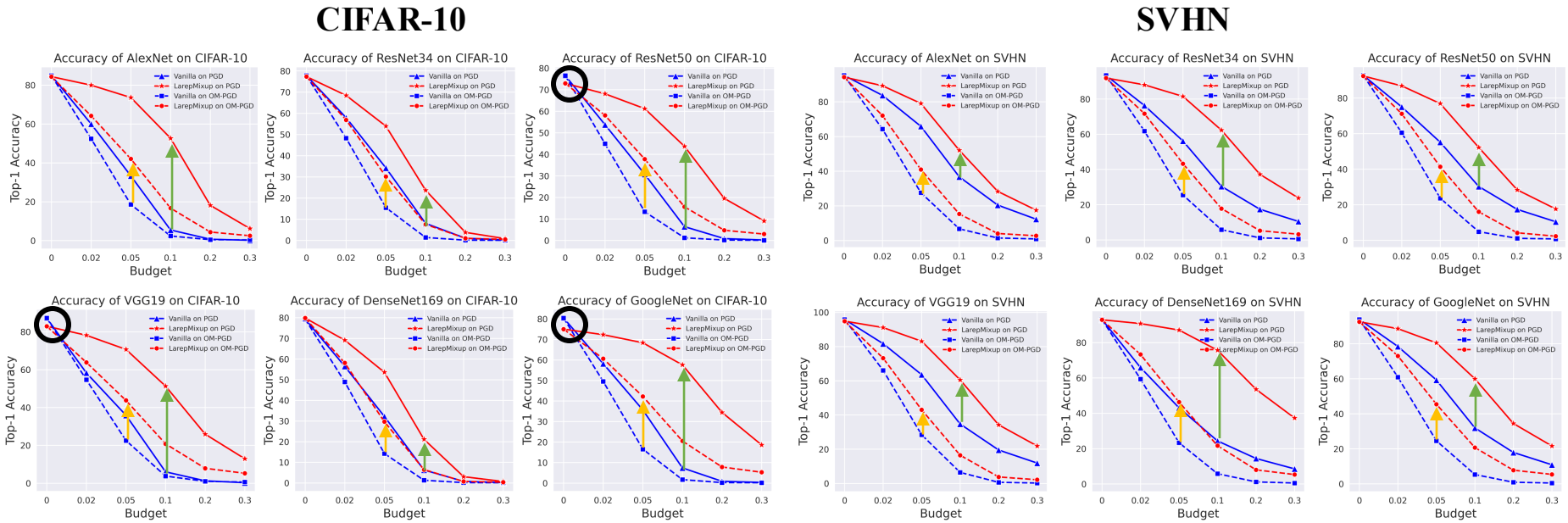
SVHN



基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 纵向对比实验

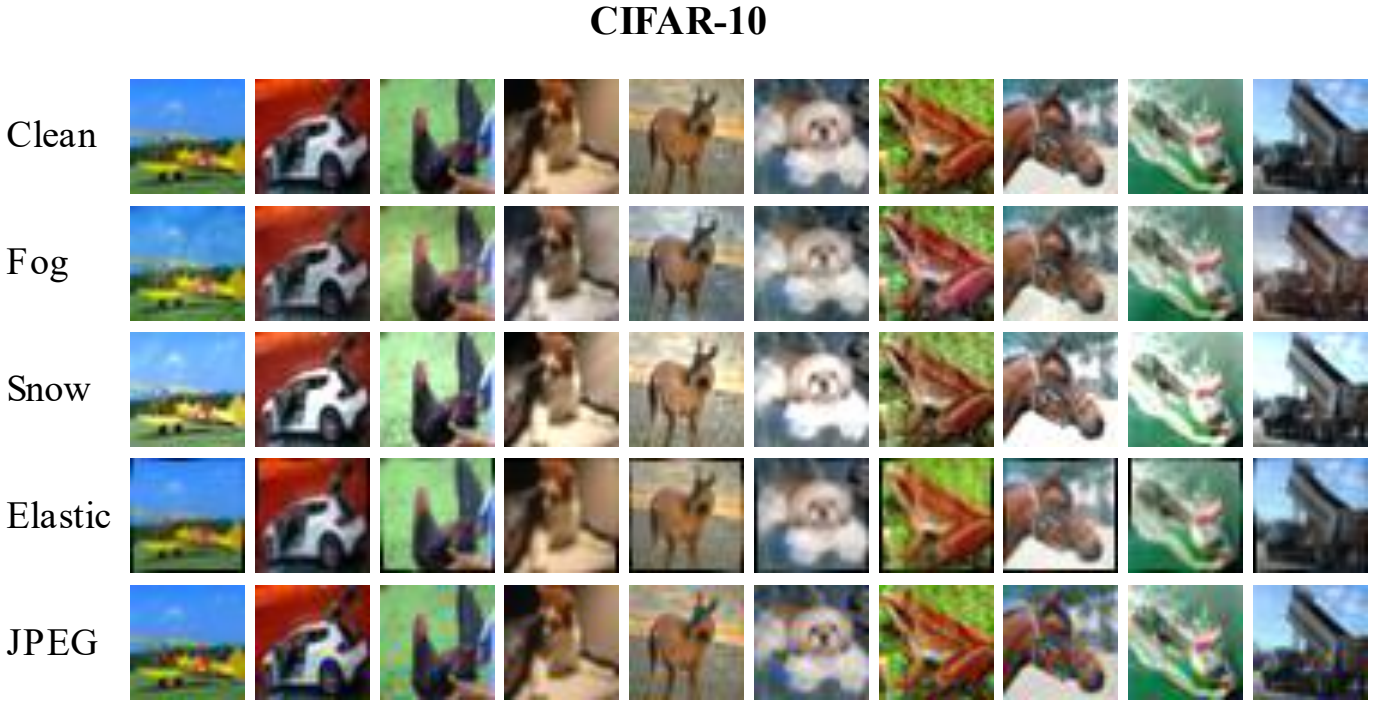
- 针对不同对抗 l_p 攻击预算的鲁棒性： $||\delta||_p \leq \epsilon$ 和 $||\zeta||_p \leq \eta$
  - 流形外扰动 δ 预算 $\epsilon \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$，单步预算为0.02。
  - 流形上扰动 ζ 预算 $\eta \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$，单步预算为0.005。



基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 纵向对比实验

- 针对非l_p感知攻击的鲁棒性。
  - 雾化、雪化、弹性、JPEG 压缩攻击

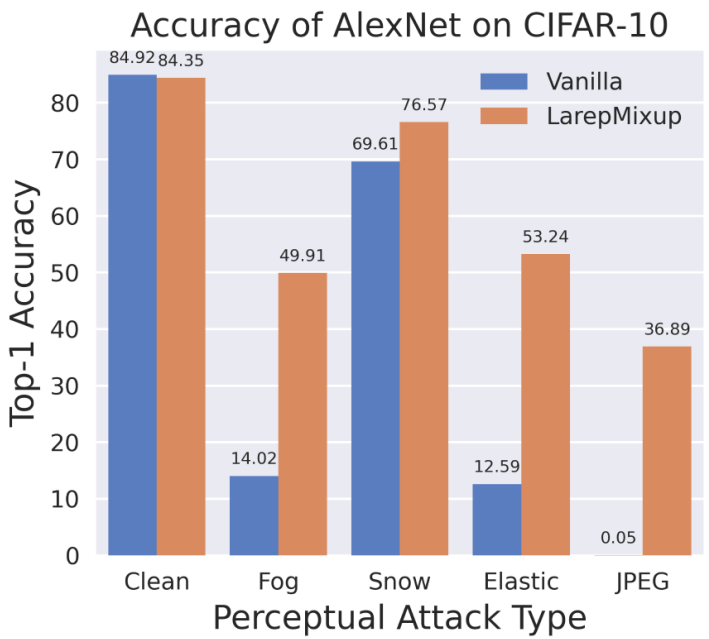


基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

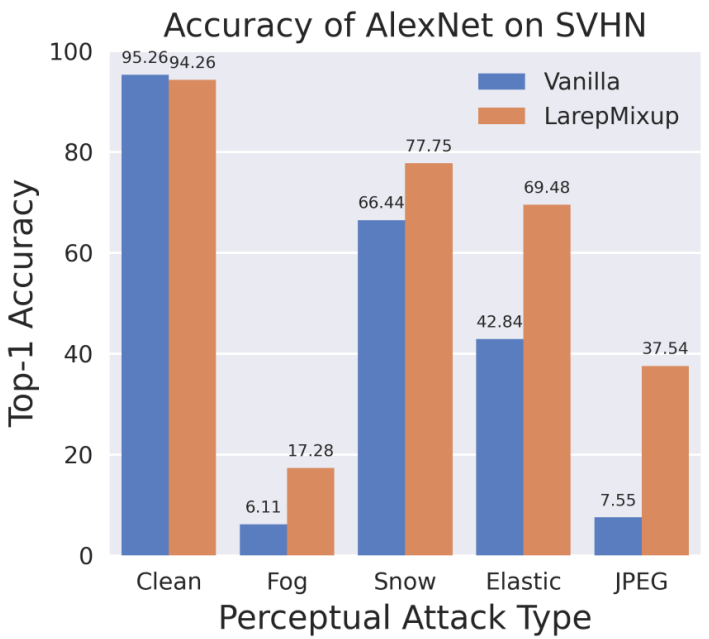
实验评估 – 纵向对比实验

- 针对非l_p感知攻击的鲁棒性。
  - 雾化、雪化、弹性、JPEG 压缩攻击

CIFAR-10



SVHN

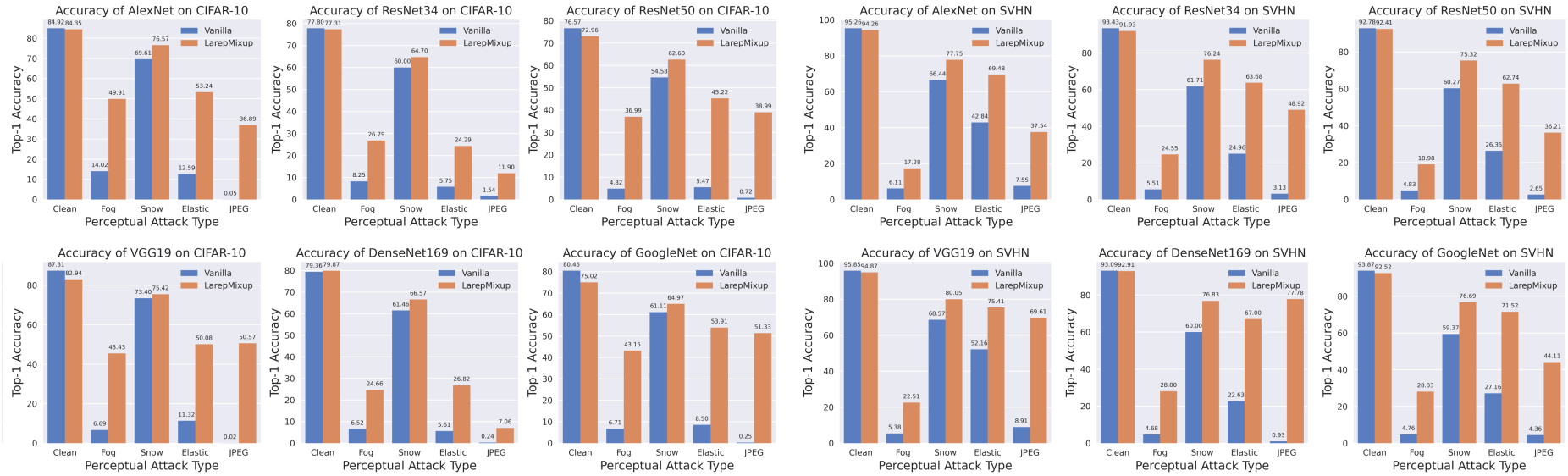


基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

□ 实验评估 – 纵向对比实验

- ➢ 针对非l_p感知攻击的鲁棒性。
  - ○ 雾化、雪化、弹性、JPEG 压缩攻击

CIFAR-10 **SVHN**



基于潜在空间表征混合的深度神经网络鲁棒性泛化方案

实验评估 – 纵向对比实验

- 不同混合模式的鲁棒性。
  - 二元 / 三元 + 线性 / 遮罩插值
  - 高维ImageNet数据集

Table 4: Robust accuracy (%) of PreActResNet18 under different mixing modes (ImageNet-Mixed10)

Method	Vanilla	Dual-LarepMixup		Ternary-LarepMixup	
		Convex	Mask	Convex	Mask
Clean	90.47	90.57±0.55	**90.89±0.35**	90.67±0.21	90.24±1.25
FGSM	13.93	17.09±0.29	16.21±0.14	16.71±0.34	**17.29±0.94**
PGD	2.00	5.38±0.81	4.68±0.45	4.73±0.69	**5.81±1.32**
AutoAttack	0.00	**3.74±0.19**	3.68±0.29	3.60±0.18	3.66±0.04
DeepFool	8.87	**85.38±0.19**	83.98±0.42	84.89±0.18	83.93±1.00
CW	0.10	**84.61±0.30**	83.16±0.52	84.19±0.47	83.28±0.62
OM-FGSM	26.90	**59.91±1.30**	28.61±5.58	57.36±1.89	28.21±0.98
OM-PGD	20.43	**58.76±1.30**	27.99±5.92	56.59±1.87	27.47±1.44

流形外

流形上

结论

- 设计了一种基于多模式流形插值的数据扩增策略，支持以多元线性混合和二值遮罩混合两种模式混合数据流形上的潜在表征，以此合成接近模型决策边界或符合训练数据潜在分布的混合样本。
- 提出了一个基于语义混合样本的多目标训练算法，利用混合样本和混合标签学习平滑的深度神经网络决策边界，增强其对边界附近扰动的鲁棒性。
- 在多种基于深度神经网络的图像分类模型和数据集上对所提的鲁棒性泛化方案进行实验评估。在白盒和黑盒场景中应对对抗攻击时，所提方法实现了像素级和表征级对抗鲁棒性的增强，提升了鲁棒性在广泛的输入空间和潜在空间扰动上的泛化能力。

Latent Representation Mixup 方案一

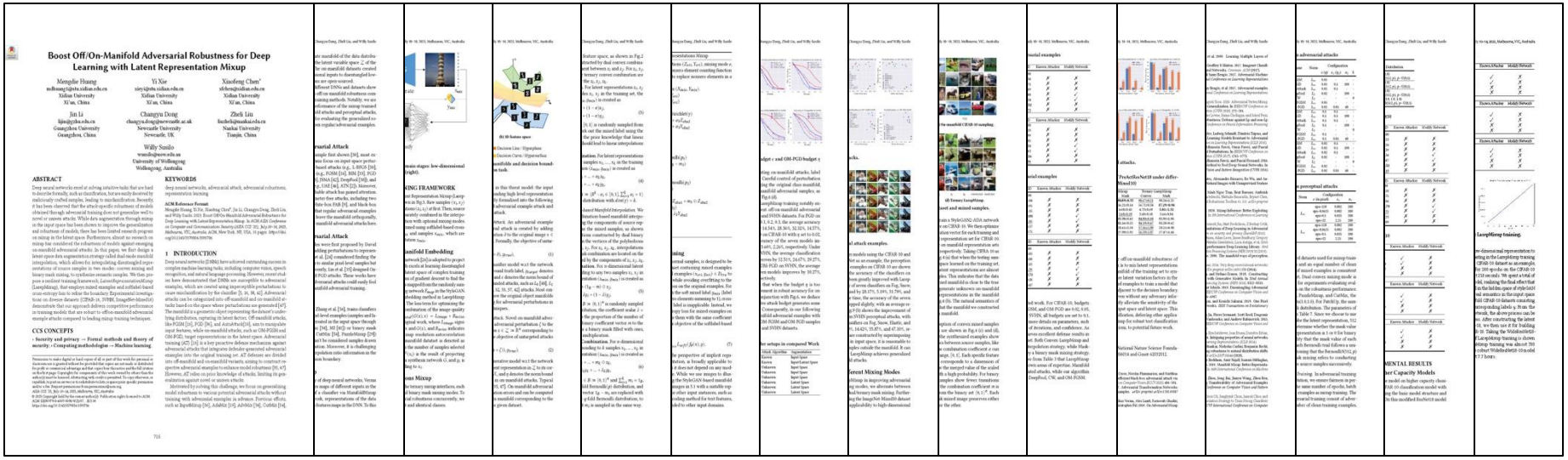
ACM ASIACCS 2023
Melbourne, Australia

□ 主要成果

主要成果发表在密码学会推荐 B 类 会议，CCF 推荐网络与信息安全 C 类 会议

ACM Asia Conference on Computer and Communications Security (AsiaCCS)

- **Mengdie Huang**, Yi Xie, **Xiaofeng Chen**, Jin Li, Changyu Dong, Zheli Liu, Willy Susilo. Boost Off/On-Manifold Adversarial Robustness for Deep Learning with Latent Representation Mixup [C]. *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*, 2023, 1(1):716-730. （2类贡献度）

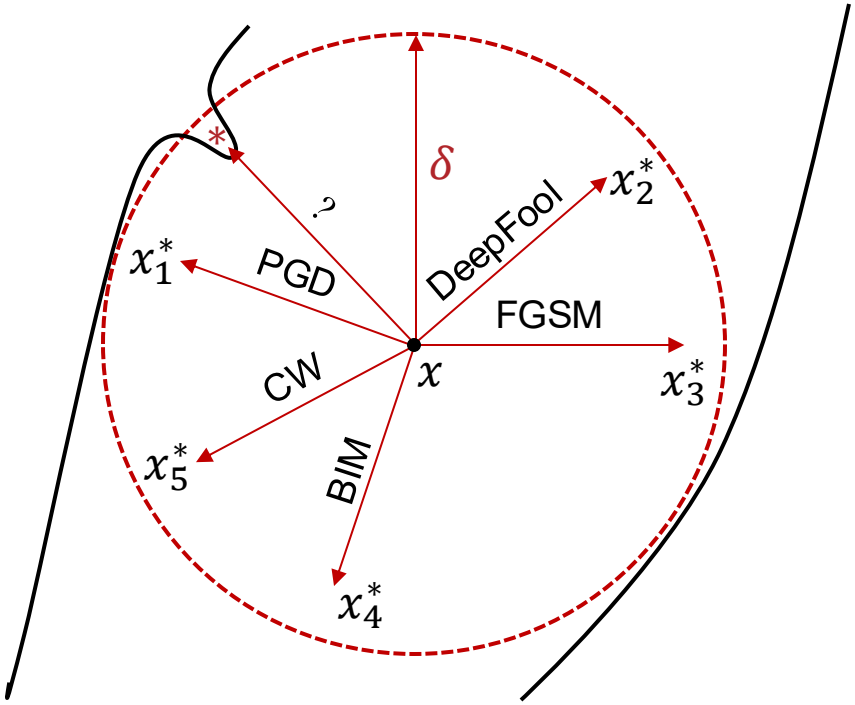

Boost Off/On-Manifold Adversarial Robustness for Deep Learning with Latent Representation Mixup

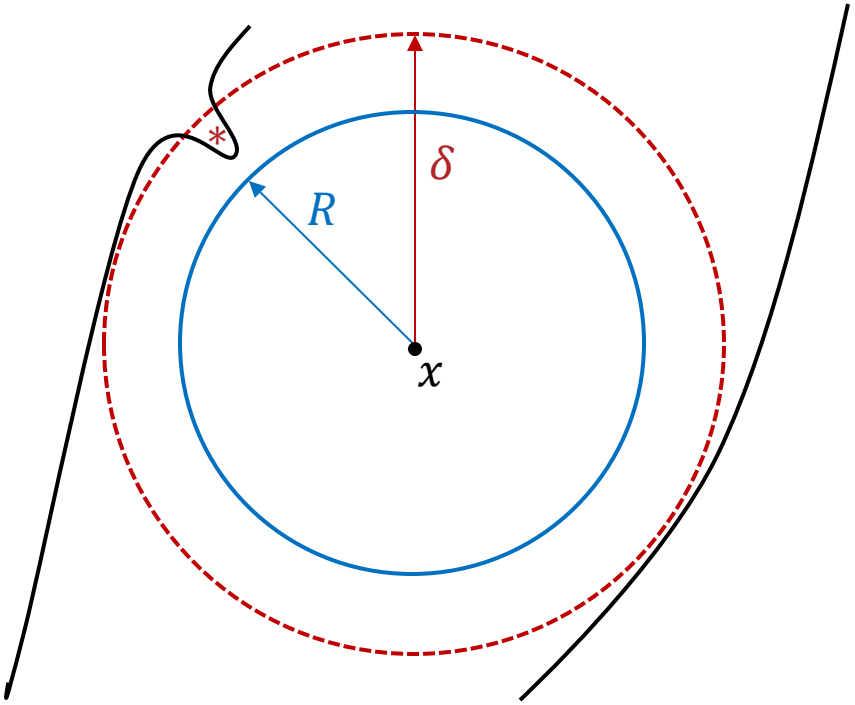
目 录

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 深度神经网络鲁棒防御角度

- ➢ 经验防御 (Empirical Defense)
  - ○ 通过启发式训练提升鲁棒性
- ➢ 可验证防御 (Certified Defense)
  - ○ 通过可验证半径提供鲁棒性证书

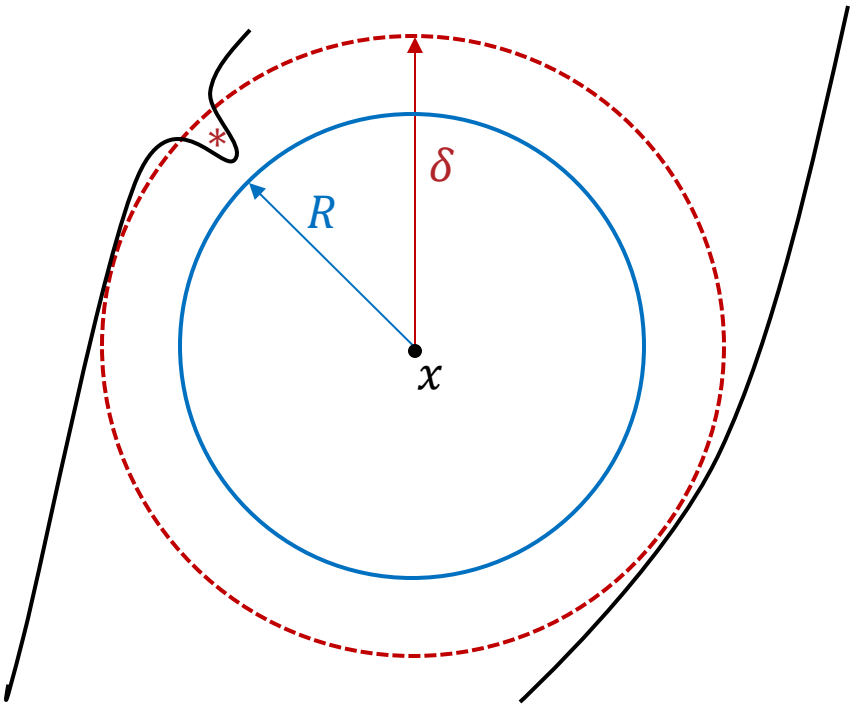




基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 深度神经网络鲁棒防御角度

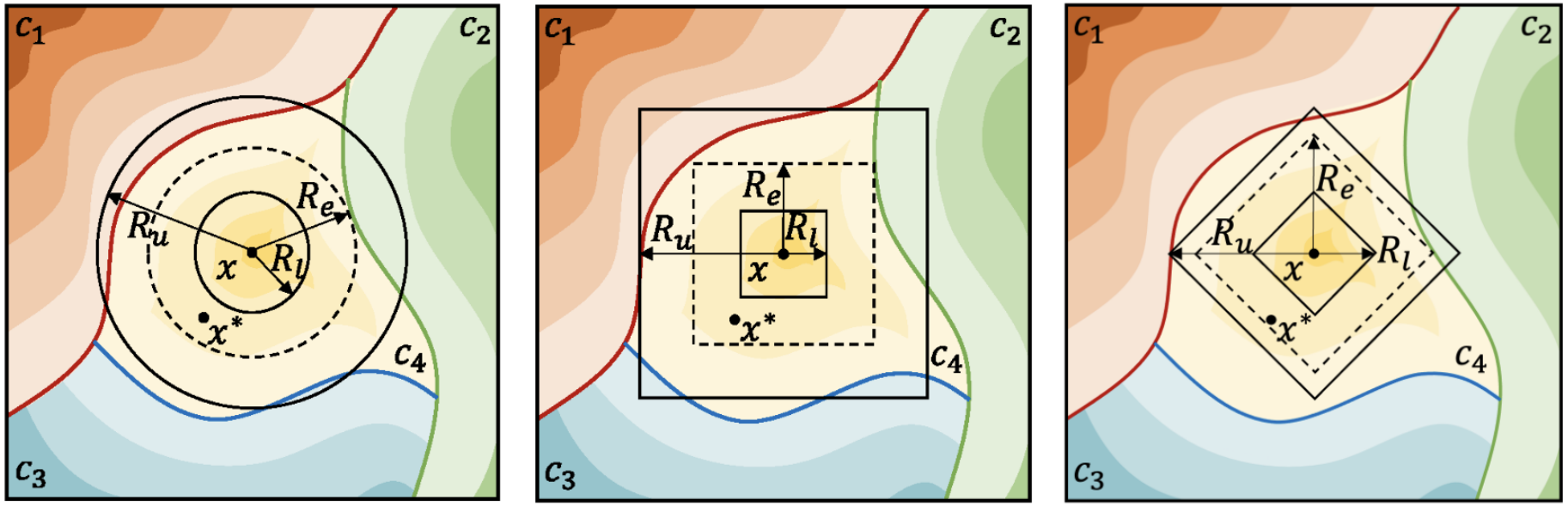
- ➢ 证书形式 (x, F, R)
  - ○ 可验证鲁棒半径R
- ➢ 验证内容
  - ○ 对于任意输入x，保证DNN分类器 F 对 x 周围以l_p范数度量的半径为 R 的领域内数据点的预测是一致的，即该区域内任意对输入x的微小扰动（如对抗攻击）都无法改变模型预测结果。
- ➢ 可验证防御优势：
  - ○ 验证半径R衡量了模型在输入空间中能抵抗对抗攻击的最大扰动范围。
- ➢ 可验证防御 (Certified Defense)
  - ○ 通过可验证半径提供鲁棒性证书



基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 可验证l_p鲁棒半径类型

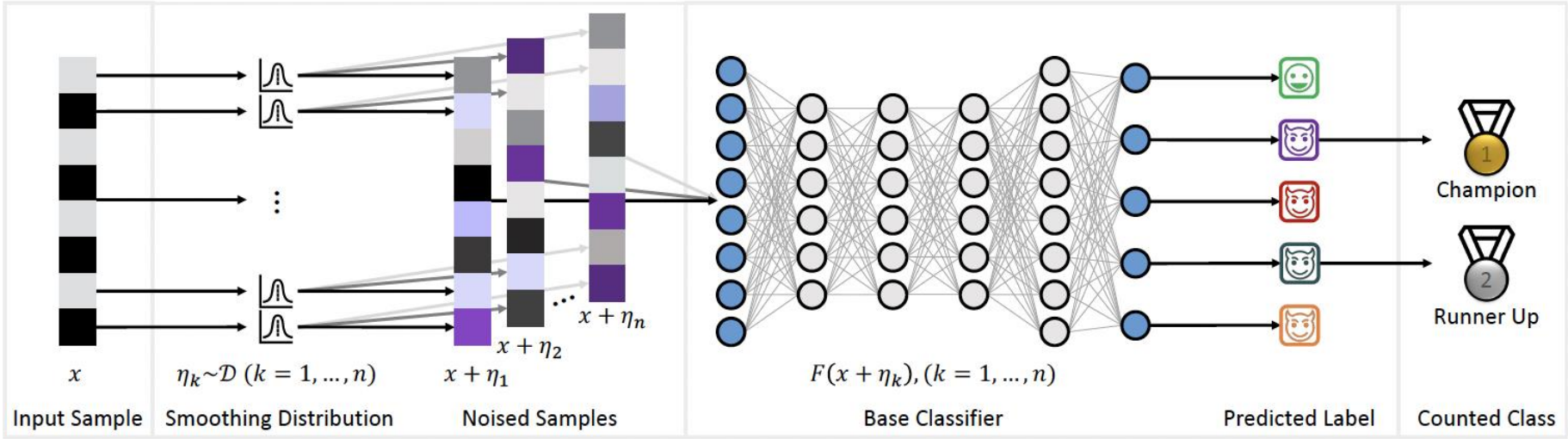
- 多类（四类）分类器在输入样本 x 上的 l_p 有界鲁棒验证半径
- 颜色越深，表示模型对输出的预测类别的置信度越高
- R_u 和 R_l 分别是模型在x上的精确鲁棒半径 R_e 的上下界



(a) l_2 radius: $||\delta||_2 < R$　　(b) l_∞ radius: $||\delta||_\infty < R$　　(c) l_1 radius: $||\delta||_1 < R$

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ Multi-Order Adaptive Randomized Smoothing (MARS) 框架



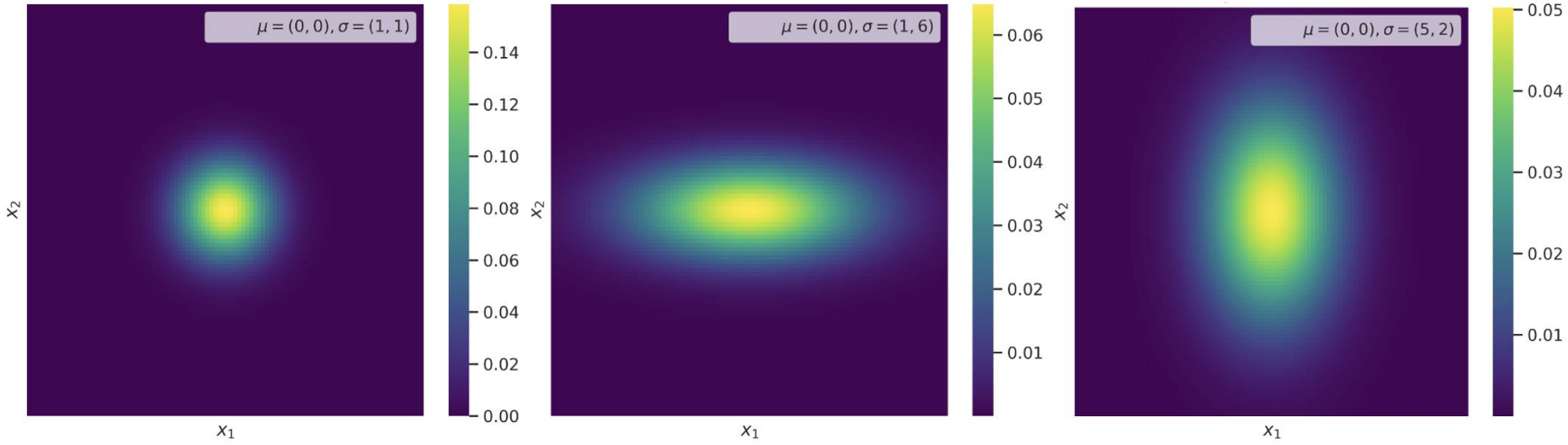
- 预测过程
- 目标：分类器预测输入 x 的类别
  - 选择平滑分布 $\mathcal{D}$
  - 采样n_{small}(100)个噪声向量η添加到 x
  - 分类器预测并确定冠军类和亚军类

- 验证过程
  - 目标：计算l_p 可验证鲁棒半径R
  - 采样n_{large} (10,000)个噪声η加到 x
  - 统计冠军类、亚军类预测频率
  - 使用统计假设检验估计鲁棒半径

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 阶段一：平滑分布参数优化

- 分布形状优化
- 分布规模优化
- 各维度具有相同或不同 σ 值的二元高斯分布 $N(\mu, \sigma)$ 的 PDF

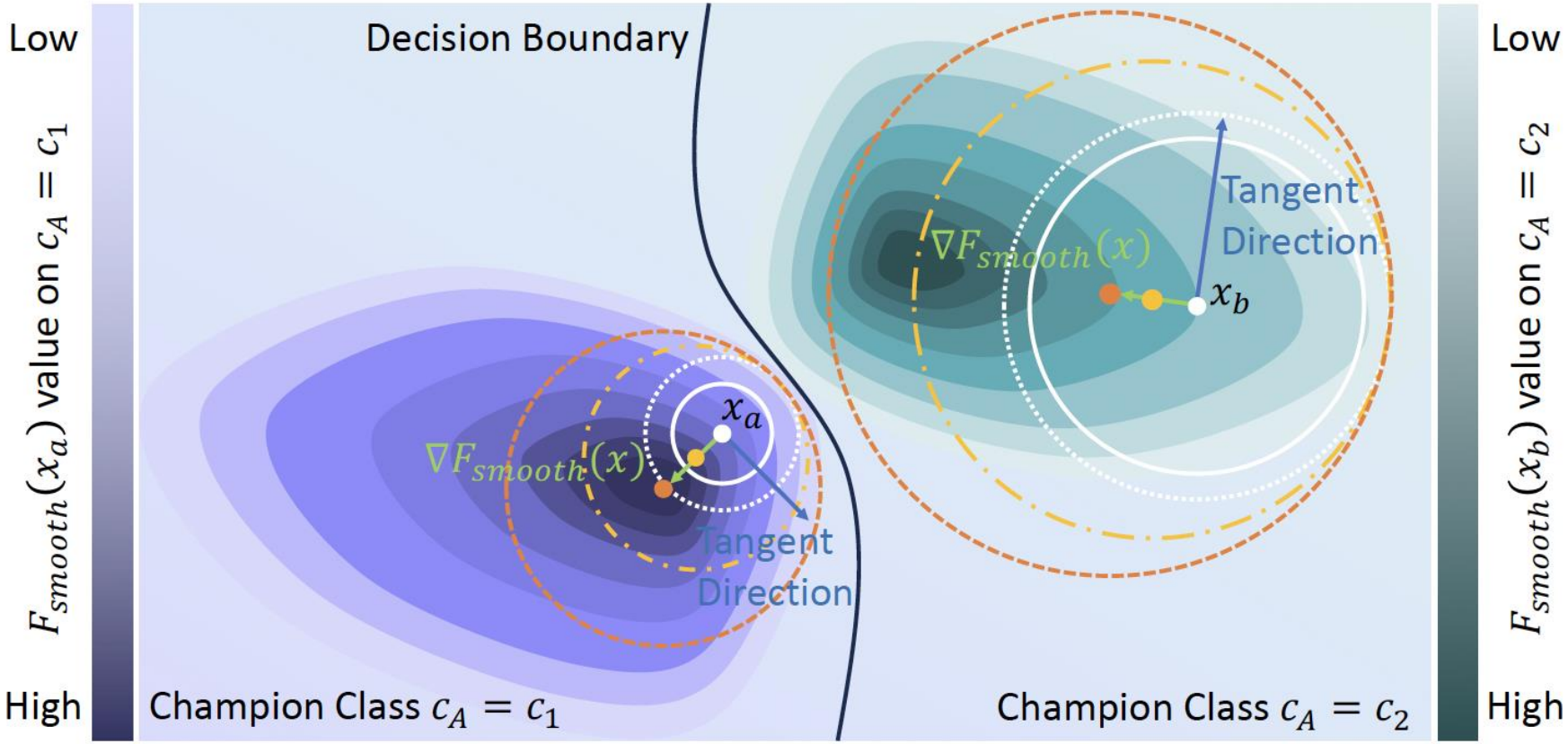


(a) Same σ across dimensions (b) Different σ across dimensions (c) Different σ across dimensions

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 阶段二：基于梯度的认证区域扩展

- 使用零阶概率信息和一阶梯度信息共同计算认证半径



基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

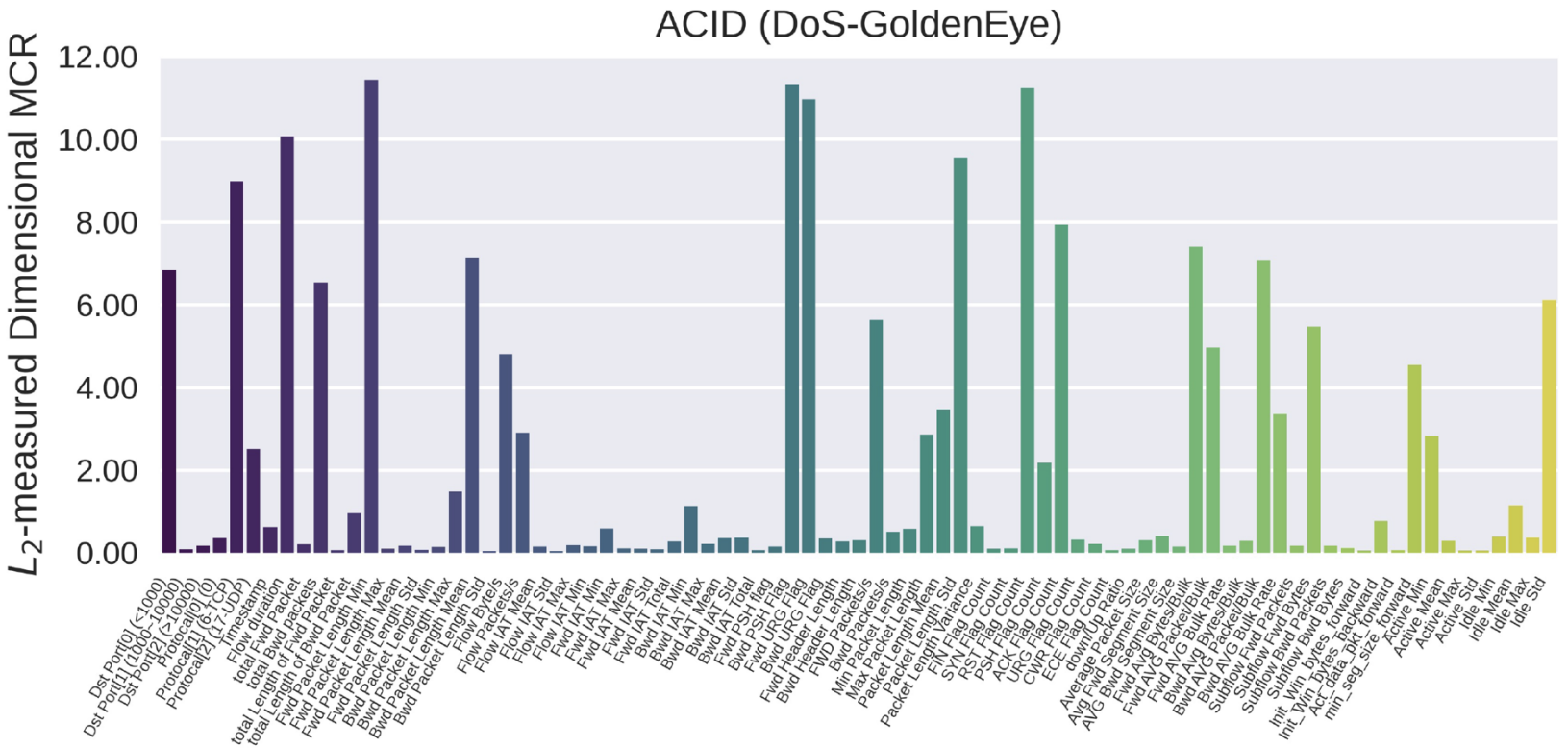
阶段三：维度半径权重计算

- 维度特征敏感性分析

$$s_i = d(f_\theta^c(x))/d(x_i) \quad s = (s_1, \ldots, s_d)$$

- 维度半径贡献量化：

$$R_i = w_i \times R, w_i = \frac{R_i}{R} = \frac{1/d}{\tilde{s}_i} = \frac{1}{d\tilde{s}_i}$$

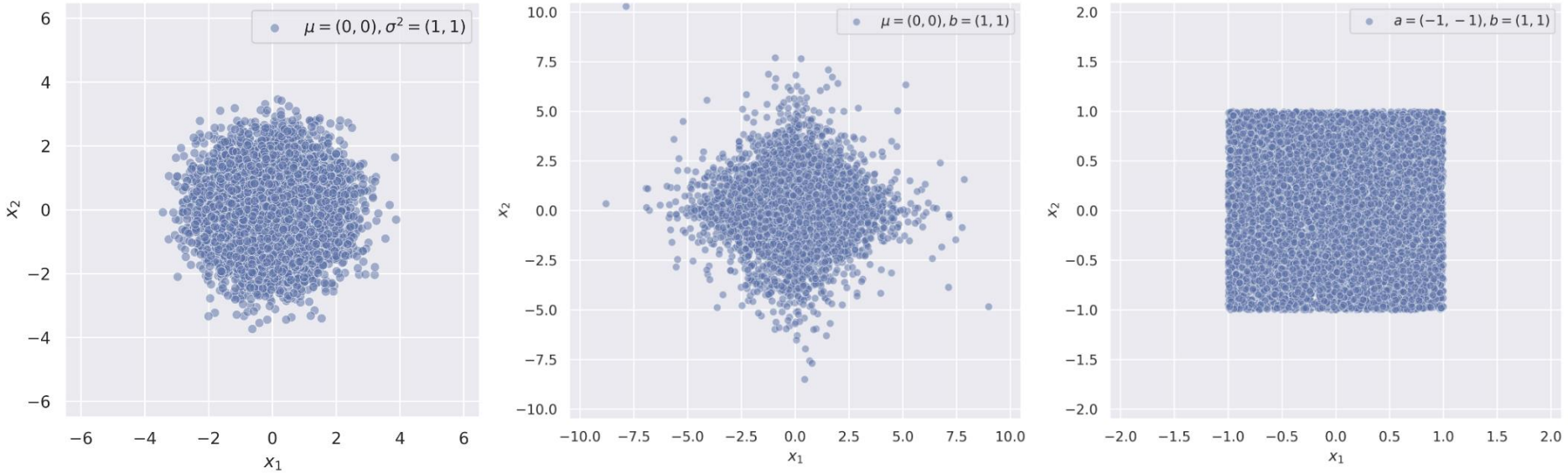


基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

阶段四：平滑分布多样性

- 多样概率分布采样区域对齐不同l_p有届区域
- 随机采样10,000个噪声样本

高斯分布对齐 l_2采样区域　拉普拉斯分布对齐 l_1 采样区域　均匀分布对齐 l_∞ 采样区域



(a) Gaussian distribution $\mathcal{N}(\mu, \sigma)$　(b) Laplacian distribution $\mathcal{L}(\mu, b)$　(c) Uniform distribution $\mathcal{U}(a, b)$

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

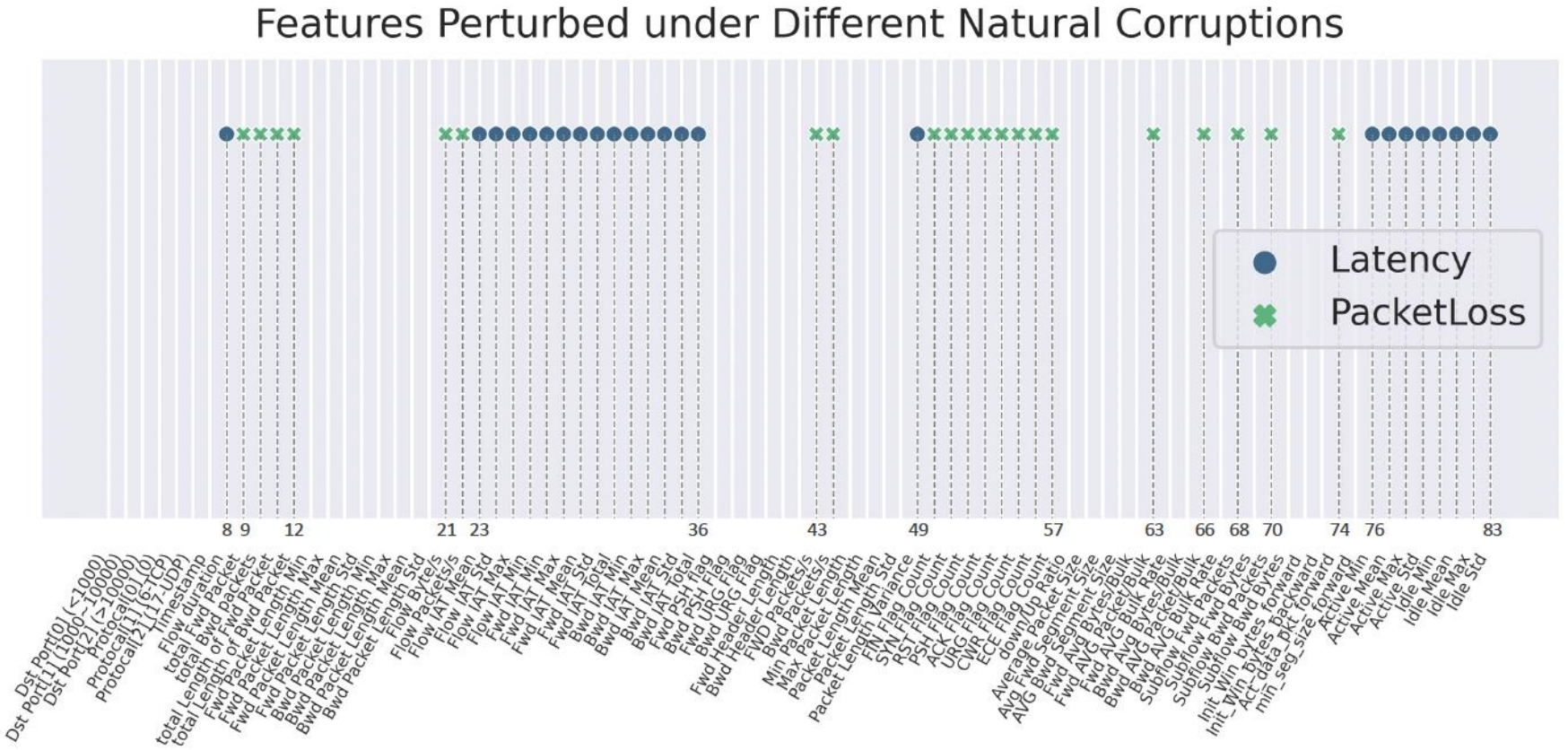
实验评估 – 实验设置

- 环境
  - PyTorch 2.0.1、SciPy V 1.11.2, CUDA V 11.7
  - NVIDIA GeForce 3090 GPU
- 数据集
  - CSE-CIC-IDS-2018
    - DosHolk-Drift Dataset
    - Inflation-Drift Dataset
    - Diverse Intrusion Dataset
    - Similar Intrusion Dataset
- 模型
  - CADE
  - ACID
- 攻击方法
  - 对抗攻击：l_2-PGD, l_1-PGD, l_∞-PGD, EAD
  - 自然损坏：Latency, Packet Loss
- 可验证防御对比方法
  - 随机平滑 (VRS)
  - 一阶随机平滑 (FRS)
  - 边界自适应随机平滑 (BARS)

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

实验评估 – 实验设置

- 自然损坏(Natural Corruption)下受到干扰的特征



基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 实验设置

➢ 评估指标

- 可验证鲁棒性 (Certified Robustness)
  - 平均验证半径
  - 认证准确率

$$Mean\ Certified\ Radius\ (MCR) = \frac{1}{N}\sum_{i=1}^{N} R_i$$

$$Certified\ Accuracy\ (CerAcc) = \frac{N_{(F_{smooth}(x)=y_{true})\&(R\geq R_{given})}}{N}$$

- 经验鲁棒性 (Empirical Robustness)
  - 对抗（恶意）样本的鲁棒准确率
  - 损坏（恶意和良性）样本的鲁棒准确率

$$Recall = \frac{TP}{TP+FN}$$

$$Robust\ Accuracy\ (RobAcc) = \frac{N_{(F_{smooth}(x^*)=y_{true})}}{N} = \frac{TP+TN}{TP+TN+FP+FN}$$

- 常规预测性能 Regular Predictive Performance
  - 干净准确率

$$Clean\ Accuracy\ (CleAcc) = \frac{N_{(F_{smooth}(x)=y_{true})}}{N}$$

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

实验评估 – 实验设置

- 用于评估的网络入侵检测 (Network Intrusion Detection,NID) 数据集信息

Dataset	CSE-CIC-IDS-2018-CADE				CSE-CIC-IDS-2018-ACID			
	DoS-Hulk-Drift Dataset		Infiltration-Drift Dataset		Diverse-Intrusions Dataset		Similar-Intrusions Dataset	
	Class	Number	Class	Number	Class	Number	Class	Number
Training	Benign	52996	Benign	52996	Benign	52996	Benign	52996
	SSH-Bruteforce	9385	SSH-Bruteforce	9385	FTP-Bruteforce	12590	DoS-GoldenEye	26565
	Infiltration	7390	DoS-Hulk	34789	DDoS-HOIC	53476	DoS-SlowHTTPTest	11191
	-	-	-	-	Bot	22584	DDoS-LOIC-HTTP	46095
Test	Benign	13249	Benign	13249	Benign	13249	Benign	13249
	SSH-Bruteforce	2346	SSH-Bruteforce	2346	FTP-Bruteforce	3148	DoS-GoldenEye	6641
	Infiltration	1894	DoS-Hulk	8697	DDoS-HOIC	13369	DoS-SlowHTTPTest	2798
	DoS-Hulk	43486	Infiltration	9327	Bot	5646	DDoS-LOIC-HTTP	11524

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

口 实验评估 – 实验设置

- 可验证防御方法比较
  - 异构性支持
  - 跨模型支持
  - 鲁棒证书多样性
  - 经验评估多样性

Table 4.4 Comparison of certified defense methods

Method	Heterogeneity	Universality	Robustness Guarantee Diversity			Adversarial Attacks			Natural Corruptions	
			l_2 Radius	l_1 Radius	l_∞ Radius	l_2 Attack	l_1 Attack	l_∞ Attack	Latency	Loss
VRS [75]	○	●	●	○	○	○	○	○	○	○
FRS [77]	○	●	●	●	●	○	○	○	○	○
BARS [78]	●	●	●	○	○	○	○	●	○	○
MARS	●	●	●	●	●	●	●	●	●	●

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

➢ l_2鲁棒性保证：平均认证半径 (MCR) 比较

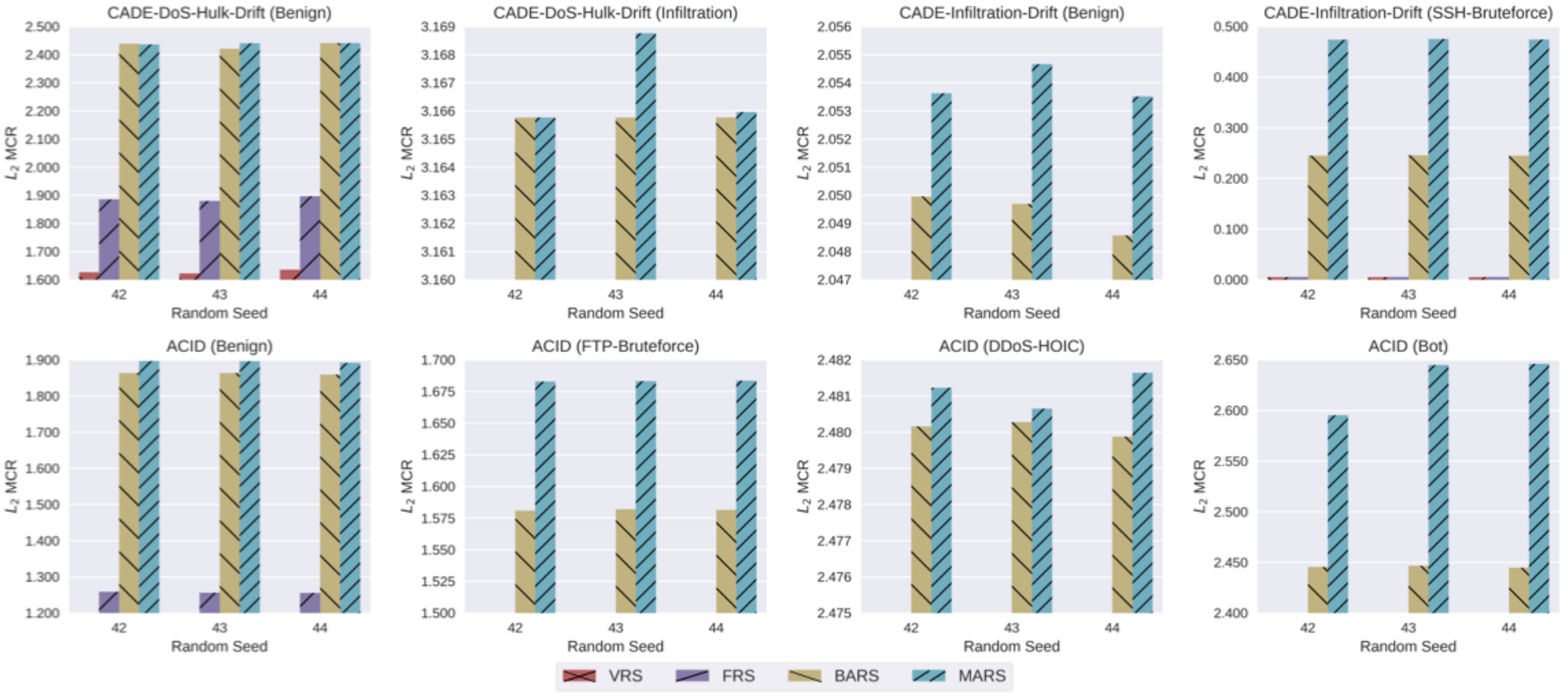
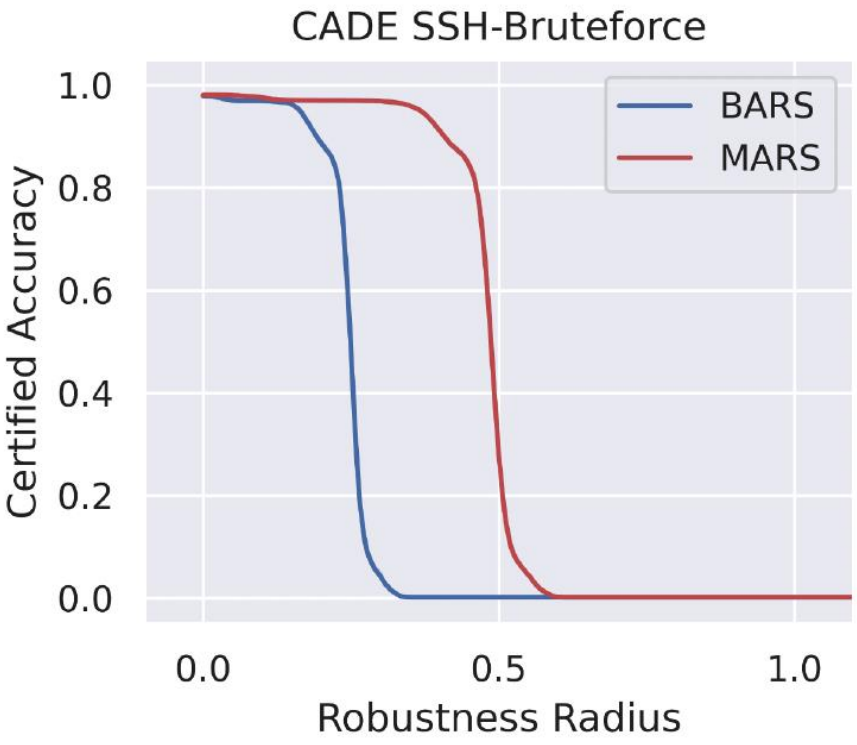


Figure 4.6 Comparison of l_2 Mean Certified Radius (MCR).

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

- l_2鲁棒性保证：可验证准确率比较



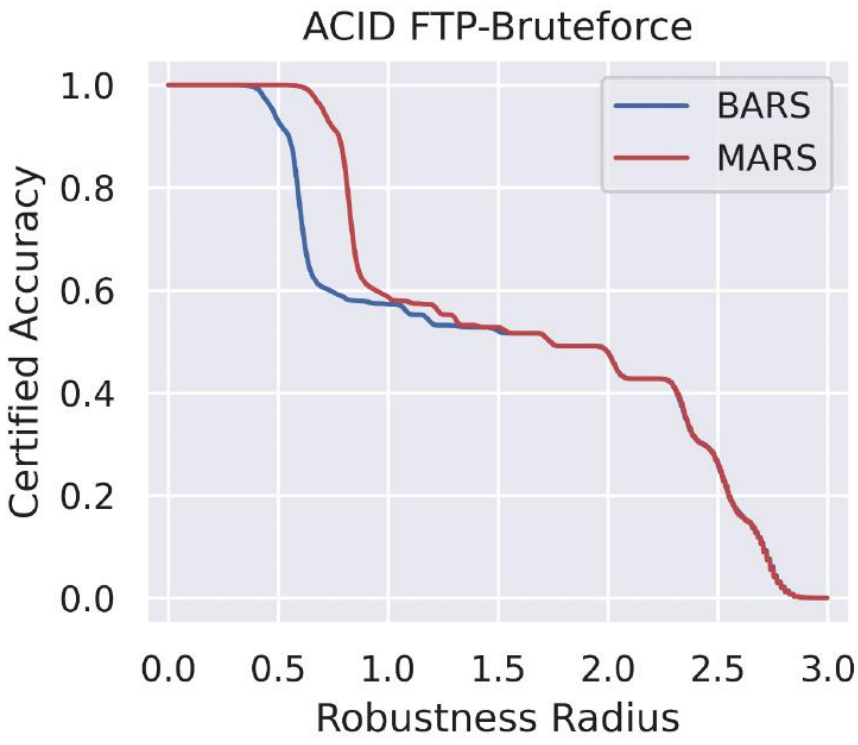


Figure 4.7 Comparison of certified accuracy of l_2 robustness guarantee.

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

➢ l_p鲁棒性保证：平均认证半径 (MCR) 比较

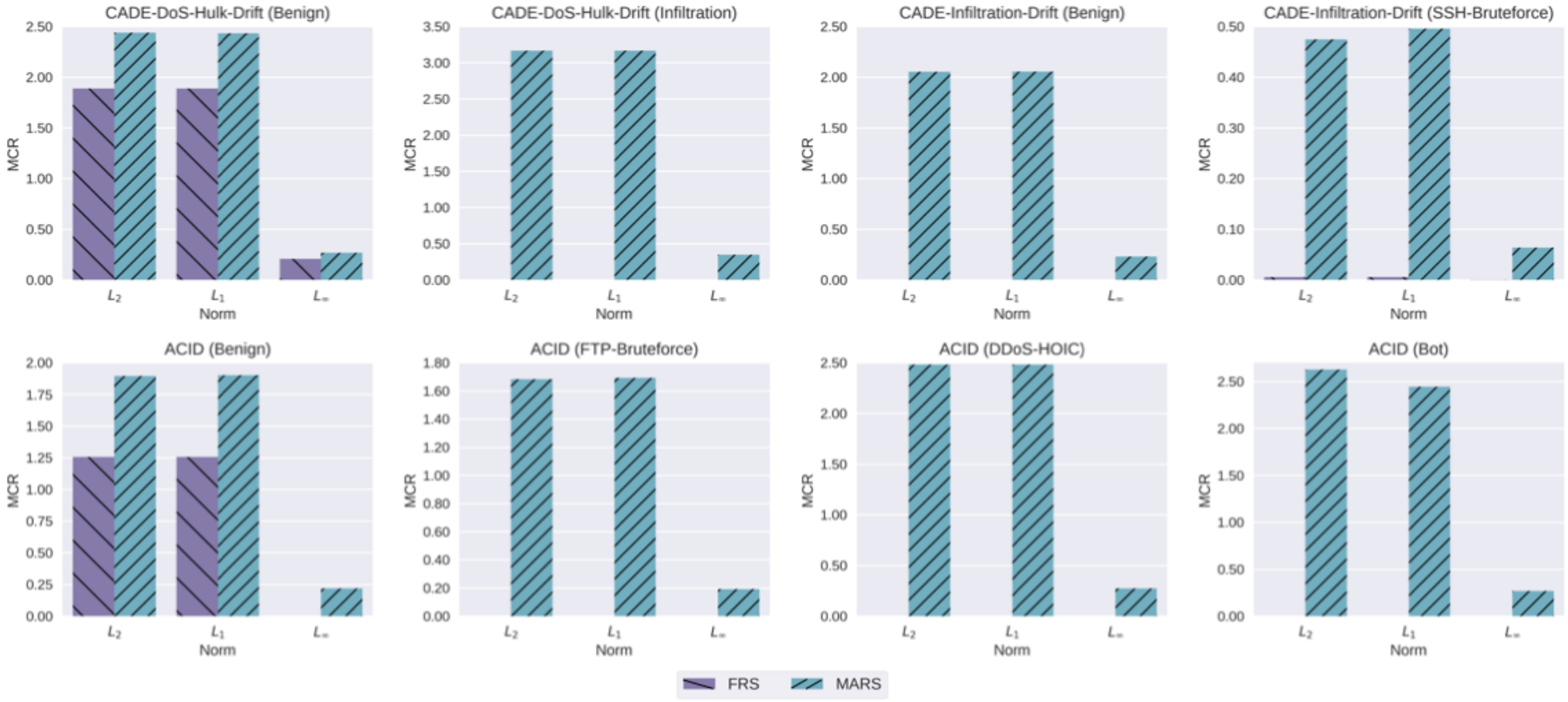


Figure 4.8 Comparison of l_p Mean Certified Radius (MCR) under the same smoothing distribution.

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

➢ 经验鲁棒性比较: 针对不同对抗攻击

Table 4.7 Comparison of empirical robustness of ACID against different adversarial attacks

Metric		Clean Accuracy	Robust Accuracy/Recall on Adversarial Examples			
Method	Seed	Clean	l_2-PGD	l_∞-PGD	l_1-PGD	l_1-EAD
Vanilla	42	1.0000	0.8395	0.5501	1.0000	0.0032
	43	1.0000	0.8395	0.5502	1.0000	0.0016
	44	1.0000	0.8395	0.5502	1.0000	0.0032
	mean±std	1.0000±0.0000	0.8395±0.0000	0.5502±0.0001	1.0000±0.0000	0.0027±0.0009
BARS [78]	42	1.0000	0.9601	0.8154	1.0000	0.0016
	43	1.0000	0.9601	0.8190	1.0000	0.0017
	44	1.0000	0.9610	0.8190	1.0000	0.0016
	mean±std	1.0000±0.0000	0.9604±0.0005	0.8178±0.0020	1.0000±0.0000	0.0016±0.0001
MARS	42	1.0000	0.9779	0.8925	1.0000	0.1031
	43	1.0000	0.9784	0.8863	1.0000	0.1021
	44	1.0000	0.9759	0.8898	1.0000	0.1031
	mean±std	1.0000±0.0000	**0.9774±0.0013**	**0.8895±0.0031**	1.0000±0.0000	**0.1028±0.0006**

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

➢ 经验鲁棒性比较: 针对不同强度自然损坏

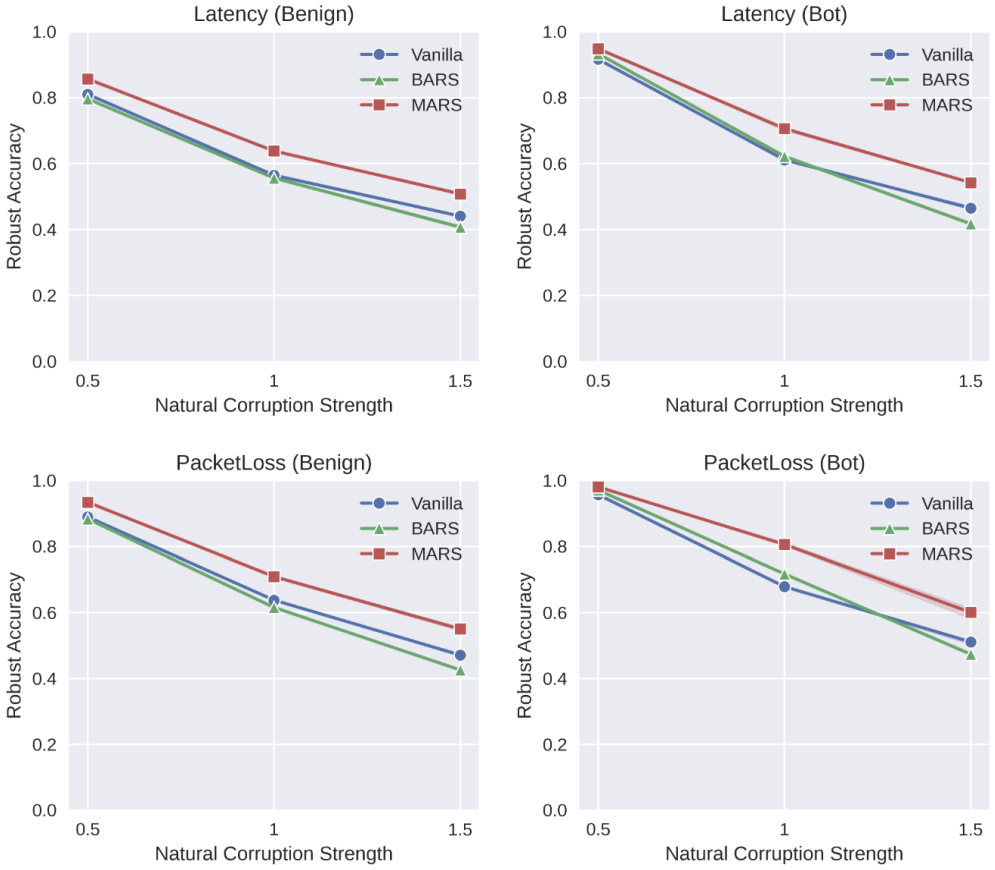


Figure 4.10 Comparison of empirical robustness of ACID against varied natural corruptions.

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

- ➢ 针对具有相似网络入侵类别的细粒度分类任务
  - DoS-GoldenEye
  - DoS-SLowHTTPTest
  - DoS-LOICHTTP

- ○ l_2, l_1, l_∞ 鲁棒性保证：比较平均认证半径 (MCR)

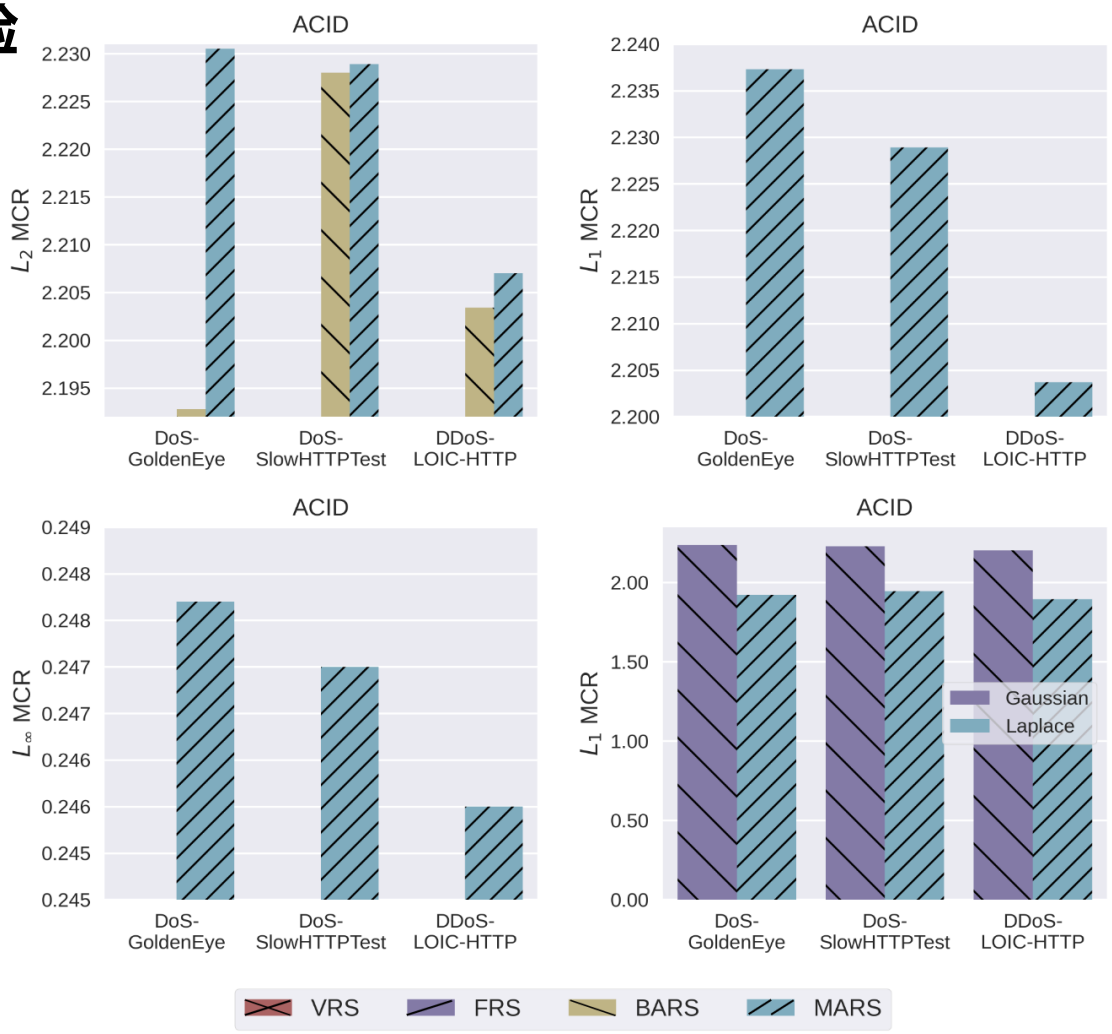


Figure 4.11 Comparison of l_p Mean Certified Radius (MCR) in fine-grained intrusion detection

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 横向对比实验

- ➢ 针对相似入侵类别的细粒度分类：经验鲁棒性比较
- ○ 针对不同自然损坏

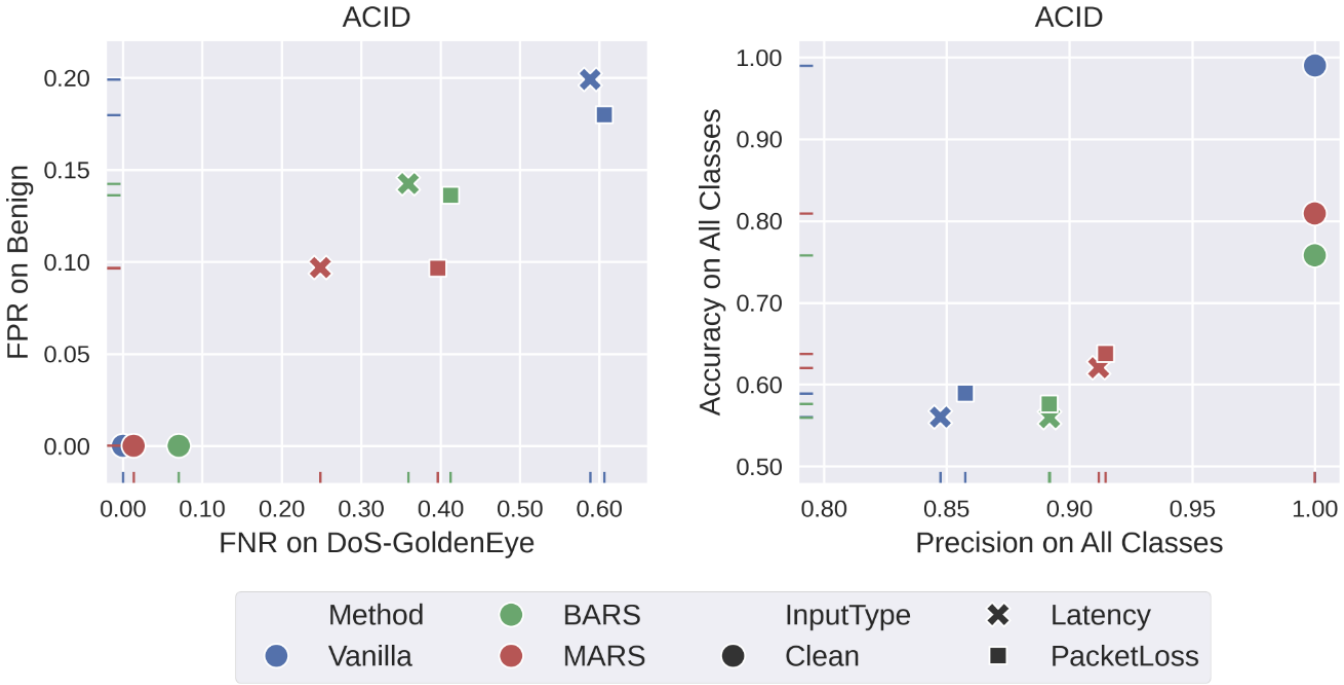


Figure 4.12 Comparison of empirical robustness of ACID against natural corruptions in fine-grained detection of similar intrusions.

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

实验评估 – 纵向对比实验

- 维度级可验证半径比较
  - Last-5 鲁棒特征
    - 最敏感的五个特征维度
    - 半径越小，表示对模型的敏感度和重要性越高，因为扰动这些特征更有可能改变模型的预测
  - Top-5 鲁棒特征
    - 最鲁棒的五个特征维度

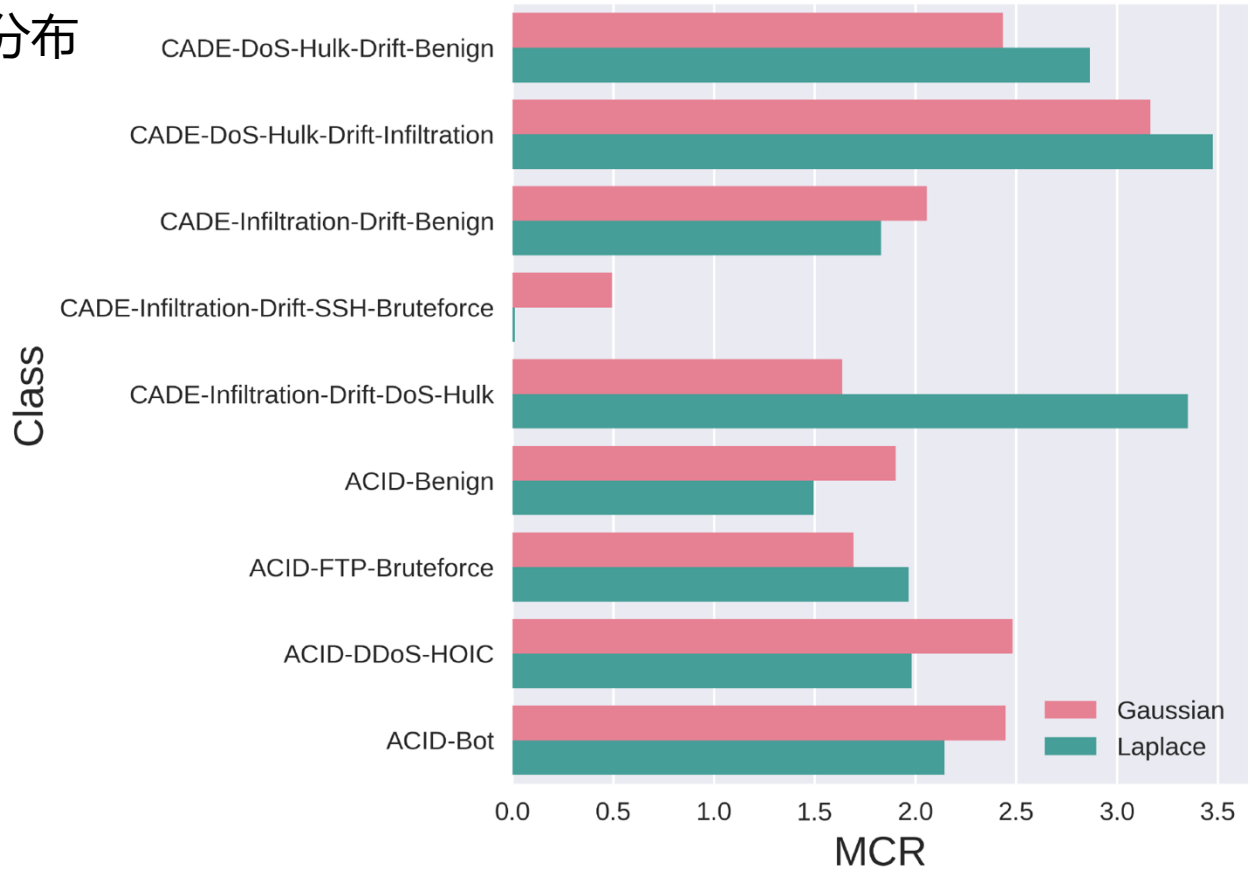
Table 4.12 Sensitive and robust features on DoS-GoldenEye

No	Radius	FeatureName	Description
24	0.0426	Flow_IAT_Std	Standard deviation time two flows.
20	0.0433	Bwd_Packet_Length_Std	Standard deviation size of packet in backward direction.
79	0.0488	Active_Std	Standard deviation time a flow was active before becoming idle.
72	0.0569	Init_Win_bytes_forward	Number of bytes sent in initial window in the forward direction.
78	0.0576	Active_Max	Maximum time a flow was active before becoming idle.
8	10.0741	Flow_Duration	Flow duration.
39	10.9644	Fwd_URG_Flag	Number of times URG flag was set in packets travelling in the forward direction (0 for UDP).
52	11.2367	RST_Flag_Count	Number of packets with RST.
38	11.3300	Bwd_PSH_Flag	Number of times PSH flag was set in packets travelling in the backward direction (0 for UDP).
13	11.4358	Fwd_Packet_Length_Min	Minimum size of packet in forward direction.
All	2.2305	MCR	Mean certified radius per class.

基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

□ 实验评估 – 纵向对比实验

- ➢ l_1鲁棒性保证：平均认证半径 (MCR) 比较
  - ○ 不同平滑分布



基于多阶自适应随机平滑的深度神经网络鲁棒性验证方案

实验评估 – 纵向对比实验

- 针对相似入侵类别的细粒度分类：
  - l_p鲁棒半径验证
  - 不同可验证防御方法
  - 平均认证半径 (MCR)
  - 平均验证时间 (ACT) (每个样本/秒)

Table 4.9 Comparison of l_2 MCR and Average Certification Time (ACT) (per sample/sec) in fine-grained detection of similar intrusions.

Method	Benign	DoS-GoldenEye	DoS-SlowHTTPTest	DDoS-LOIC-HTTP
VRS [75]	0.1950 (0.0028)	0.0001 (0.0039)	0.0000 (0.0053)	0.0000 (0.0033)
FRS [77]	1.2077 (0.0238)	0.0024 (0.0309)	0.0014 (0.0459)	0.0000 (0.0269)
BARS [78]	1.9036 (0.0029)	2.1928 (0.0040)	2.2280 (0.0051)	2.2034 (0.0033)
MARS	**1.9260 (0.0253)**	**2.2305 (0.0331)**	**2.2289 (0.0361)**	**2.2070 (0.0279)**

Table 4.10 Comparison of l_p MCR and Average Certification Time (ACT) (per sample/sec) in fine-grained detection of similar intrusions under the same smoothing distribution.

Norm	Method	Benign	DoS-GoldenEye	DoS-SlowHTTPTest	DDoS-LOIC-HTTP
l_2	FRS [77]	1.2077 (0.0238)	0.0024 (0.0309)	0.0014 (0.0459)	0.0000 (0.0269)
	MARS	**1.9260 (0.0253)**	**2.2305 (0.0331)**	**2.2289 (0.0361)**	**2.2070 (0.0279)**
l_1	FRS [77]	1.2077 (0.0236)	0.0000 (0.0311)	0.0000 (0.0365)	0.0000 (0.0272)
	MARS	**1.9317 (0.0256)**	**2.2373 (0.0334)**	**2.2289 (0.0378)**	**2.2037 (0.0283)**
l_∞	FRS [77]	0.0000 (0.0240)	0.0000 (0.0321)	0.0000 (0.0377)	0.0000 (0.0279)
	MARS	**0.2196 (0.0263)**	**0.2482 (0.0336)**	**0.2475 (0.0383)**	**0.2460 (0.0293)**

结论

- 提出了一个基于多阶信息的自适应随机平滑算法，利用平滑分类器的零阶输出和一阶梯度信息，搜索多种范数度量下的鲁棒半径，获得了比现有方法更紧的深度神经网络对抗鲁棒性下界。
- 设计了一种基于特征敏感性的逐维鲁棒半径度量算法，通过量化输入特征各维度的鲁棒性权重计算维度级鲁棒半径，实现了适用于具有异构特征的输入样本的细粒度对抗鲁棒性验证。
- 本文在多种基于深度神经网络的网络入侵检测模型和数据集上对所提的鲁棒性验证方案进行了实验评估。结果表明，所提方法在更大的l_p范数约束的扰动区域内成功验证了模型的对抗鲁棒性，增强了模型针对多种对抗攻击和自然损坏的鲁棒性。

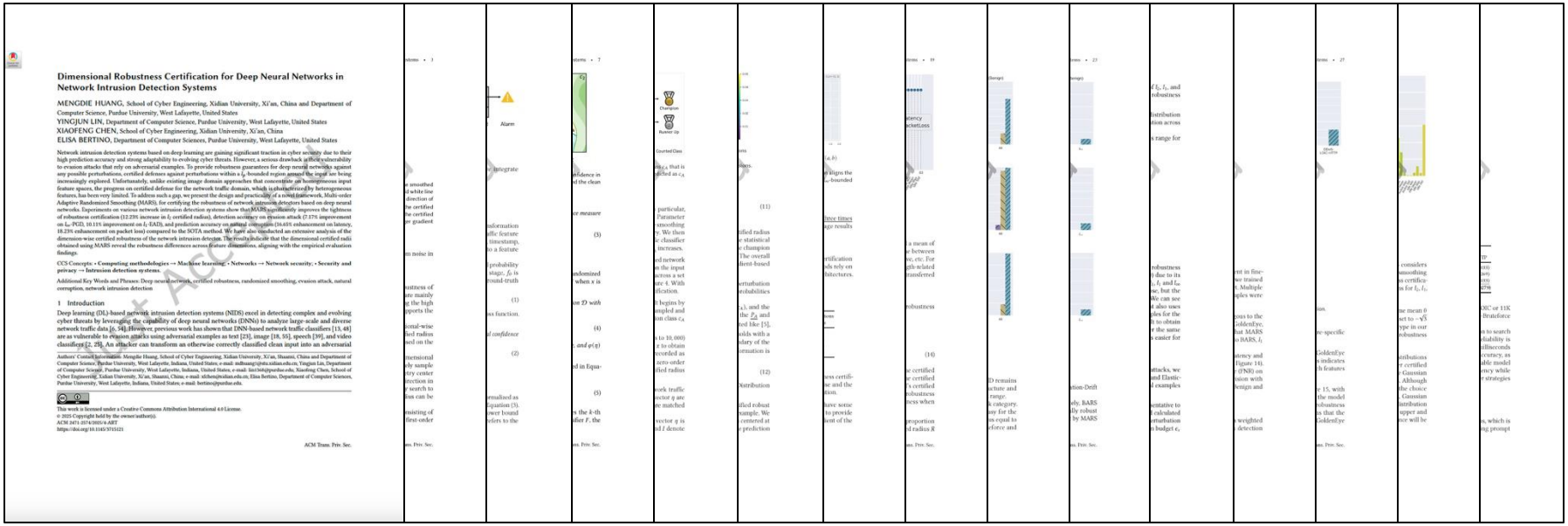
□ 主要成果

ACM Transactions on Privacy and Security

主要成果已发表在CCF推荐网络与信息安全 B 类、中科院JCR 三区 期刊

ACM Transactions on Privacy and Security (TOPS)

➢ **Mengdie Huang**, Yingjun Lin, **Xiaofeng Chen**, Elisa Bertino. Dimensional Robustness Certification for Deep Neural Networks in Network Intrusion Detection Systems [J]. *ACM Transactions on Privacy and Security (TOPS)*, 2025, 1-33. （1类贡献度）


Dimensional Robustness Certification for Deep Neural Networks in Network Intrusion Detection Systems

MENGDIE HUANG, School of Cyber Engineering, Xidian University, Xi'an, China and Department of Computer Science, Purdue University, West Lafayette, United States
YINGJUN LIN, Department of Computer Science, Purdue University, West Lafayette, United States
XIAOFENG CHEN, School of Cyber Engineering, Xidian University, Xi'an, China
ELISA BERTINO, Department of Computer Sciences, Purdue University, West Lafayette, United States

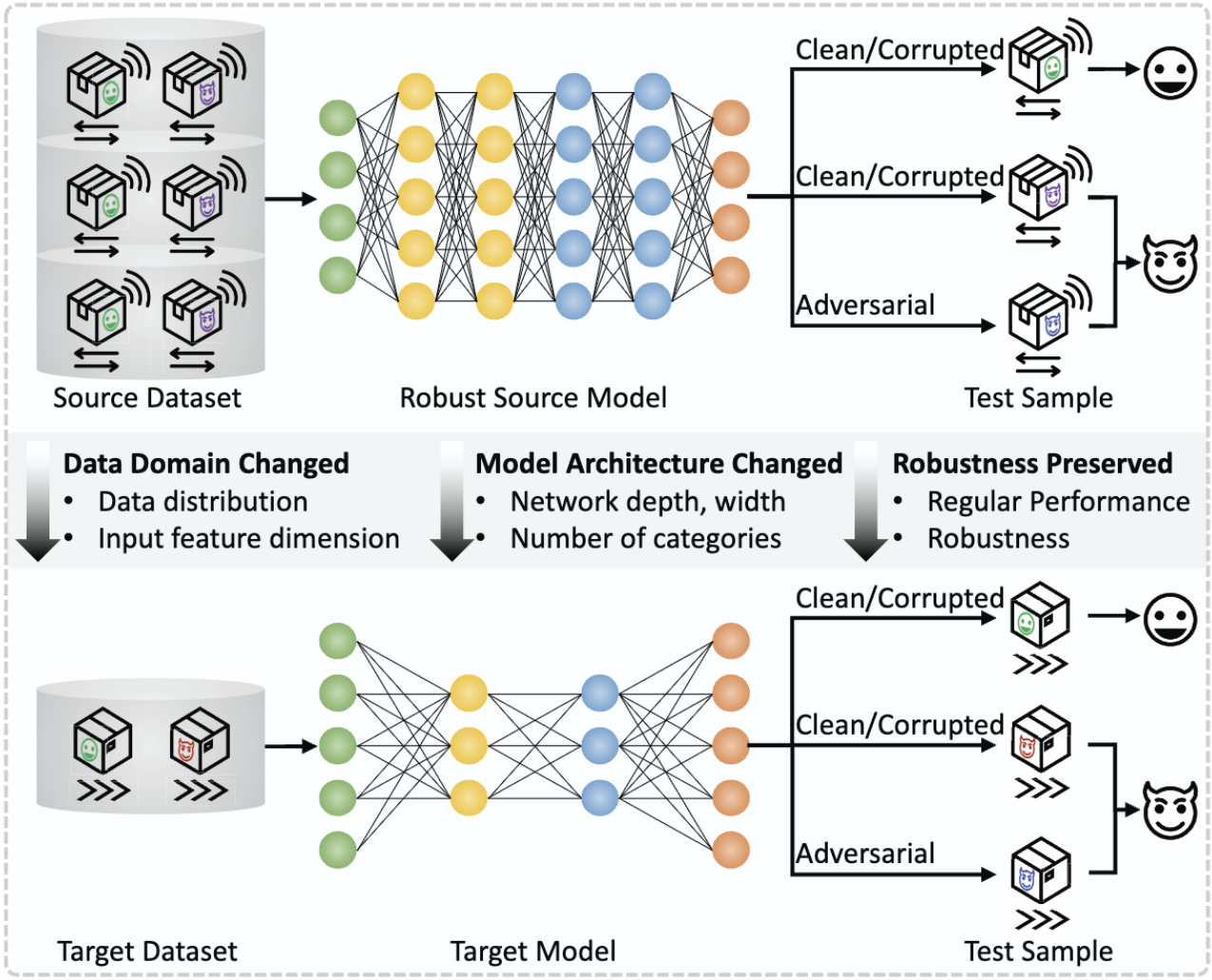
目 录

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 多种迁移场景

跨数据域迁移

跨模型迁移

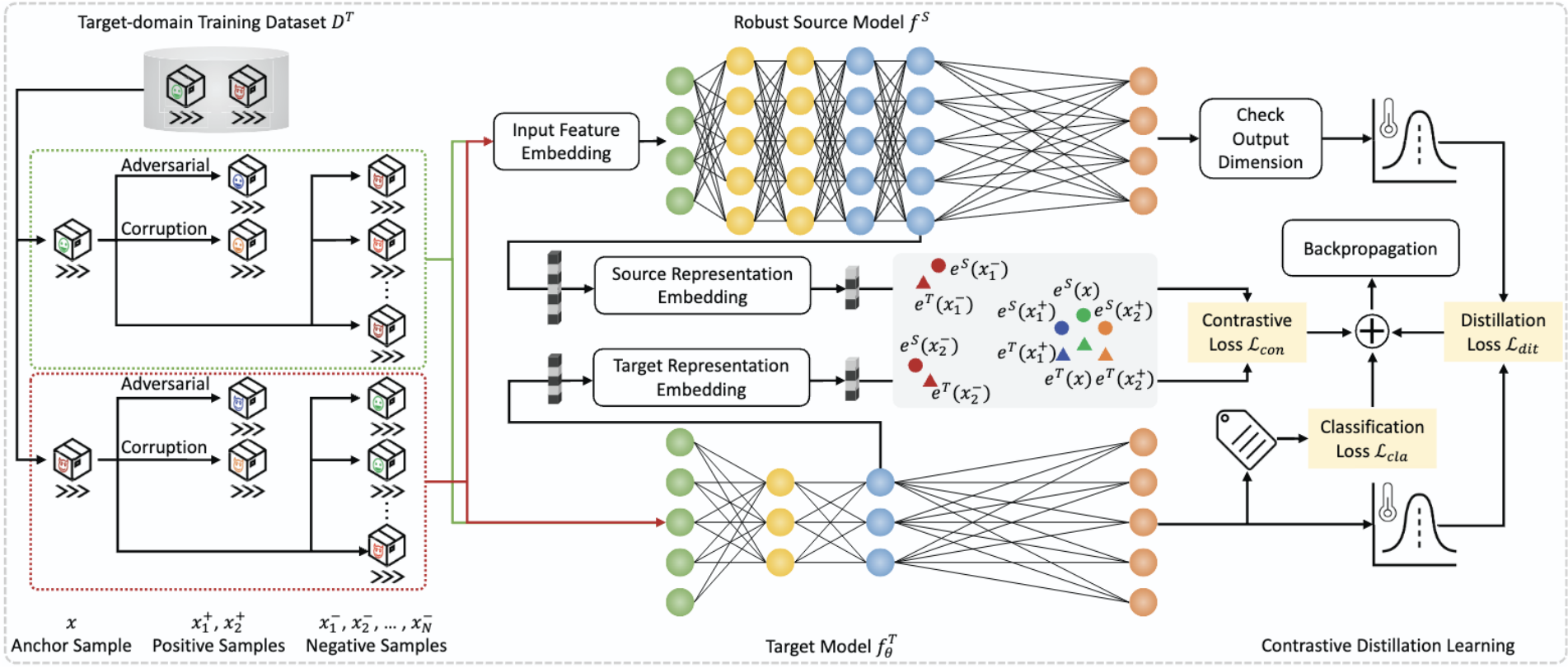


常规性能

鲁棒性能

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ Contrastive Adversarial Representation Distillation (CARD) 框架



(a)鲁棒性感知视图构建　(b)自适应维度对齐　(c)对比表征蒸馏学习

研究适用多种迁移场景的深度神经网络的对抗鲁棒性迁移技术具有重要意义

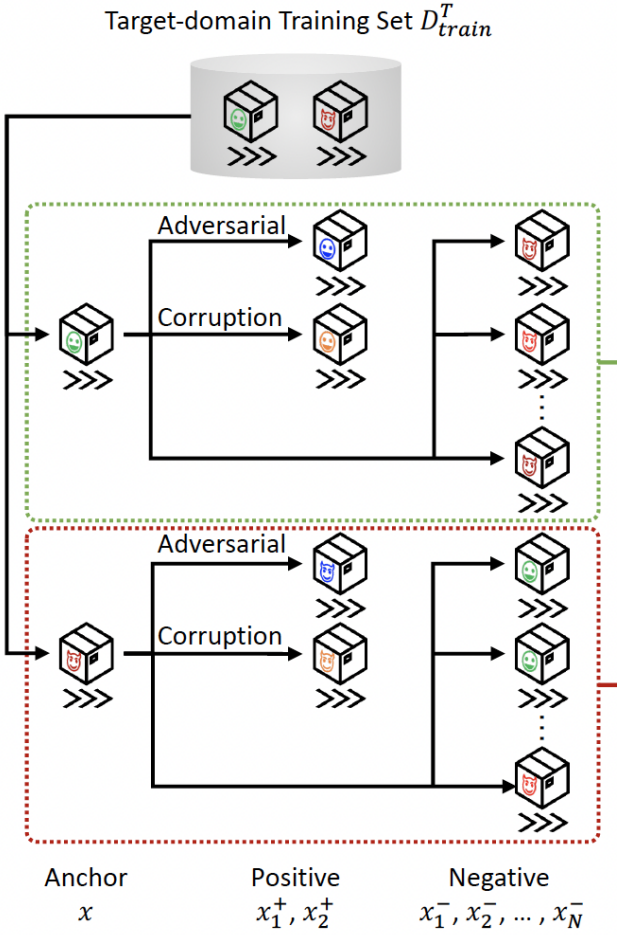
基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 阶段一：鲁棒性感知对比视图构建

- ➢ 对比学习的目标
  - ○ 锚样本： x
  - ○ 正样本对 (Positive Pair)： (x, x^+)

    鼓励对相似的样本学习到相近的特征表示
  - ○ 负样本对 (Negative Pair): (x, x^-)

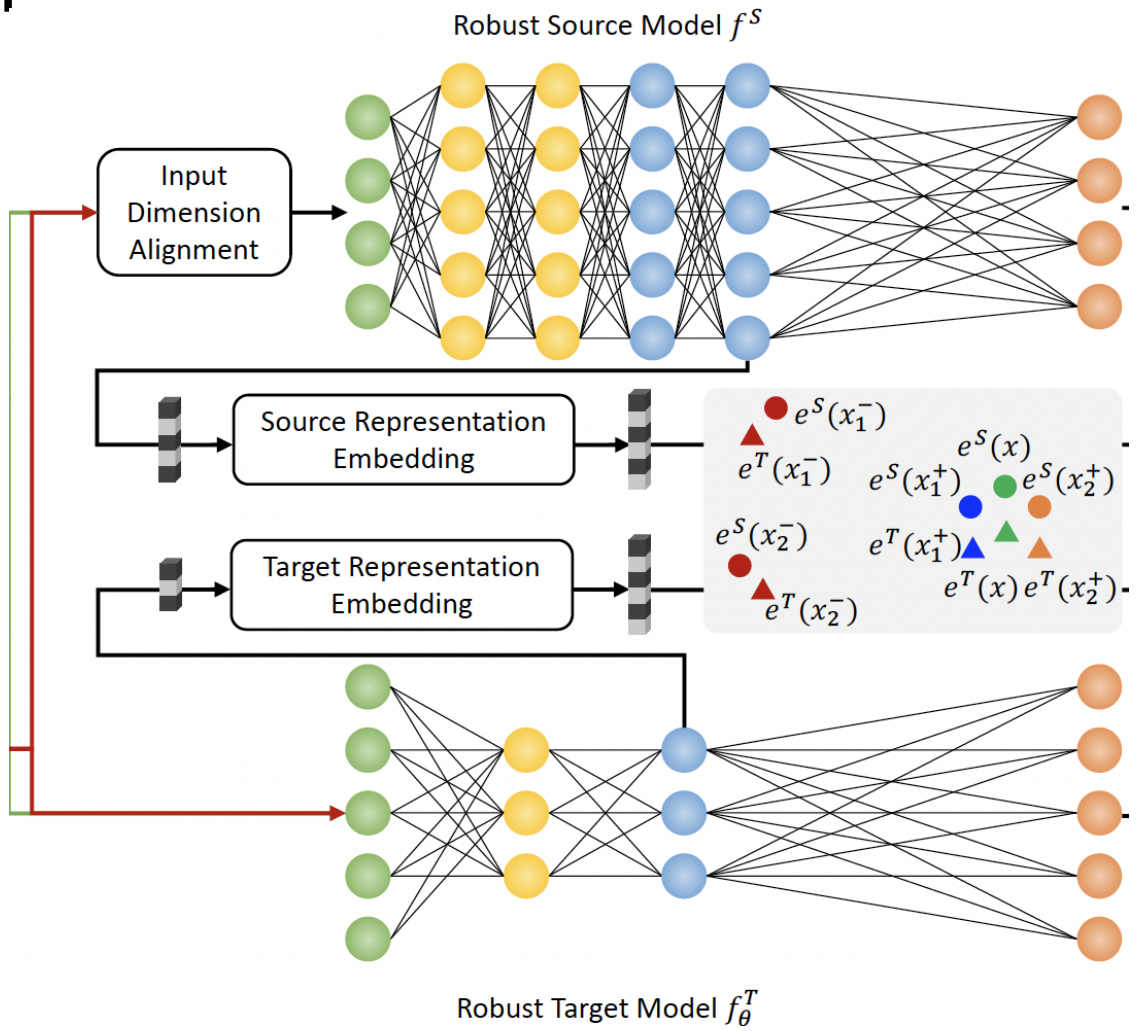
    鼓励对不相似的样本学习到远离的特征表示
- ➢ 常规对比学习的正视图 x^+ 来源
  - ○ 旋转或翻转等数据扩增
  - ○ 鲁棒性感知对比学习
  - ○ 正视角一：对抗扰动 $(x, x_1^+) = (x, x^*) = x, x + \delta$
  - ○ 正视角二：自然损坏 $(x, x_2^+) = (x, \tilde{x}) = x, x + \delta$



基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

阶段二：自适应维度对齐

- 跨数据域迁移学习
  - 源域数据集输入空间维度与目标域数据集输入空间维度不匹配
  - 输入空间维度对齐
- 跨模型迁移学习
  - 源域模型隐藏层表征维度与目标域模型隐藏层表征维度不匹配
  - 隐层表征空间维度对齐



基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 阶段三：对比表征蒸馏学习

- 对比损失：
  - 目标域模型学习源域模型鲁棒表征能力

$$\mathcal{L}_{con} = \mathcal{L}^T_{con}(e^S, e^T, e^{+T}, e^{-T}) + \mathcal{L}^S_{con}(e^T, e^S, e^{+S}, e^{-S})$$

- 蒸馏损失 (在仅跨模型的迁移中生效)
  - 目标域模型学习源域输入类比分布

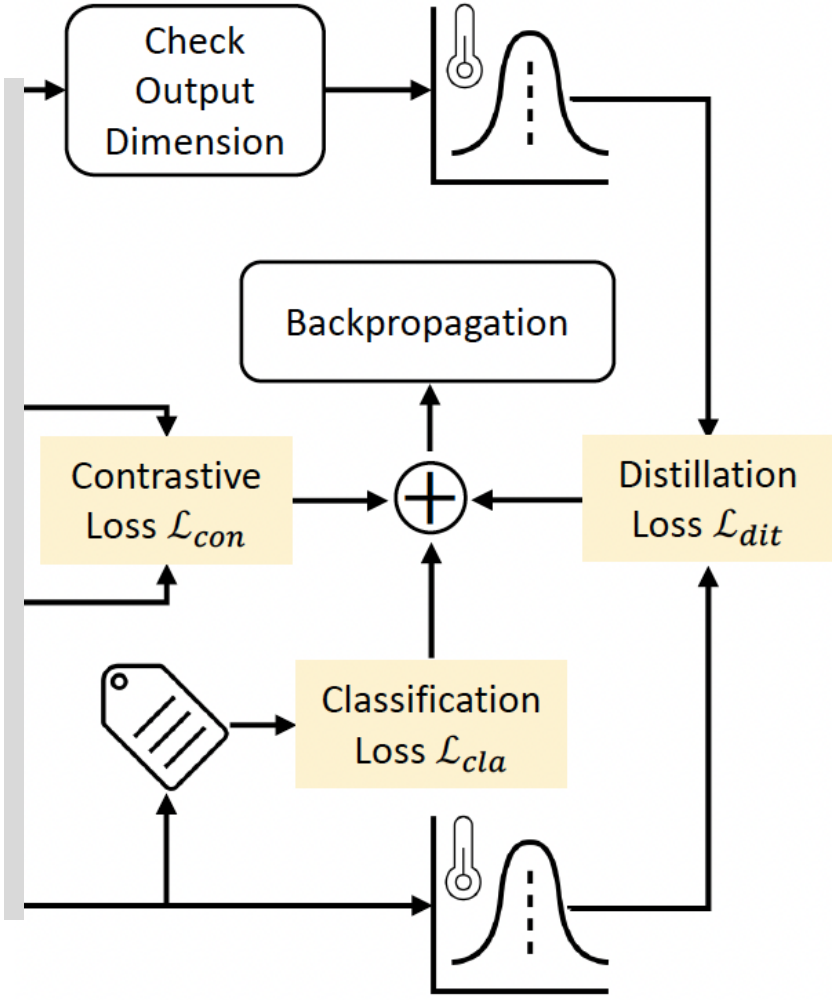
$$\mathcal{L}_{dit}(o^T, o^S, o^T_{adv}, o^S_{adv}) = \mathrm{KL}(o^{S\tau}|o^{T\tau}) + \mathrm{KL}(o^{S\tau}|o^{T\tau}_{adv}) + \mathrm{KL}(o^{S\tau}_{adv}|o^{T\tau}_{adv}))$$

- 分类损失
  - 目标域模型自主学习干净样本和对抗样本

$$\mathcal{L}_{cla}(o^T, y_{true}) = \mathcal{L}_{CE}(o^T, y_{true}) + \mathcal{L}_{CE}(o^T_{adv}, y_{true})$$

- 综合损失

$$\min_{\theta,\vartheta^I,\vartheta^S,\vartheta^T} \mathcal{L}_{card} = \alpha \cdot \mathcal{L}_{con} + \beta \cdot \mathcal{L}_{dit} + \gamma \cdot \mathcal{L}_{cla}$$



基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

口 实验评估 – 实验设置

➢ 环境

○ PyTorch 2.2.0、CUDA V 12.1、NVIDIA GeForce RTX 3090 GPU

➢ 模型

Table 5.1 Model architecture and parameter information

Role	Architecture	Block Type	Total Params	Forward Size	Params Size	Total Size
Source Model	WideResNet-34-10	Residual	46,159,545	7.38M	176.08M	183.47M
Target Model	ResNet-18	Residual	11,172,297	1.29M	42.62M	43.91M
Target Model	MobileNet	Separable	3,215,625	1.69M	12.27M	13.95M

➢ 数据集

○ 源域数据集 (2)：UNSW-NB15 (NoExploit)， NSL-KDD (All).

○ 目标域数据集 (2)：UNSW-NB15 (WithExploit)，UNSW-NB15 (All).

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

口 实验评估 – 实验设置

➢ 数据集

Table 5.2 Dataset information

Role	Name	ClassNum	Percentage of Target Train	Num of Limited Train-Benign	Num of Limited Train-Malicious	Num of Limited Train
Target Dataset	UNSW-NB15 (WithExploit)	2	5% 10%~50%	2595 5189~25945	978 1956~9776	3573 7145~35721
Target Dataset	UNSW-NB15 (All)	10	5% 10%~50%	2595 5189~25945	2704 5408~27040	5299 10597~52985
Source Dataset	UNSW-NB15 (NoExploit)	9	5% 10%~50%	2595 5190~25950	1727 3454~17270	4322 8644~43220
Source Dataset	NSL-KDD (All)	5	5%	3368	2931	6299

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

口 实验评估 – 实验设置

➢ 评估指标

- 准确率 (Acc)、F1 分数 、召回率、精确率 (Precis)、假正率 (FPR) 和假负率 (FNR)。
- 在干净样本（e.g. CleAcc）、对抗样本（ e.g. AdvAcc）和损坏样本（ e.g. CorAcc）上的结果分别揭示了目标模型的干净性能、对抗鲁棒性和自然鲁棒性。
- 每个报告值代表三个随机种子（41，42，43）设定下的平均结果。

➢ 攻击配置

- 对对抗攻击的评估包括三部分：A(结合了 DeepFool 和 Auto-PGD) , 自适应攻击。

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

口 实验评估 – 实验设置

➢ 对比方法

○ 标准微调FT， 知识蒸馏KD；对抗性微调：FRFE, TWINS；对抗性蒸馏：VAD, AAD.

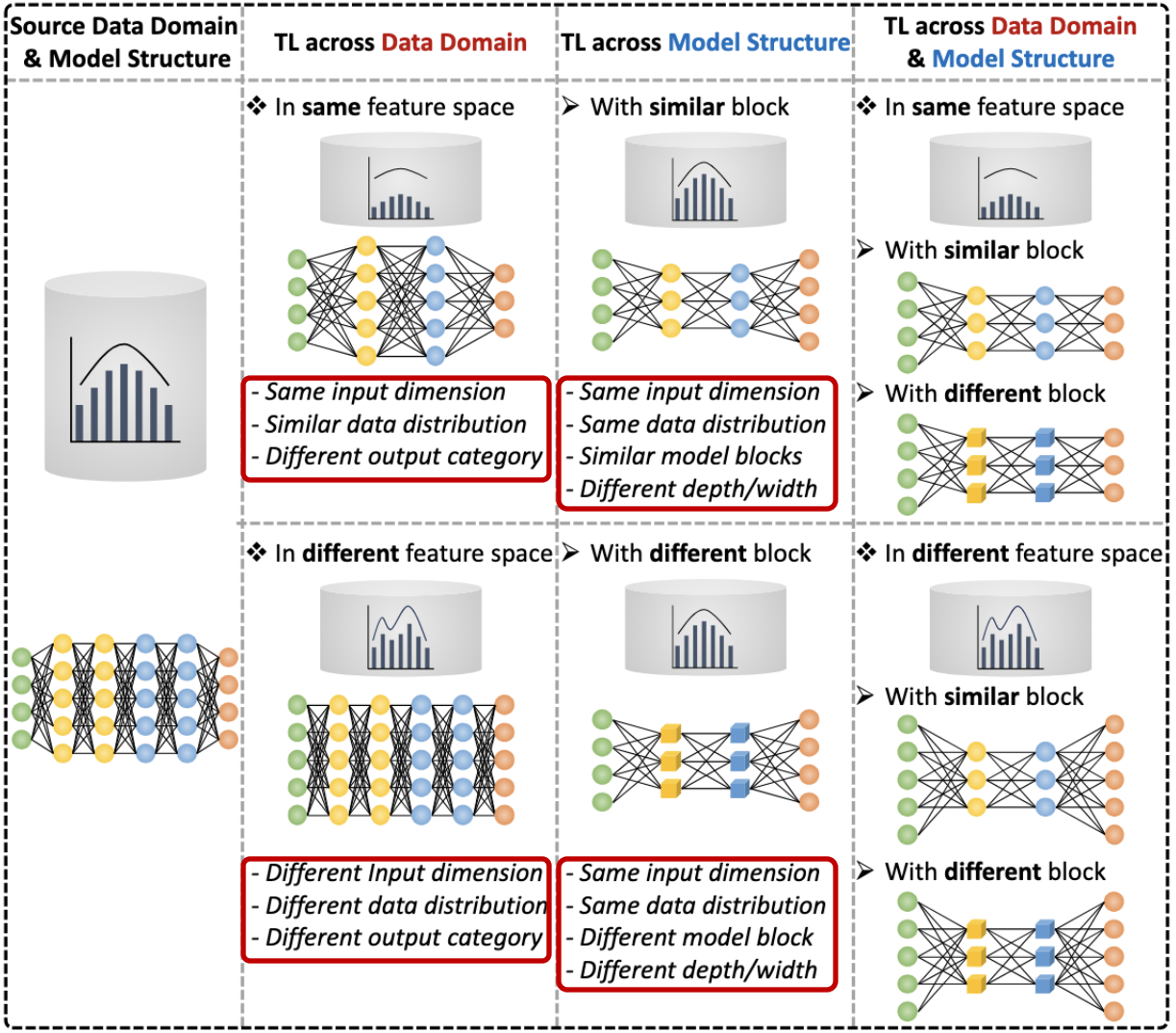
Table 5.3 Comparison of standard/robustness-preserving Transfer Learning (TL) methods

Transfer Method	Robustness Transferring	Designed for NID	Evaluate Binary Classification	Evaluate Multi-Class Classification	Differences between Source and Target Tasks: Same Model Different Domains	Different Models Same Domain	Different Models Different Domains
FT [146]	○	●	●	○	●	○	○
KD [149]	○	●	●	○	○	●	○
FRFE [81]	●	○	○	●	●	○	○
TWINS [84]	●	○	○	●	●	○	○
VAD [87]	●	○	○	●	○	●	○
AAD [93]	●	○	○	●	○	●	○
CARD	●	●	●	●	●	●	●

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 实验评估 – 实验设置

- ➢ 迁移场景划分
  - ○ 目标域任务与源域任务的数据域相似性
  - ○ 目标域任务与源域任务的模型相似性



基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

实验评估 – 实验设置

- 迁移场景划分

  - 相同输入特征空间的鲁棒性迁移
    - 目标域NID数据集与源域NID数据集的输入特征空间相同，即输入特征定义和维度数相同。对于此场景，使用 UNSW-NB15 (NoExploit) 作为源域数据集，使用UNSW-NB15 (WithExploit) 作为目标域数据集。

  - 不同输入特征空间的鲁棒性迁移
    - 目标域NID数据集与源域NID数据集的输入特征空间不同，即输入特征定义和维度数均不同。对于此场景，使用 NSL-KDD (All) 作为源域数据集，使用UNSW-NB15 (All) 作为目标域数据集。

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

实验评估 – 实验设置

- 迁移场景划分

- 具有相同基础块结构的模型间的鲁棒性迁移
  - 目标模型与源域模型的基础块结构相同。对于此场景，使用 WideResNet-34-10作为源模型，使用ResNet-18作为目标模型。

- 具有不同基础块结构的模型间的鲁棒性迁移
  - 目标模型与源域模型的基础块结构不同。对于此场景，使用 WideResNet-34-10作为源模型，使用MobileNet作为目标模型。

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 实验评估 – 实验设置

- ➢ 迁移场景划分
  - ○ 相同输入特征空间的迁移 UNSW-NB15 (NoExploit) → UNSW-NB15 (WithExploit)
    - 目标域与源域数据集的输入特征空间相同，即特征定义和维度数都相同。
  - ○ 不同输入特征空间的迁移NSL-KDD (All) → UNSW-NB15 (All)
    - 目标域与源域NID数据集的输入特征空间不同，即特征定义和维度数均不同。
- ➢ 网络流量分类 (入侵检测粒度)划分
  - ○ 多分类模型间迁移
    - 9 分类UNSW-NB15 (NoExploit) → 2分类UNSW-NB15 (WithExploit)
    - 5分类NSL-KDD (All) → 10 分类UNSW-NB15 (All)
  - ○ 二分类模型间迁移
    - 2 分类UNSW-NB15 (NoExploit) →2 分类UNSW-NB15 (WithExploit)
    - 2 分类NSL-KDD (All) → 2 分类UNSW-NB15 (All)

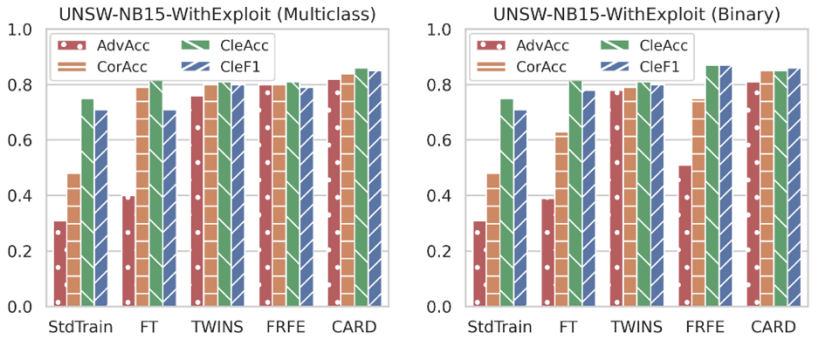
基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%

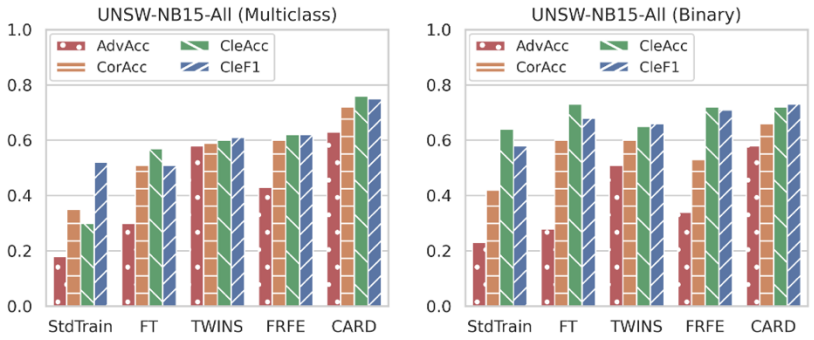
实验评估 – 横向对比实验

- 针对跨数据域的迁移学习任务：与 SOTA 微调技术的性能比较

综合性能对比



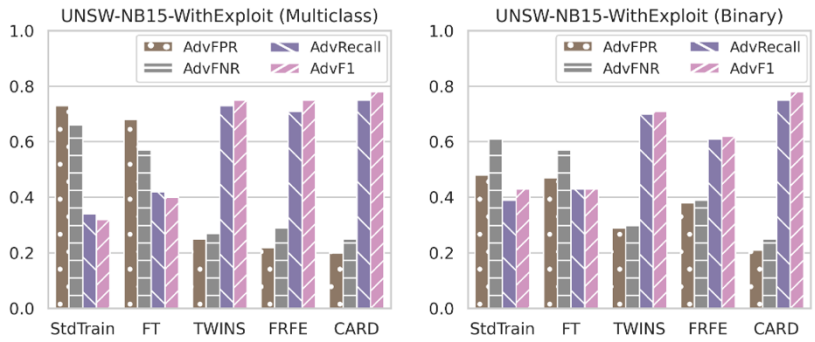
(a) TL across domains with the same input space.



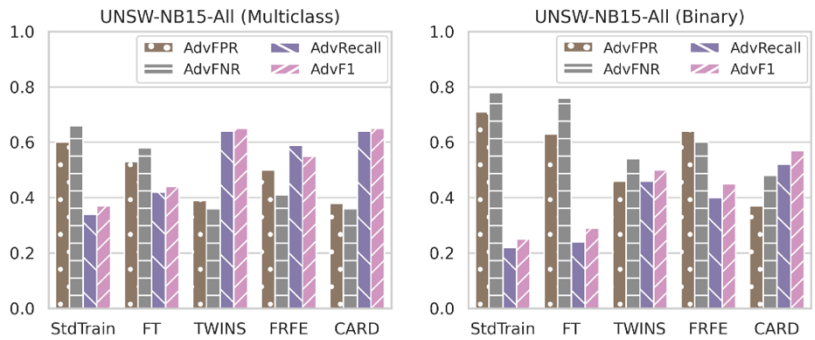
(b) TL across domains with different input spaces.

Figure 5.4 Comparison with SOTA fine-tuning methods on cross-domain TL.

对抗鲁棒性细粒度性能对比



(a) TL across domains with the same input space.



(b) TL across domains with different input spaces.

Figure 5.5 Adversarial comparison with SOTA fine-tuning methods on cross-data domain TL.

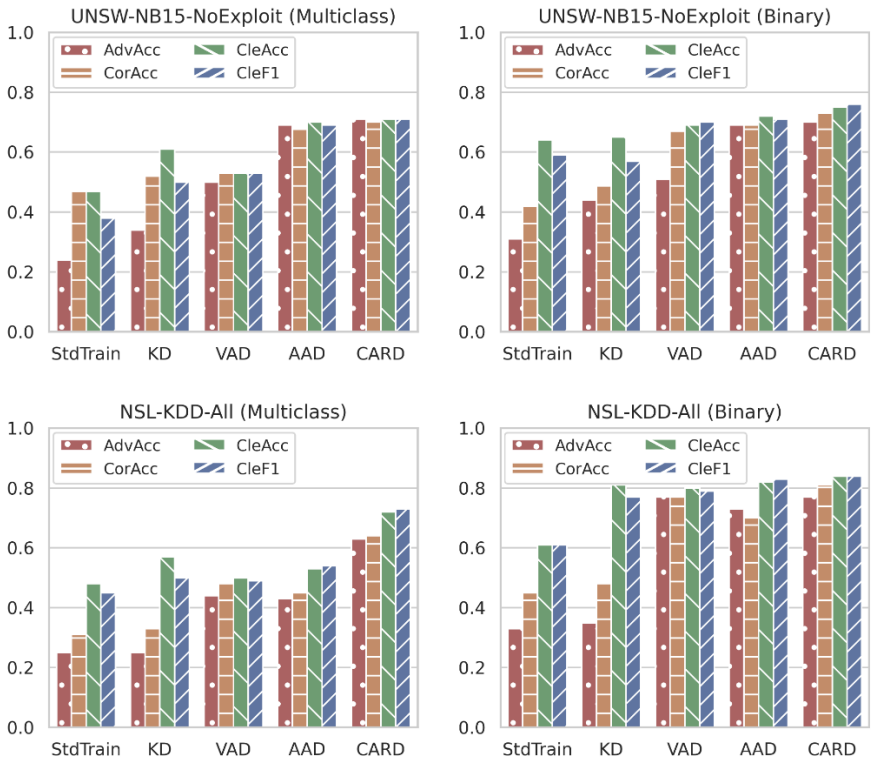
基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%

□ 实验评估 – 横向对比实验

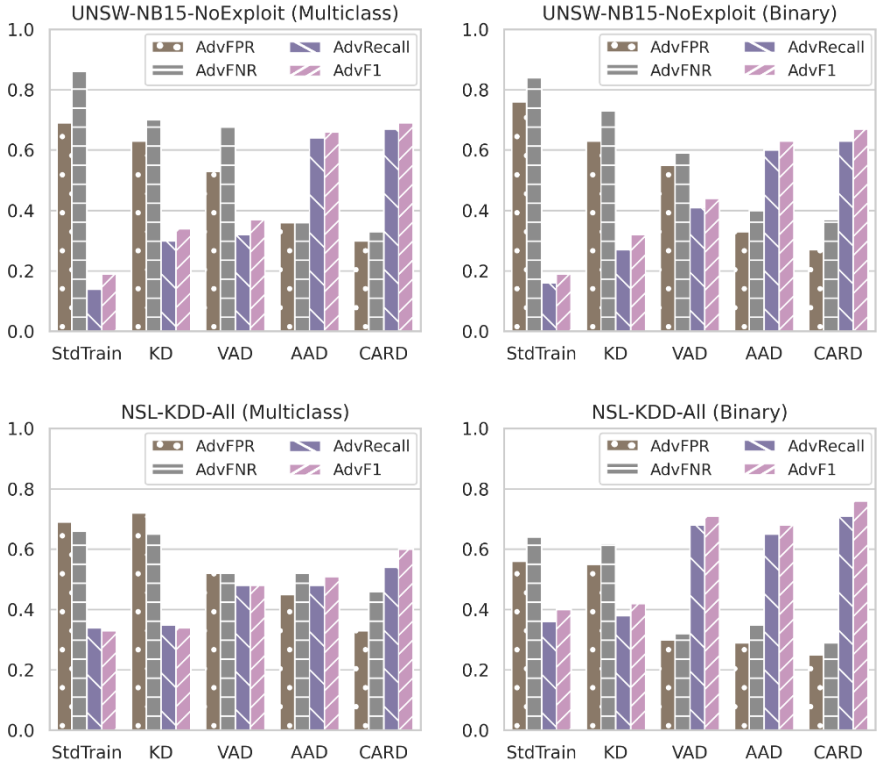
➢ 针对跨模型结构的迁移学习任务：与SOTA蒸馏技术的比较 （ResNet-18）

综合性能对比



(a) TL across models with similar building blocks.

对抗鲁棒性细粒度性能对比



(a) TL across models with similar building blocks.

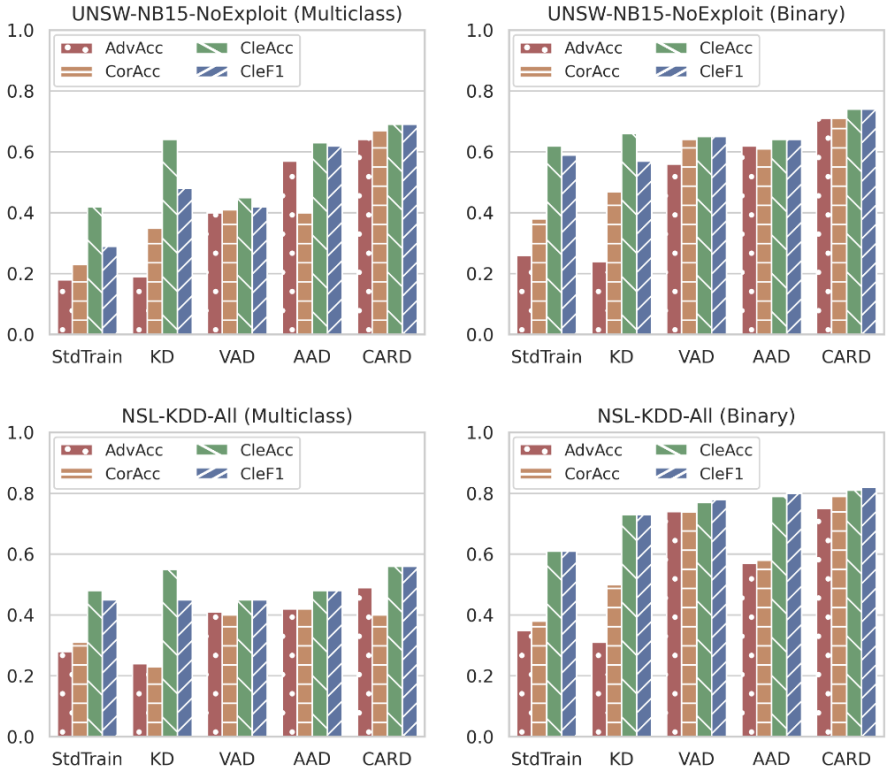
基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%

口 实验评估 – 横向对比实验

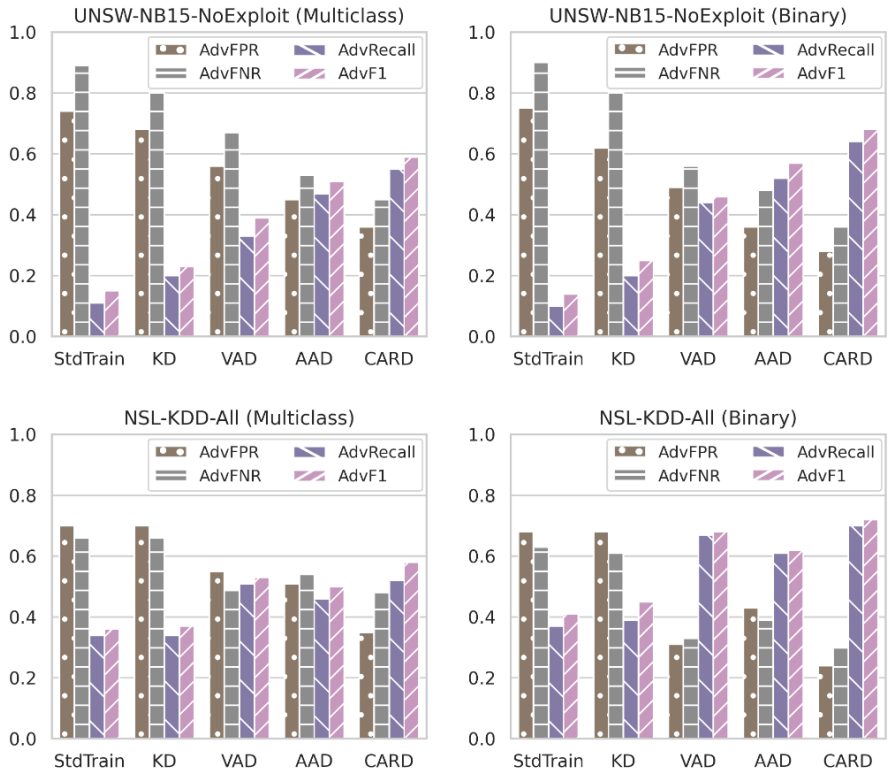
➢ 针对跨模型结构的迁移学习任务：与SOTA蒸馏技术的比较 （MobileNet）

综合性能对比



(b) TL across models with different building blocks.

对抗鲁棒性细粒度性能对比



(b) TL across models with different building blocks.

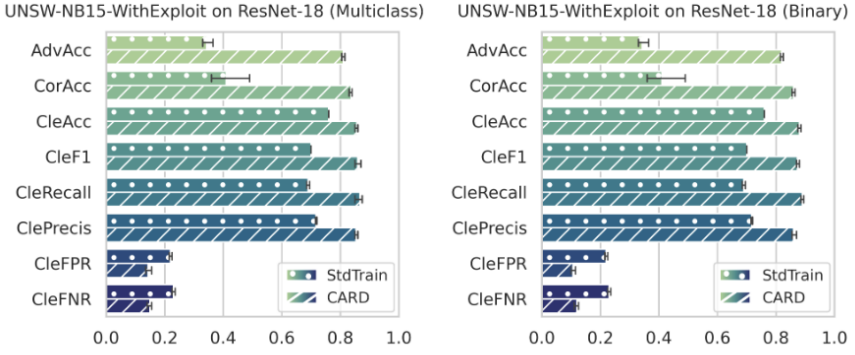
基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%

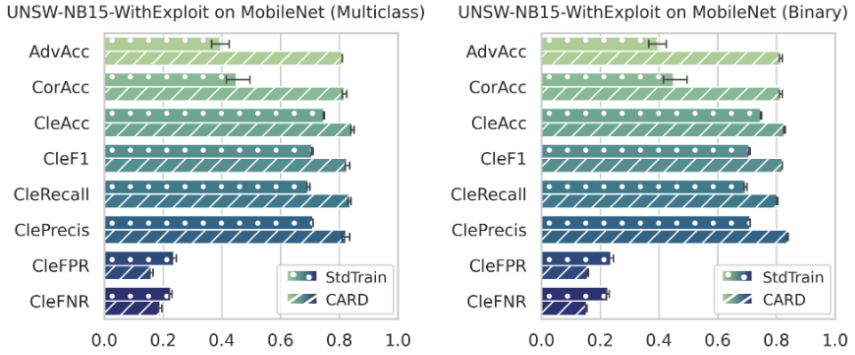
□ 实验评估 – 横向对比实验

➢ 针对跨数据域和模型结构的迁移学习任务：与标准从零训练技术的比较

综合性能对比

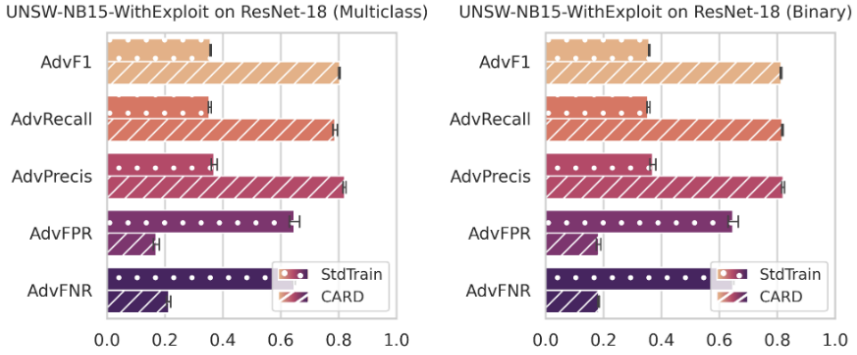


(a) TL across domains with same input space and models with similar blocks.

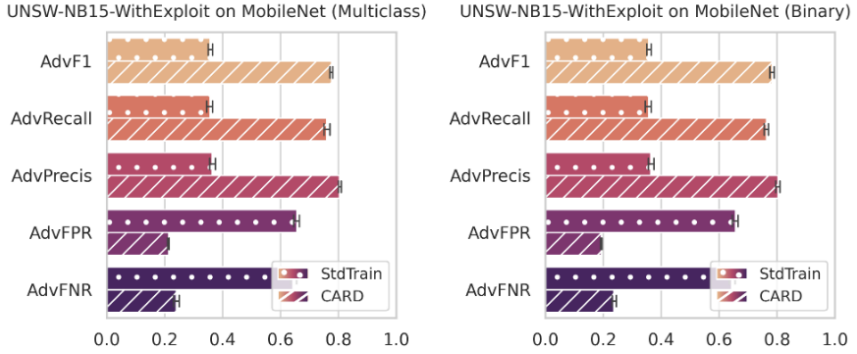


(b) TL across domains with same input space and models with different blocks.

对抗鲁棒性细粒度性能对比



(a) TL across domains with same input space and models with similar blocks.



(b) TL across domains with same input space and models with different blocks.

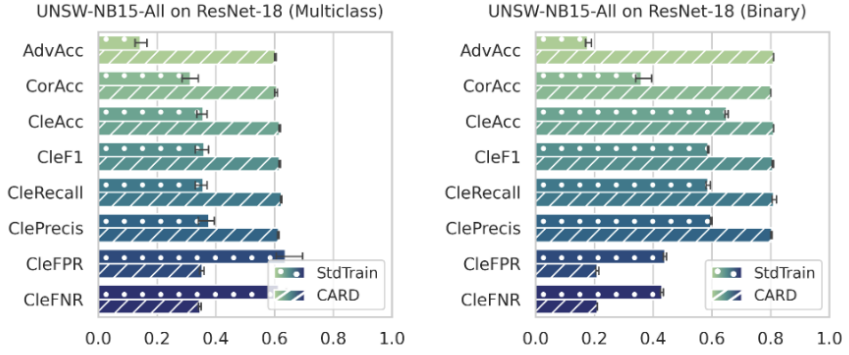
基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%

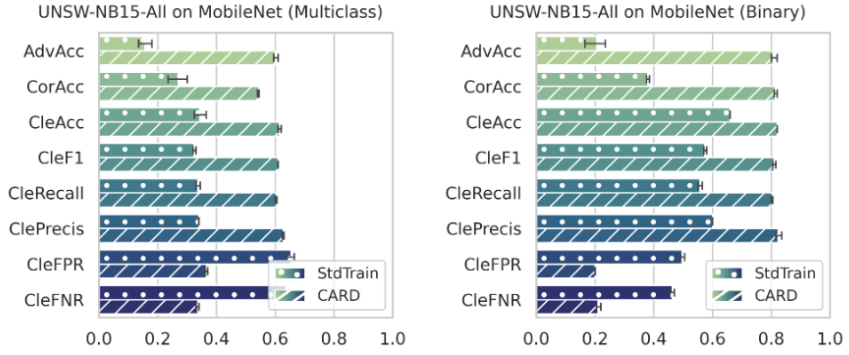
口 实验评估－横向对比实验

➢ 针对跨数据域和模型结构的迁移学习任务：与标准从零训练技术的比较

综合性能对比

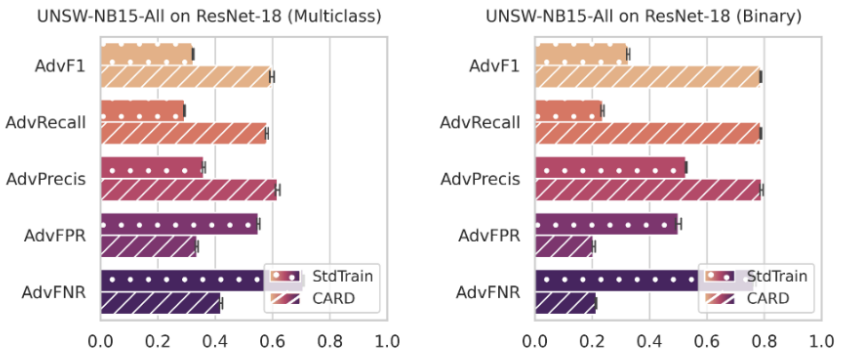


(c) TL across domains with different input spaces and models with similar blocks.

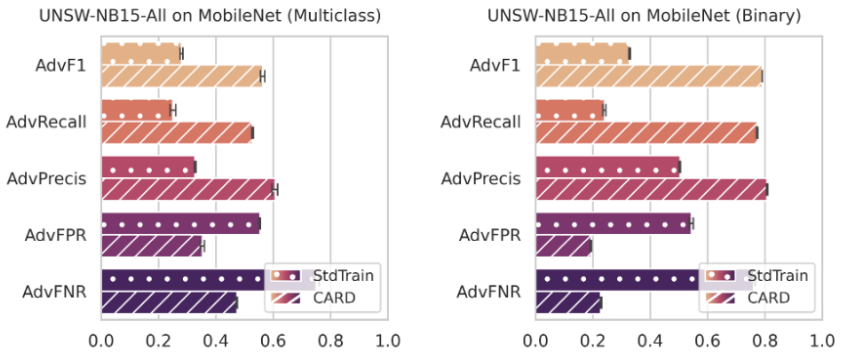


(d) TL across domains with different input spaces and models with different blocks.

对抗鲁棒性细粒度性能对比



(c) TL across domains with different input spaces and models with similar blocks.



(d) TL across domains with different input spaces and models with different blocks.

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%~50%

□ 实验评估 – 纵向对比实验

- ➢ 目标域训练数据数量的影响
  - ○ 针对跨数据域迁移学习任务

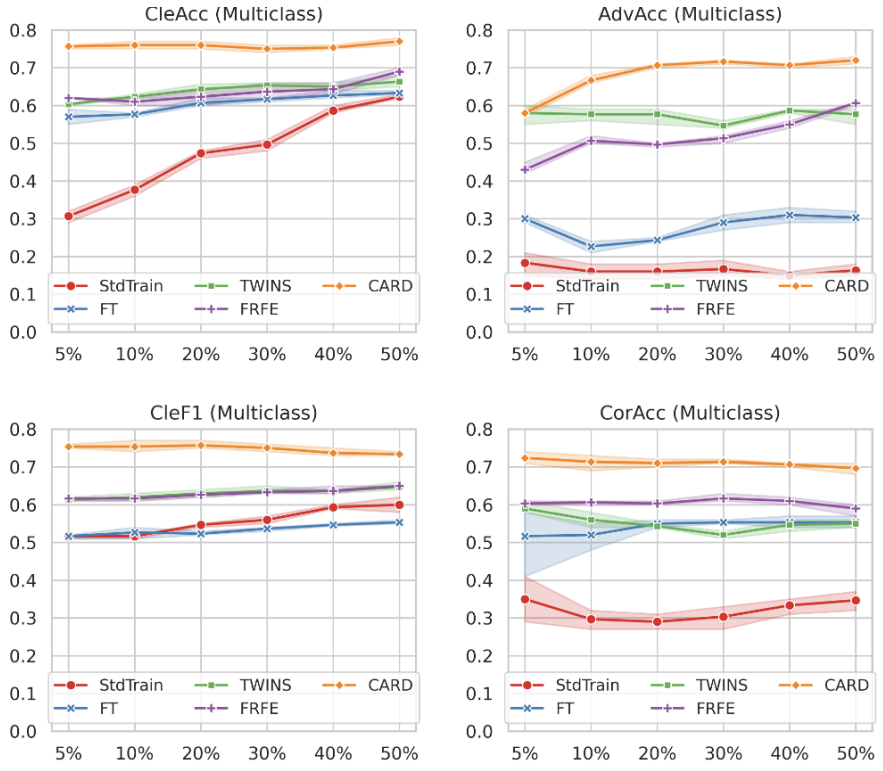


Figure 5.10 Impact of target-domain training data amount on cross-domain TL.

- ➢ 目标域训练数据数量的影响
  - ○ 针对跨模型结构迁移学习任务

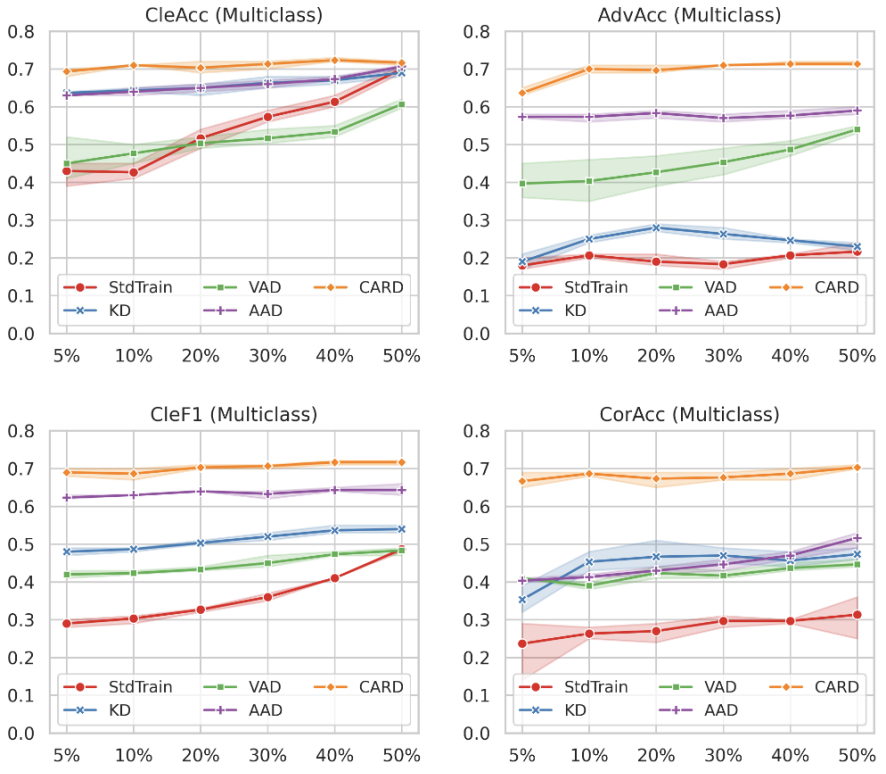


Figure 5.11 Impact of target-domain training data amount on cross-model TL.

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%~50%

□ 实验评估 – 纵向对比实验

- ➢ 目标域训练数据数量的影响
  - ○ 针对跨数据域和模型结构的迁移学习任务
  - ○ 从基于WideResNet-34-10的鲁棒5分类NSL-KDD(All)探测器生成基于MobileNet的10分类 UNSW-NB15(All)检测器

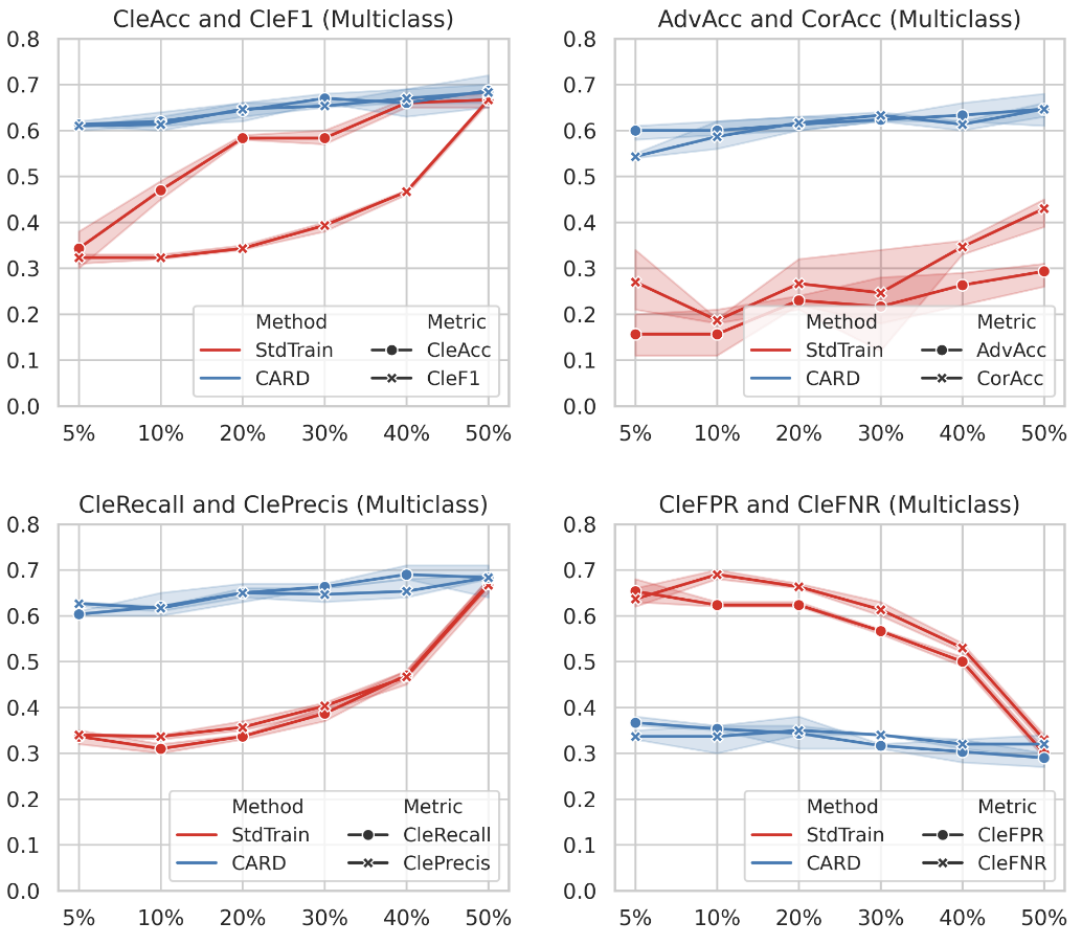


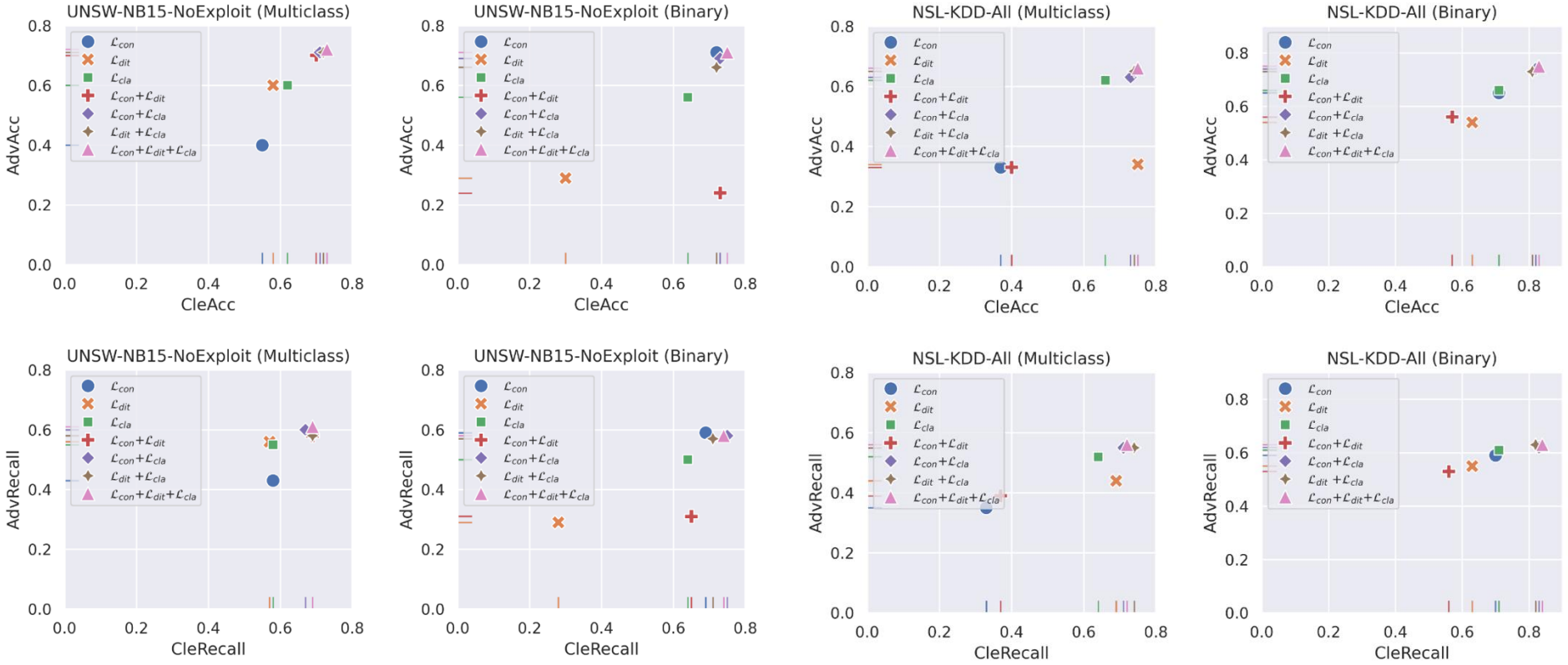
Figure 5.12 Impact of target-domain training data amount on cross-domain-and-model TL.

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

5%

口 实验评估 – 纵向对比实验

➢ 跨模型结构迁移学习任务中损失函数的消融研究

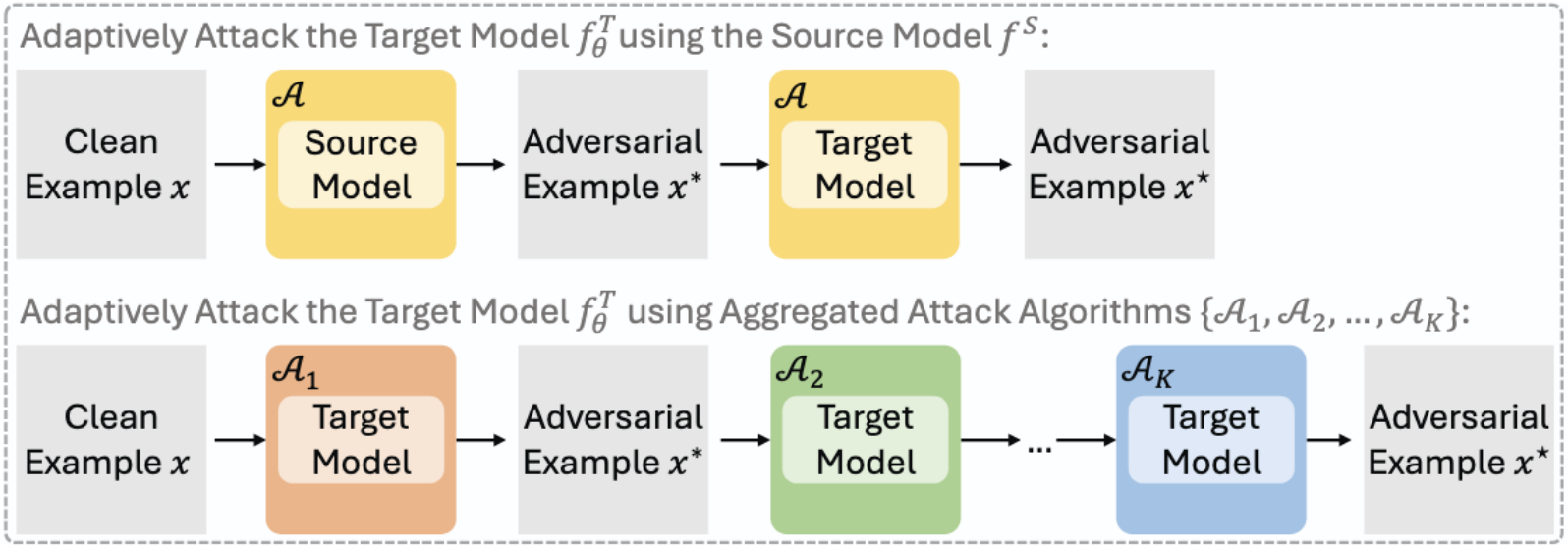


(a) Cross-model TL on UNSW-NB15 (NoExploit)　　(b) Cross-model TL on NSL-KDD (All)

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

实验评估 – 纵向对比实验

- 自适应攻击
  - 策略1：基于源模型的自适应攻击（SM-Adapt）
  - 策略2：基于扰动聚合的自适应攻击（AA-Adapt）



基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 实验评估－纵向对比实验

- ➢ 自适应攻击
  - ○ 策略1：基于源模型的自适应攻击（SM-Adapt）
  - ○ 策略2：基于扰动聚合的自适应攻击（AA-Adapt）

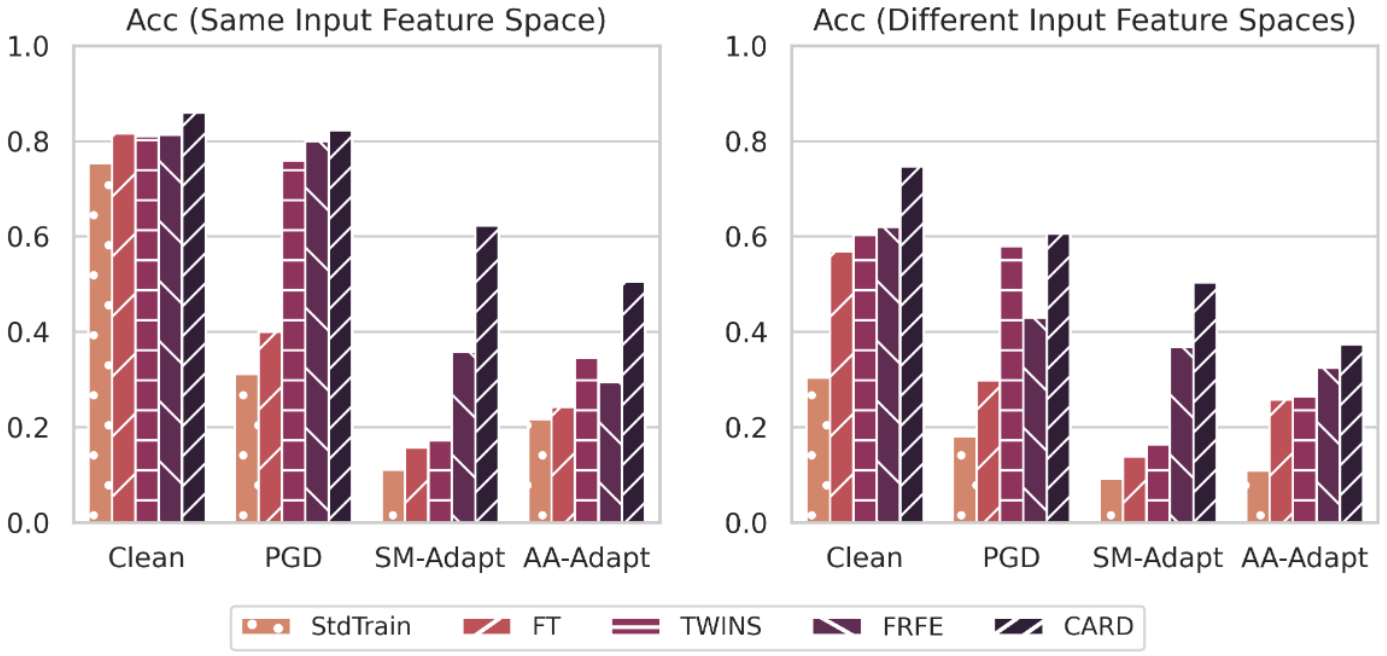


Figure 5.16 Robustness against Adaptive Attacks in Cross-Domain TL

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

□ 实验评估－纵向对比实验

➢ 自适应攻击

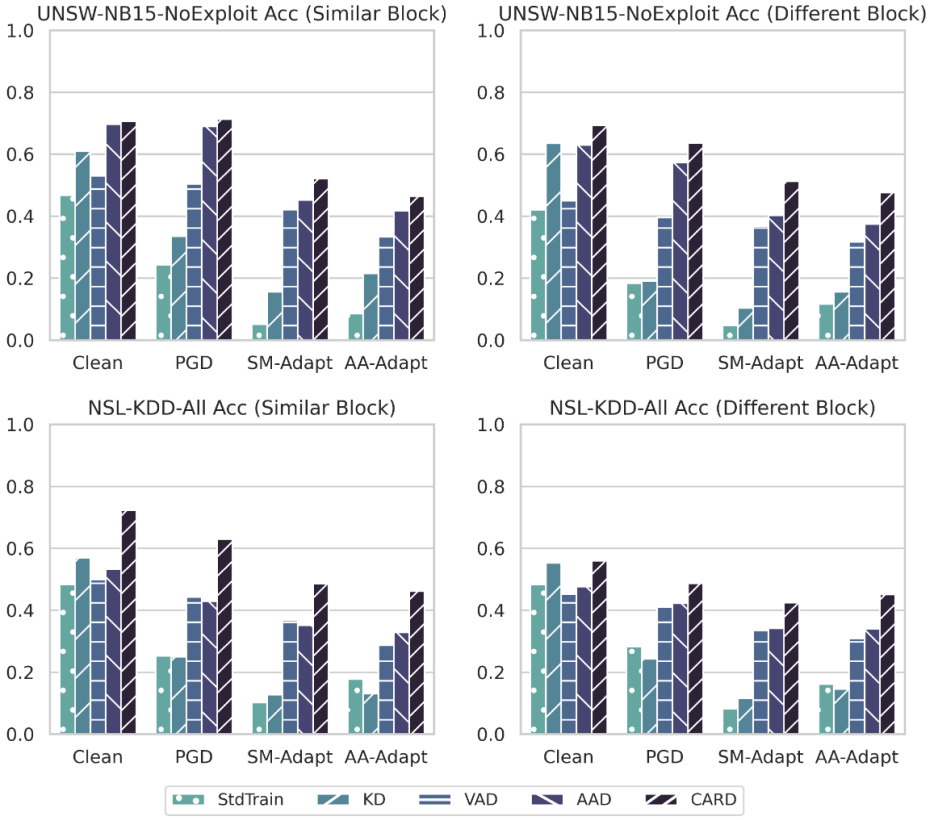


Figure 5.17 Robustness against Adaptive Attacks in Cross-Model TL

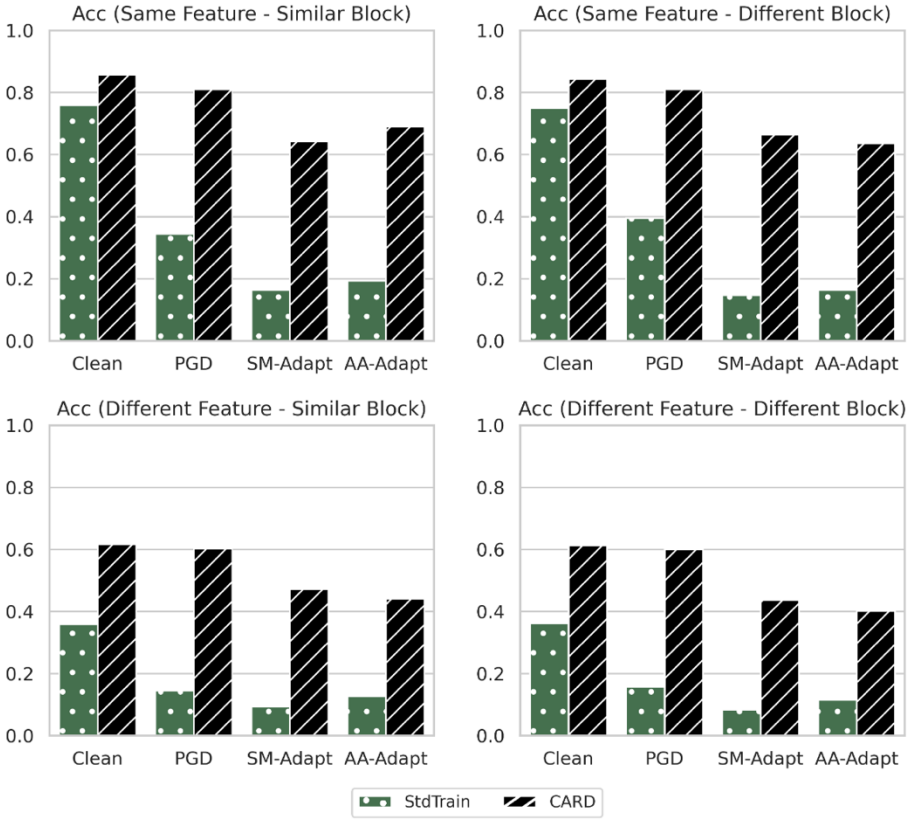


Figure 5.18 Robustness against Adaptive Attacks in Cross-Domain&Model TL

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

实验评估 – 纵向对比实验

- 时间评估

Table 1: Comparison of Time Cost on Cross-Domain TL Tasks

Source Dataset	Target Dataset	TL Method	CleAcc	AdvAcc	CorAcc	Time (sec/epoch)
9-class UNSW-NB15 (NoExploit)	2-class UNSW-NB15 (WithExploit)	StdTrain	0.75	0.31	0.48	0.95
		FT	0.82	0.40	0.79	0.89
		TWINS	0.81	0.76	0.80	32.44
		FRFE	0.81	0.80	0.80	**0.93**
		CARD	**0.86**	**0.82**	**0.84**	16.99
5-class NSL-KDD (All)	10-class UNSW-NB15 (All)	StdTrain	0.30	0.18	0.35	**1.30**
		FT	0.57	0.30	0.51	1.33
		TWINS	0.60	0.58	0.59	47.59
		FRFE	0.62	0.43	0.60	1.36
		CARD	**0.75**	**0.61**	**0.72**	24.17

Table 3: Comparison of Time Cost on Cross-Domain&Model TL Tasks.

Source Dataset	Target Dataset	Target Model	TL Method	CleAcc	AdvAcc	CorAcc	Time (sec/epoch)
UNSW-NB15 (NoExploit)	UNSW-NB15 (WithExploit)	ResNet-18	StdTrain	0.76	0.34	0.41	**0.30**
			CARD	**0.86**	**0.81**	**0.84**	3.61
		MobileNet	StdTrain	0.75	0.40	0.45	**0.29**
			CARD	**0.84**	**0.81**	**0.82**	2.86
NSL-KDD (All)	UNSW-NB15 (All)	ResNet-18	StdTrain	0.36	0.15	0.32	**0.38**
			CARD	**0.62**	**0.60**	**0.61**	5.15
		MobileNet	StdTrain	0.36	0.16	0.27	**0.36**
			CARD	**0.61**	**0.60**	**0.54**	4.12

基于对比对抗表征蒸馏的深度神经网络鲁棒性迁移方案

实验评估 – 纵向对比实验

- 时间评估

Table 2: Comparison of Time Cost on Cross-Model TL Tasks

Dataset	Target Model	TL Method	CleAcc	AdvAcc	CorAcc	Time (sec/epoch)
9-class UNSW-NB15 (NoExploit)	ResNet-18	StdTrain	0.47	0.24	0.47	**0.34**
		KD	0.61	0.34	0.52	0.72
		VAD	0.53	0.50	0.53	2.82
		AAD	0.70	0.69	0.68	14.10
		CARD	**0.71**	**0.71**	**0.70**	4.39
	MobileNet	StdTrain	0.42	0.18	0.23	**0.33**
		KD	0.64	0.19	0.35	0.71
		VAD	0.45	0.40	0.41	2.74
		AAD	0.63	0.57	0.40	14.27
		CARD	**0.69**	**0.64**	**0.67**	3.53
5-class NSL-KDD (All)	ResNet-18	StdTrain	0.48	0.25	0.31	**0.37**
		KD	0.57	0.25	0.33	0.81
		VAD	0.50	0.44	0.48	3.47
		AAD	0.53	0.43	0.45	16.98
		CARD	**0.72**	**0.63**	**0.64**	5.22
	MobileNet	StdTrain	0.48	0.28	0.31	**0.38**
		KD	0.55	0.24	0.23	0.81
		VAD	0.45	0.41	0.40	3.53
		AAD	0.48	0.42	**0.42**	17.37
		CARD	**0.56**	**0.49**	0.40	4.47

结论

- 设计了一种基于自适应维度对齐的蒸馏策略，引入嵌入网络对齐目标模型与源模型的输入维度以及隐层表征维度，以支持数据域和模型变化时的知识迁移。
- 提出了一种基于双重鲁棒感知的对比迁移学习算法，利用输入样本的对抗操纵和自然损坏视图，捕获和学习源表征空间中的领域不变鲁棒信息，以实现通用对抗鲁棒性的迁移。
- 在多种基于深度神经网络的网络入侵检测模型和数据集上对所提的鲁棒性迁移方案进行了实验评估。结果表明，所提方法增强了对抗鲁棒性在跨数据域和跨模型任务间的迁移效果，在数据有限的轻量级模型中实现了领先的对抗鲁棒性。

Contrastive Adversarial Representation Distillation

方案三

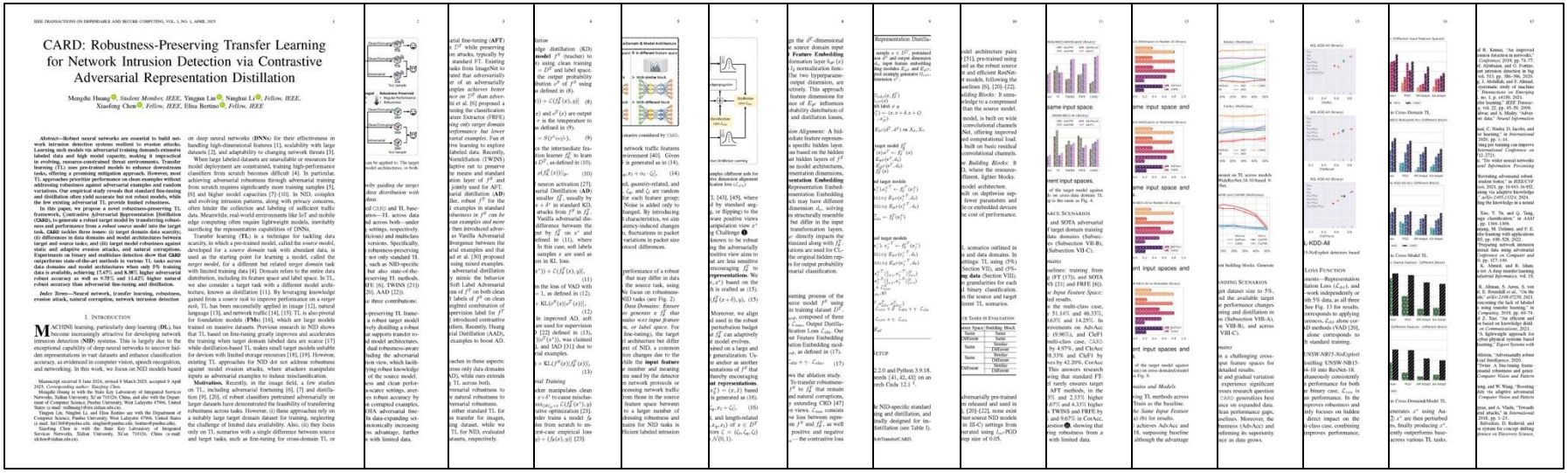
IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTIN

□ 主要成果

主要成果已发表在CCF推荐网络与信息安全 A 类、中科院JCR 二区 期刊

IEEE Transactions on Dependable and Secure Computing (TDSC)

➢ **Mengdie Huang**, Yingjun Lin, Ninghui Li, **Xiaofeng Chen**, Elisa Bertino. CARD: Robustness-Preserving Transfer Learning for Network Intrusion Detection via Contrastive Adversarial Representation Distillation [J]. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2025, 1-18.（1类贡献度）


CARD: Robustness-Preserving Transfer Learning for Network Intrusion Detection via Contrastive Adversarial Representation Distillation

目 录

结论与展望

结论

- 提出基于潜在空间表征混合的深度神经网络对抗鲁棒性泛化方案 Latent Representation Mixup (LarepMixup)。
- 提出基于多阶自适应随机平滑的深度神经网络对抗鲁棒性验证方案 Multi-Order Adaptive Randomized Smoothing (MARS)。
- 提出基于对比对抗表征蒸馏的深度神经网络对抗鲁棒性迁移方案 Contrastive Adversarial Representation Distillation (CARD)。

展望

- 研究非图像数据域潜在空间混合训练技术。
- 研究多模态基础模型的对抗鲁棒性可验证防御方法。
- 研究基础模型到下游模型的鲁棒性迁移技术。

结论与展望

□ 主要成果

- **Mengdie Huang**, Yingjun Lin, Ninghui Li, **Xiaofeng Chen**, Elisa Bertino. CARD: Robustness-Preserving Transfer Learning for Network Intrusion Detection via Contrastive Adversarial Representation Distillation [J]. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2025, 1-18. （CCF-A, 1类贡献度）

- **Mengdie Huang**, Yingjun Lin, **Xiaofeng Chen**, Elisa Bertino. Dimensional Robustness Certification for Deep Neural Networks in Network Intrusion Detection Systems [J]. *ACM Transactions on Privacy and Security (TOPS)*, 2025, 1-33. （CCF-B, 1类贡献度）

- **Mengdie Huang**, Yi Xie, **Xiaofeng Chen**, Jin Li, Changyu Dong, Zheli Liu, Willy Susilo. Boost Off/On-Manifold Adversarial Robustness for Deep Learning with Latent Representation Mixup [C]. *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*, 2023, 1(1):716-730. （CACR-B, 2类贡献度）

- **Mengdie Huang**, Yingjun Lin, Xiaofeng Chen*, Elisa Bertino. MARS: Robustness Certification for Deep Network Intrusion Detectors via Multi-Order Adaptive Randomized Smoothing [C]. *IEEE International Conference on Trust, Security, and Privacy in Computing and Communications (TrustCom)*. 2024. 767-774. (CCF-C, 3类贡献度)

目 录

专家质询问题

盲审专家一

- 对抗攻击的样本形式是否会对未见的对抗攻击产生影响？

回答要点：

- 对于常规对抗训练而言，不同攻击方法生成的对抗样本具有不同的扰动分布特性，因此对抗训练过程中采用的对抗样本生成形式可能会对目标模型面对未见对抗攻击时的鲁棒性表现产生影响。
- 对于我们提出的对抗鲁棒性泛化和验证方案而言，由于防御过程不涉及任何指定类型的对抗样本，因此目标模型在面对未见的对抗攻击时仍然具有良好的鲁棒性。
- 对于我们提出的对抗鲁棒性迁移方案，下游目标模型的鲁棒性来自于对源模型的继承和适应，并且强化了对通用鲁棒表征的学习，因此目标模型在面对未见的两类自适应对抗攻击时也表现出了良好的鲁棒性。

专家质询问题

盲审专家一

- 样本攻击的数据形式是否会对对抗鲁棒性产生影响？

回答要点：

- 样本的数据形式包含样本的模态结构和分布特征等。
- 对于样本的模态结构，通过我们对对抗鲁棒性验证方案的实验评估可知，针对具有同构特征的图像数据提出的鲁棒性验证基准方案在具有异构特征的网络流量数据上有效性会下降。因此可以得到初步结论，面对同一防御算法，样本攻击的模态结构会对模型的对抗鲁棒性产生一定影响，因此需要特定于样本模态的设计。
- 对于样本的分布特征，通过我们对对抗鲁棒性迁移方案的实验评估可知，模型的鲁棒性常在标准的跨域迁移学习后下降。但是通过特定于目标域的鲁棒性适应和强化算法，可以进一步缓解其对模型的对抗鲁棒性的影响。

专家质询问题

盲审专家二

- 估算这三项工作的计算开销?

回答要点:

- 工作一：对抗鲁棒性泛化。以CIFAR-10数据集为例，在CIFAR-10上训练StyleGAN模型280个epoch，每个epoch耗时约218秒；训练WideResNet28-10模型40个epoch，每个epoch耗时约700秒。
- 工作二：对抗鲁棒性验证。MARS防御的CADE对每个样本的平均认证时间为8.0～14.6毫秒，ACID为24.4～36.1毫秒。MARS防御的CADE认证2K个渗透（Infiltration）或暴力破解（SSH-Bruteforce）恶意流量样本大约需要28秒，认证13K个良性流量样本或10K个DoS-Hulk恶意流量样本需要大约125秒。
- 工作三：对抗鲁棒性迁移。跨数据域迁移平均每个epoch的时间为20.58秒；跨模型迁移为4.40秒，跨数据域和模型迁移为3.15秒。

专家质询问题

评审专家一 禹勇教授

- 论文的三四五章所介绍的三个方案之间有什么关联性?

回答要点:

- 从对抗鲁棒性泛化、验证、迁移三个关键角度，系统地构建了一个相对完整的对抗鲁棒性研究框架，三者相互独立又互为补充。
- 目标一致性：三章均围绕提升深度神经网络在面对对抗攻击时的可靠性展开，尽管侧重点不同，但目标都是提高模型在真实复杂环境中的安全性。
- 层层递进的技术路径：训练阶段增强模型对未知对抗样本的鲁棒性；从理论角度评估模型鲁棒性，提供了安全性下界；在模型轻量化或数据有限的情况下，将已有鲁棒性迁移至新任务。
- 方法协同：三者不仅可以单独使用，也可以组合形成“训练→验证→迁移”的闭环机制。LarepMixup用于生成鲁棒模型，MARS为其鲁棒性提供理论保障，CARD则将该鲁棒能力迁移至其他系统中。

专家质询问题

评审专家一 禹勇教授

- 论文所介绍的三个方法是否适用于现实环境中的部署?

回答要点:

- 算法设计兼顾性能与开销：LarepMixup使用混合潜在表征的数据增强策略，在训练过程中引入的计算开销较低；MARS通过高效的随机平滑技术进行鲁棒性验证，具有较强的适配性和可扩展性；CARD采用蒸馏和对比学习策略，在数据稀缺和计算资源有限的现实条件下依然能迁移鲁棒性，适合轻量化部署。
- 系统性与模块化强：三个方法均可作为模块集成至已有AI系统中：LarepMixup用于模型训练阶段，MARS用于上线前验证，CARD用于模型升级与迁移部署，具有良好的工程集成性。
- 综上，论文中提出的三个方法设计上充分考虑了现实部署的可行性与落地需求，具备在工业系统中实际应用的潜力。

专家质询问题

评审专家二 陈晓江教授

- 虽然集中于神经网络的对抗鲁棒性相关方法研究。但研究内容一是针对图像分类场景，研究内容二和三是针对网络入侵场景。研究内容一的对抗性泛化方法是否可以应用到网络入侵场景中？
  - 不同任务提供不同视角。图像任务中对抗样本主要依赖各维同构的视觉扰动，网络入侵检测中则是各维异构的结构化数据，多场景研究有助于构建对抗鲁棒性的更全面理论基础与实践体系，增强研究成果的广度与实用性。
  - 可以，但需要适当调整。LarepMixup依赖的潜在表征混合和多模态插值策略本质上并不依赖于图像数据的结构，而是依赖于DNN中间层的语义表征空间。因此，其核心思想在结构化的网络流量数据场景中也具备应用潜力。

专家质询问题

评审专家二 陈晓江教授

- 在研究内容一中，所提出的方法能够泛化到未知类型的对抗攻击。何谓未知类型的对抗攻击？是否有对应的验证性实验？

回答要点：

- 未知类型（unseen/unknown）的对抗攻击指的是模型在训练阶段未曾见过的攻击方式或攻击算法生成的对抗样本。
- 我们设计了专门的实验来验证这一点。

- 训练：未使用任何对抗攻击算法生成对抗样本以供训练。
- 测试：引入了多种攻击方式。

Perturbation Space	Name	Norm
Off-Manifold	FGSM	l_∞
	PGD	l_∞
	AutoAttack	l_∞
	DeepFool	l_2
	CW	l_2
On-Manifold	OM-FGSM	l_∞
	OM-PGD	l_∞

Perturbation Space	Name	Norm
Off-Manifold	Fog	l_∞
	Snow	
	Elastic	
	JPEG	
Off-Manifold	Fog	l_∞
	Snow	
	Elastic	
	JPEG	

专家质询问题

评审专家二 陈晓江教授

- 解释选择WideResNet-34-10作为(鲁棒的)源模型，选择ResNet-18和MobileNet作为目标模型的理由。

回答要点：

- 旨在评估对抗鲁棒性迁移中不同模型结构组合对鲁棒性迁移效果的影响，因此选择了两对具有代表性的源-目标模型架构。
- WideResNet-34-10→ResNet-18：此设置模拟迁移学习任务中，将鲁棒知识从尺寸较大的模型迁移到基础块结构相似尺寸更小的模型。能够评估在保持相似结构的情况下，模型压缩或轻量化是否保留鲁棒性迁移性能。
- WideResNet-34-10→MobileNet：此设置模拟更具挑战性的应用场景，例如在资源受限环境中部署鲁棒模型。MobileNet代表了在移动端或嵌入式设备中部署模型的典型选择。用于评估在构建块差异较大的情形下，鲁棒特征能否迁移并保留有效性。

专家质询问题

评审专家二 陈晓江教授

- 深度神经网络包含了多种架构的网络模型，如CNN、RNN、Seq2seq、Transformer、GNN等，解释说明所提方法是否可以应用于所有的网络模型。

回答要点：

- 所提方法在设计上具有一定的通用性，核心思想具有模型无关性（architecture-agnostic），因此原则上可推广到多种深度神经网络架构。
- 但由于不同模型架构间在特征表示、训练机制和输入形式上的差异，实际应用中仍需针对目标模型进行机制适配与验证实验，以确保方法的有效性。

- CNN 的中间层特征是局部空间感知的图像特征。
- RNN / Seq2Seq 注重时间序列或语义依赖的隐藏状态。
- Transformer 使用自注意力机制，特征分布形式不同。
- GNN 使用图结构中的邻居聚合机制，特征嵌入方式独特。

专家质询问题

评审专家三 王子龙教授

- 在对抗鲁棒性泛化方案中，提到了多目标训练算法，该算法相较于现有其他算法有什么独特之处和优势？

回答要点：

- LarepMixup 提出了一个基于语义混合样本的多目标训练算法， 鼓励模型在预测混合样本的类别时按照合成信息比例来学习不止一个目标标签，从而平滑深度神经网络的决策边界。这一方法克服了传统的基于硬标签的对抗训练方法容易产生过拟合的问题，提高了模型在未知样本上的泛化能力。
- 此外，相较已有混合训练算法，通过潜在表征空间混合生产的混合样本更具有真实的语义，满足给定数据集的底层特征结构，因此更有利于模型学习有意义的通用鲁棒表征。

专家质询问题

评审专家三 王子龙教授

- 在对抗鲁棒性验证方案中，随机平滑是常见验证防御手段，多阶自适应随机平滑算法（MARS）是否对比了最新的现有方法，理论上如何保证优势？

回答要点：

- 在基于同构输入样本的图像分类任务中，基础的随机平滑及其扩展方法被广泛研究。但对于具有异构输入的网络流量分类模型而言，可验证防御技术的发展十分有限。
- 所提方法已与最新且唯一的适用于网络流量任务的前沿方法BARS进行了全方位对比，并展现出稳定的鲁棒性验证半径及准确率优势。
- 理论上，通过结合模型在预测结果上的一阶梯度信息和二分查找，该算法能够在从输入样本出发的更大的范围内，搜索有可能突破当前预测结果的扰动样本，进而提供更紧的模型鲁棒性下届。

专家质询问题

评审专家三 王子龙教授

- 在对抗鲁棒性迁移方案中，文中的自适应维度对齐的蒸馏策略在面对特征尺度低的问题中是否会导致模型参数增加？

回答要点：

- 所引入的自适应维度对齐层仅为迁移学习的辅助组件，不影响目标模型的推理路径和部署体积。执行自适应维度对齐的模块是基于单层线性变换层，参数开销相对极小。在特征尺度较低的情况下，线性变换的输入维度本身就较小，对应的参数量和计算量进一步降低，不会显著增加目标模型的复杂度或存储开销。

- 在对比迁移学习中不同的迁移时机与迁移手段对鲁棒性的影响如何？

回答要点：

- 针对一个无法获得源域数据集的鲁棒源模型，所提方法相较以微调、蒸馏为基础的不同迁移学习手段在多种跨数据域、跨模型的任务中都具有优势。

专家质询问题

评审专家四 吕锡香教授

- 第四章研究了模型自然随机噪声下的鲁棒性验证，一般来说，训练数据中自然就有噪声，模型在训练中应该较好地学习到了自然噪声知识，好的模型对自然噪声是具有较高的容忍性的，那么在模型鲁棒性验证中考虑自然噪声所增加的代价多大？是不是值得？

回答要点：

- 鲁棒性验证的核心目标是计算模型有关任意输入样本上的预测结果的鲁棒半径，以评估模型对任何可能存在的对抗攻击或自然扰动的鲁棒性下届。
- 在模型预测时，通过引入自然噪声以实现可靠的预测和鲁棒性验证确实增加了一定的时间成本，但代价相对较小，仅为毫秒级，且存储开销无增加。
- 而其带来的验证价值和可靠性意义显著高于代价本身。因此，在鲁棒性验证中通过随机平滑引入自然噪声是相对值得的，有利于构建可信的AI系统。

恳请各位专家老师批评指正

谢谢！