计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(6): 1506−1521, 2026

提高轻量级模型鲁棒性的对抗排序蒸馏方法

老明瑞[1] 李 正[1] 丁兆云[1] 张雪毅[1] 岳祥虎[2] 郭延明[1]
[1](国防科技大学系统工程学院 长沙 410073)
[2](天津大学智能与计算学部 天津 300072)
(laomingrui17@nudt.edu.cn)

Improving Lightweight Model Robustness via Adversarial Ranking Distillation Method

Lao Mingrui[1], Li Zheng[1], Ding Zhaoyun[1], Zhang Xueyi[1], Yue Xianghu[2], and Guo Yanming[1]
[1] (*College of Systems Engineering*, *National University of Defense Technology*, *Changsha* 410073)
[2] (*College of Intelligence and Computing*, *Tianjing University*, *Tianjin* 300072)

Abstract With the widespread application of lightweight models in resource-constrained scenarios such as edge computing, their insufficient adversarial robustness has become increasingly prominent. While adversarial distillation serves as a primary approach to enhance lightweight models' robustness, existing methods predominantly employ KL (Kullback-Leibler) divergence-based rigid constraints for robust knowledge transfer, which still suffer from flawed knowledge modeling and limited transfer efficiency. To address these issues, we propose an adversarial ranking distillation (ARD) method that significantly improves lightweight models' robustness through a priority ranking constraint mechanism and a multi-level consistency framework. Specifically, the priority ranking constraint mechanism organizes the output elements of vision models on adversarial samples by importance ranking, enforces teacher-student ranking consistency, and implements differentiable loose coupling constraints by approximating discontinuous ranking processes through hyperbolic tangent functions. Building upon this, the multi-level ranking consistency distillation framework models transfer robustness knowledge from three perspectives: categorical semantic correlation, sample semantic correlation, and adversarial discrepancy correlation, enabling multi-perspective transmission of teachers' adversarial defense capabilities while synergistically improving lightweight models' clean sample accuracy and adversarial robustness. Extensive experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets validate the effectiveness of our method, demonstrating substantial performance advantages over state-of-the-art adversarial distillation approaches. Furthermore, the proposed ARD exhibits enhanced adaptability and generalization capabilities under scenarios of limited training data and black-box attacks, providing an efficient security solution for lightweight model deployment in edge computing environments.

Key words adversarial robustness; lightweight model; knowledge distillation; adversarial defense; deep learning

摘 要 随着轻量级模型在边缘计算等资源受限场景中的广泛应用,其对抗鲁棒性不足的问题日益凸显。对抗蒸馏作为提升轻量级模型鲁棒性的重要手段,虽然现有方法普遍采用基于KL(Kullback-Leibler)散度的硬性约束实现鲁棒知识传递,但仍面临知识建模不完善和传递效率受限等问题。为此,该研究提出

收稿日期: 2025-05-23; **修回日期**: 2026-03-04
基金项目: 湖南省重点研发计划基金项目(2025QK3004); 国家自然科学基金青年科学基金项目(62502530); 国防科技大学科研基金项目(ZZCX-JDZ-39, ZK24-27)
This work was supported by the Key Research and Development Program of Hunan Province (2025QK3004), the National Natural Science Foundation of China for Young Scientists (62502530), and the Foundation of NUDT (ZZCX-JDZ-39, ZK24-27).

一种对抗排序蒸馏方法，通过优先级排序约束机制与多层次一致性框架，显著提升轻量级模型面向对抗攻击的抵御能力。具体地，前者旨在将视觉模型基于对抗样本的输出元素的重要性排序，进而约束教师-学生模型间的元素优先级保持一致，并通过函数近似非连续排序过程实现可导的松耦合约束。在此基础上，多层次排序一致性蒸馏框架从类别语意关联、样本语意关联以及对抗差异关联3个方面对鲁棒性进行建模与知识迁移，实现教师对抗防御能力的多视角传递，协同提升轻量级模型的干净样本准确率与对抗鲁棒性。在CIFAR-10, CIFAR-100, Tiny ImageNet数据集上实验验证了该方法的有效性，且相比于现有的先进对抗蒸馏方法有显著性能优势。同时，对抗排序蒸馏方法在训练数据受限和黑盒攻击场景下展现出更强的适应性与泛化能力，为轻量级模型在边缘部署中的安全防御提供了高效解决方案。

关键词 对抗鲁棒性；模型轻量化；知识蒸馏；对抗防御；深度学习

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550339 **CSTR:** 32373.14.issn1000-1239.202550339

随着神经网络[1]和深度学习[2]技术的快速发展，以图像分类[3]和目标识别[4]为代表的计算机视觉研究在多个应用领域取得了显著进展。受益于大模型技术[5]及扩展法则(scaling law)[6]的影响，视觉模型逐步朝着大规模网络框架与结构演进。然而，在实际应用场景中，视觉识别系统通常部署在资源受限的用户端或边缘设备上。因此，开发参数量级较小且性能优越的轻量级模型显得尤为必要。此外，由于轻量级模型在应用中用户交互和边缘设备部署的特点，相较于大规模视觉模型，它们更易受到攻击者的恶意攻击[7]，进而影响模型的鲁棒性和可靠性[8]。因此，提升轻量级模型的对抗鲁棒性具有重要的研究价值。

在众多对抗防御的方法中，对抗训练[9]是运用最广泛且高效的策略，其原理是通过在训练过程中引入对抗样本，使模型在面对潜在攻击时能够提高其对抗鲁棒性。从数学形式化角度而言，对抗训练过程可建模为一个双层优化问题，其本质表现为最小-最大优化框架。在该框架下，内层优化通过最大化对抗扰动以生成最具攻击性的对抗样本，而外层优化则通过最小化模型在这些对抗样本上的经验风险来实现鲁棒性提升。然而，对抗训练在小模型的对抗防御场景中仍存在一定的局限性[10-12]：一方面，轻量级模型由于参数量受限，其对抗训练的表达能力和样本学习能力较弱，难以捕捉对抗样本的复杂特征，训练过程也相对不稳定；另一方面，轻量级模型在同时处理噪声扰动和干净样本时缺乏足够的灵活性与适应性，难以在对抗训练中生成高效的对抗样本。近年来，对抗鲁棒蒸馏[13]逐渐成为提升小模型对抗训练能力的关键研究技术。其旨在基于教师-学生网络的知识迁移框架[14-15]，充分利用经过对抗训练的大模型的知识，以引导轻量级模型进行有效的对抗训练。通过这种方式，轻量级模型能够借鉴与模仿大模型面向对抗样本的抵御和判别能力，提升其在实际应用中应对恶意攻击的安全性与可靠性。

自面向轻量级模型的对抗鲁棒性蒸馏范式[13]被提出以来，大量工作对其进行了深入探索和优化，其核心研究点主要聚焦在3个方面：1)标签改进技术。通过如鲁棒软标签[16]、标签混合[17]、标签公平性[18]等手段增强轻量级模型对大模型的对抗鲁棒性的模仿能力。2)教师选择方法。通过结合干净样本训练教师[19]、学生模型自蒸馏[20]等手段多教师联合引导轻量级模型的鲁棒性学习。3)训练改进策略。通过对内层/外层优化[21-22]的持续改进提升轻量级模型的对抗样本生成质量与防御能力。尽管这些方法一定程度提升了对抗鲁棒蒸馏的性能，但目前几乎所有先进方法在蒸馏模式上都贯彻了一致的思路：基于给定教师模型的响应(如优化后的软标签)，利用KL(Kullback-Leibler)散度[23]实现大模型向轻量级模型的对抗防御知识传递，其本质是约束教师模型与学生模型间的响应/预测值保持一致。

然而，本文认为，基于KL散度对教师-学生模型预测值施加的硬性约束在对抗鲁棒蒸馏中仍存在一定局限性：首先，与干净样本不同，对抗样本为了实现强欺骗性，其攻击方向通常指向与正确答案高度相关的类别[24]，而KL散度下的类别预测值难以准确捕捉这种攻击方向上的类别语义关系，从而限制了对抗鲁棒性知识的传递；其次，由于教师模型与学生模型的参数量存在显著差异，两者在对抗样本的特征学习与表征能力上差距明显，对预测值进行硬约束可能进一步增加轻量级模型提升对抗鲁棒性的难度。

针对这一问题，本文提出了一种新颖且高效的

轻量级模型鲁棒性提升方法——对抗排序蒸馏。具体而言,与现有方法通过对教师-学生模型的输出元素(如类别预测值)施加刚性约束的蒸馏模式不同,本研究提出的优先级排序约束机制通过以下方式实现鲁棒知识迁移:首先对视觉模型基于对抗样本的输出元素进行重要性排序,然后强制保持教师模型与学生模型间的排序一致性。受冒泡排序启发,该方法构建优先级对比函数实现两两元素间重要性的比较(重要则输出0,反之输出1),并整合所有元素对的函数输出得到每个元素的优先级排序。然而,这种理想化的优先级对比函数是非连续的,其无法直接通过梯度反向传播方法来优化学生模型进行训练。针对这一问题,我们提出用双曲正切函数近似排序过程,使整个约束过程具备可导性。我们所提出的优先级排序约束机制在对抗鲁棒性蒸馏范式中有2个方面优势:1)面向元素(如类别预测)的排序相比于其输出数值能够更有效地建模与表达对抗鲁棒性知识,如基于特定对抗样本2种不同预测类别的重要程度及对抗样本在类别攻击方向上发生的偏移等;2)基于优先级排序的约束在教师-学生网络间具有松耦合性且易于实现,能够在在学生网络参数量与表征能力受限的情况下,显著降低轻量级网络学习对抗鲁棒性知识的难度。

受益于优先级排序约束机制的对抗鲁棒性建模与高效知识传递能力,对抗排序蒸馏方法构建了一个多层次排序一致性蒸馏框架,旨在从类别语义关联、样本语义关联以及对抗差异关联3个方面对鲁棒性进行建模与知识迁移。1)类感知的排序一致性蒸馏。基于特定对抗样本输入,根据模型对不同类别预测的重要性进行优先级排序,其中排序元素是所有预测类别。2)实例感知的排序一致性蒸馏。基于特定预测类别,根据一个训练批次内样本与该类别的预测相似性进行排序,此时排序元素是训练批次中的全部对抗样本。3)差异感知的排序一致性蒸馏。基于特定训练样本,根据模型对其干净样本与对抗样本的预测差异进行类别排序(如对“飞机”预测的变化幅度大于“汽车”的变化幅度),此排序反映了模型预测面向对抗攻击的鲁棒性偏移。不同于主流方法局限于类别预测的知识迁移,本文提出的对抗排序蒸馏方法能够从多视角传递教师模型应对恶意攻击的防御能力,协同提升小模型的干净样本准确率与对抗鲁棒性。总体而言,本文贡献有3个方面。

1)从对抗蒸馏的鲁棒性知识建模以及学生模型的鲁棒性学习难度出发,首次提出以响应元素排序的视角实现教师-学生模型间对抗防御能力的高效传递。

2)提出了一种用于提升轻量级模型鲁棒性的全新策略——对抗排序蒸馏,具体包括2部分:优先级排序约束机制构建了可导的输出元素排序策略,实现教师-学生网络响应的松耦合约束;多层次排序一致性则基于前者的对抗鲁棒性表达能力,实现从类别语义关联、样本语义关联以及对抗差异关联3个维度的多视角鲁棒知识蒸馏。

3)为了可靠地评估对抗排序蒸馏在轻量级模型对抗防御上的有效性,本文在对抗训练领域广泛使用的CIFAR-10, CIFAR-100, Tiny ImageNet数据集上进行了大量实验分析。结果表明,相比已有方法,对抗排序蒸馏能够在稳定维持轻量级模型干净样本准确率的同时,显著提升对抗鲁棒性能。同时,在训练样本数量受限以及不同外部最小化函数设定的环境中,相比于刚性约束的对抗鲁棒蒸馏策略,本文方法具备更强的适应力与通用性。

1 相关工作

深度神经网络在多个研究应用领域展现出卓越性能,而轻量级模型凭借其高效的计算性能和边缘部署优势获得了广泛关注。然而,研究表明深度神经网络存在显著的对抗脆弱性问题[7]。具体而言,攻击者通过在原始数据中添加人眼难以察觉的微小扰动生成对抗样本,这些样本会导致神经网络产生错误输出,从而引发实际应用中的安全威胁。值得注意的是,由于参数规模有限和表征能力不足,轻量级模型的对抗脆弱性问题更为突出。

为提升深度神经网络的对抗鲁棒性,对抗训练[9]作为一种典型的启发式解决方案,采用攻击与防御协同优化的框架。该策略通过以下2个关键机制实现鲁棒性提升:首先,利用对抗攻击方法生成具有强干扰性的对抗样本;其次,将这些对抗样本纳入训练过程,优化模型使其能够正确识别对抗样本,从而增强模型对对抗攻击的防御能力。这种“以毒攻毒”的训练范式已被证明能有效提升模型的鲁棒性。针对对抗训练的持续改进和优化方法主要聚焦在对抗样本生成以及模型防御训练2个方面:对抗样本生成聚焦于在对抗训练的攻击阶段,生成难以感知攻击隐蔽性强的图像空间攻击样本,主要方法包括FGSM[25], PGD[26], CW[27], AA[28]等方法;模型防御训练则关注对抗训练的防御阶段,其基本问题是利用已生成的对抗样本进行训练,并通过分类或检测损失最小化来

更新模型，其通常结合参数自适应、模型增强、损失正则项以及早期停止等技术进一步提升对抗鲁棒性。然而，由于对抗样本生成质量以及数据拟合能力的差异，研究表明[29]对抗训练方法在容量大、高度参数化网络中的表现显著强于轻量级模型，限制了对抗防御技术在移动端与边缘设备的应用发展。

受知识蒸馏在模型压缩领域成功应用的启发，基于教师-学生架构的对抗鲁棒蒸馏已成为提升轻量级模型对抗防御能力的重要范式。该范式采用经过对抗训练的大型模型作为教师模型，通过知识迁移机制，在保持干净样本预测精度的同时，指导轻量级学生模型学习对抗鲁棒性特征[30]。后续对抗鲁棒蒸馏的相关工作主要从3个方面进行创新：1)标签优化策略，旨在教师-学生知识传递中产生更为高效的训练标签来指导学生进行鲁棒性强化。AKD方法[17]较早探索了标签混合对于鲁棒性转移的影响，而RSLAD[16]提出使用鲁棒软标签作为监督信号，能够更好地捕捉教师模型的鲁棒行为。Fair-ARD方法[18]则将鲁棒公平性引入对抗蒸馏，并提出了基于几何视角的最少对抗步数度量方法，以对标签类别权重进行动态重加权。2)教师选择策略，旨在筛选多样化教师进行学生网络鲁棒蒸馏。由于教师模型在学生生成对抗样本上表现出的不稳定现象，IAD方法[20]提出教师蒸馏与自蒸馏结合的多教师蒸馏策略。不局限于单个鲁棒教师模型，Zhao等人[19]提出额外引入一个干净样本训练下的教师模型来指导对抗蒸馏，通过结合2个教师模型各自的优势，来综合改善学生模型的性能。B-MTARD[31]则在此基础上通过增加一个额外的熵平衡模块，来平衡不同教师模型之间的知识规模。3)训练改进方法，旨在从对抗蒸馏内层/外层优化的训练模式上进行持续提升。Huang等人[21]将教师模型引入内层优化过程，自适应地搜索学生模型和教师模型在扰动邻域中的最大预测差异点以改进蒸馏效果。DGAD方法[22]提出了动态指导的对抗蒸馏框架，通过动态调整样本权重优化知识传递，并解决教师模型的误分类问题。SmaraAD[32]则设计了一种归因区域对齐机制，从缓解训练干扰以及注意力行为理解不足的视角持续提高学生模型鲁棒性。Wu等人[33]提出了一种基于混合解耦蒸馏的小容量模型对抗防御方法，通过高效提取预训练教师模型中多层次、细粒度的鲁棒性知识，显著提升小容量模型的对抗鲁棒性。相关工作及方法特点总结如表1所示。

不同于常规基于KL散度的刚性蒸馏方法，本文提出一种松弛性更高的元素排序策略来实现对抗鲁棒蒸馏，得益于在鲁棒性建模与知识传递方面的高效性，本文方法获得使轻量级学生模型能够从3个视角协同继承教师大模型抵御对抗攻击的能力：面向特定样本的类别预测、面向特定类别的样本语义关联以及面向干净样本与对抗样本的类别预测差异。

Table 1 Summary and Comparison of Adversarial Distillation Relative Methods

表1 对抗蒸馏相关方法总结与对比

方法	年份	蒸馏改进策略			蒸馏知识类型			蒸馏方式	
		标签改进	教师选择	训练优化	类别预测	样本关联	对抗差异	KL散度	排序约束
文献[13]	2020							✔	
文献[16]	2021	✔			✔			✔	
文献[17]	2022	✔			✔			✔	
文献[18]	2024	✔			✔			✔	
文献[19]	2022		✔					✔	
文献[20]	2022		✔		✔			✔	
文献[21]	2023			✔	✔			✔	
文献[22]	2025			✔	✔	✔		✔	
文献[31]	2024		✔		✔	✔		✔	
文献[32]	2024			✔	✔	✔		✔	
文献[33]	2024			✔	✔	✔		✔	
本文				✔	✔	✔	✔		✔

2 预备知识

本文方法主要基于对抗训练和对抗蒸馏理论，下面就相关概念和基本知识予以介绍。

2.1 对抗训练

为抵御对抗样本的恶意攻击，对抗训练结合对抗样本生成与防御性训练技术，进而生成鲁棒视觉识别模型。这个过程可以被定义为一个最小最大优化问题，表示如下

$$\min_{\theta} E_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\Omega}\mathcal{L}_{\mathrm{ce}}\left(f\left(x_{\mathrm{cle}}+\delta;\theta\right),y\right)\right], \tag{1}$$

其中$f(\cdot;\theta)$代表一个权重θ的深度神经网络，$\mathcal{D}$是包括了干净训练样本x_{cle}及其对应的类别标签y的数据分布，$\mathcal{L}_{\mathrm{ce}}$为交叉熵损失函数。$\delta$是对抗扰动，而$\Omega$是一个定义为$\Omega=\{\delta:\|\delta\|\leqslant\epsilon\}$的界限，其中最大扰动范围为$\epsilon$。对抗样本则表示为干净样本与对抗扰动的叠加$x_{\mathrm{adv}}=x_{\mathrm{cle}}+\delta$。为了在式(1)中的内部最大化中生成对抗样本，对抗训练普遍采用PGD(projected gra-

dient descent)方法通过多次迭代逐步优化扰动 δ。

$$\delta^{n+1}=\Pi_{\Omega}\left(\delta^{n}+\alpha\cdot\mathrm{sgn}(\nabla_{x}\mathcal{L}(f(x_{\mathrm{cle}}+\delta^{n};\theta),y))\right)，\quad(2)$$

其中δ^{n}是第 n 次优化步骤的对抗扰动，而$\Pi(\cdot)$和α分别代表映射函数和步长。最终的对抗扰动δ^{*}是通过初始化δ^{0}为均匀分布并在 PGD 算法中进行 N 次总步数的迭代获得的。

2.2 对抗蒸馏

标准知识蒸馏范式通过最小化教师模型和学生模型预测之间的分布差异，将知识从预先训练好的教师模型转移到学生模型。形式上，对于学生和教师网络输出逻辑$S(x)\in\mathbb{R}^{N^{\mathrm{b}}\times C}$和$T(x)\in\mathbb{R}^{N^{\mathrm{b}}\times C}$，其中$N^{\mathrm{b}}$和$C$分别表示训练批次大小和预测总类别数，知识蒸馏的损失表示为

$$\mathcal{L}_{\mathrm{kd}}(T(x);S(x;\theta))=\frac{\tau^{2}}{N^{\mathrm{b}}}\sum_{i=1}^{N^{\mathrm{b}}}KL(\sigma(T(x)_{i,:},\tau);\sigma(S(x)_{i,:},\tau))=\sum_{i=1}^{N^{\mathrm{b}}}\sum_{j=1}^{C}\ln\left(\frac{\sigma(T(x)_{i,j},\tau)}{\sigma(S(x)_{i,j},\tau)}\right)，\quad(3)$$

其中$KL(\cdot;\cdot)$指 KL 散度，用于衡量学生模型与教师模型输出分布的相似性。参与蒸馏学生及教师模型的预测分布为

$$\begin{aligned}\sigma(S(x)_{i,:},\tau)&=softmax\left(\frac{S(x)_{i,:}}{\tau}\right)，\\ \sigma(T(x)_{i,:},\tau)&=softmax\left(\frac{T(x)_{i,:}}{\tau}\right)，\end{aligned}\quad(4)$$

其中τ是用于控制预测平滑程度的温度因子。在对抗防御中，学生模型不仅要继承教师模型对自然样本预测的准确性，还要模仿其抵御恶意攻击样本的对抗鲁棒性。因此，作为知识蒸馏在对抗防御领域的扩展，对抗蒸馏旨在结合对抗训练技术，通过教师模型的指导提升学生模型在对抗样本上的鲁棒性表现。其核心思想是引入对抗样本作为蒸馏过程中的重要组成部分，使学生模型能够模仿教师模型在对抗样本上的输出分布，从而学习到教师模型的鲁棒性表现。对抗蒸馏方法的优化目标可以写为

$$\begin{aligned}&\arg\min_{\theta}(1-\beta)\cdot\mathcal{L}_{\mathrm{ce}}(S(x_{\mathrm{cle}}),y)+\beta\cdot\mathcal{L}_{\mathrm{kd}}(S(x_{\mathrm{adv}};\theta),T(x_{\mathrm{cle}}))，\\ &x_{\mathrm{adv}}=\arg\max_{\|\delta\|_{p}<\epsilon}\mathcal{L}_{\mathrm{ce}}(S_{\theta}(x_{\mathrm{cle}}+\delta),y)，\end{aligned}\quad(5)$$

其中x_{adv}，x_{cle}分别指代干净与对抗样本，β为任务和蒸馏损失间的权重调整参数。此外，为了充分增强学生模型生成样本在教师模型中的适应性，自适应对抗蒸馏技术(AdaAD)将教师模型融入对抗样本的生成过程，即在最大化学生模型与教师模型在对抗扰动输入上的预测差异，具体公式可表达为

$$\begin{aligned}&\arg\min_{\theta}(1-\beta)\cdot\mathcal{L}_{\mathrm{kd}}(S_{\theta}(x_{\mathrm{cle}})\|T(x_{\mathrm{cle}}))+\\ &\quad\beta\cdot\mathcal{L}_{\mathrm{kd}}(S_{\theta}(x_{\mathrm{adv}})\|T(x_{\mathrm{adv}}))，\\ &x_{\mathrm{adv}}=\arg\max_{\|\delta\|_{p}<\epsilon}\mathcal{L}_{\mathrm{kd}}(S_{\theta}(x_{\mathrm{cle}}+\delta),T(x_{\mathrm{cle}}+\delta))。\end{aligned}\quad(6)$$

3 对抗排序蒸馏

鉴于当前对抗蒸馏方法普遍忽视了教师-学生网络间刚性约束的局限性，本文提出了一种新颖且有效的对抗排序蒸馏方法，该方法包括 2 个核心部分：优先级排序约束机制以及多层次排序一致性蒸馏框架。前者面向对抗鲁棒性构建了松耦合且高效的知识建模与传递机制，而后者在此基础上分别从类别语意关联、样本语意关联以及对抗差异关联 3 种不同层次挖掘教师模型对抗鲁棒性，并引导其协同提升轻量级模型的干净样本下的准确率以及对抗攻击下的韧性。

3.1 优先级排序约束机制

尽管现有对抗蒸馏研究在监督标签改进、教师模型选择等方面上进行了广泛研究与优化，但几乎所有这些方法在教师-学生网络间蒸馏模式上均采用基于 KL 散度的硬性约束，即监督学生模型的输出元素值(例如响应)与教师模型保持一致。然而，值得注意的是，相较于干净样本，对抗样本的预测数值普遍因对抗攻击的方向而发生偏移，此时类别间的优先级关系往往比预测数值本身更为重要。其次，由于大模型与小模型在参数量及表征能力上的显著差异，学生模型在模拟教师模型对抗样本响应时面临的难度明显高于模拟干净样本，这导致对抗训练的难度显著增加。

针对上述问题，本文提出了一种优先级排序约束机制，如图 1 所示，不同于 KL 散度对教师模型与学生模型输出响应值的分布约束，该机制关注响应元素优先级排序并以排序一致性作为约束。

具体地，假设教师网络与学生网络的输出元素序列分别为$\boldsymbol{Z}^{\mathrm{t}}=(z_{1}^{\mathrm{t}},z_{2}^{\mathrm{t}},\cdots,z_{N}^{\mathrm{t}})\in\mathbb{R}^{d}$以及$\boldsymbol{Z}^{\mathrm{s}}=(z_{1}^{\mathrm{s}},z_{2}^{\mathrm{s}},\cdots,z_{N}^{\mathrm{s}})\in\mathbb{R}^{d}$，其中$d$为元素序列维度(即元素个数)。首先需要对响应值$\boldsymbol{Z}^{\mathrm{t}}$，$\boldsymbol{Z}^{\mathrm{s}}$内元素的重要性排序，受冒泡排序策略启发，构建优先级对比矩阵$\boldsymbol{D}^{\mathrm{t}}$，$\boldsymbol{D}^{\mathrm{s}}\in\mathbb{R}^{N\times N}$可表达为

$$\boldsymbol{D}^{\mathrm{t}}=I\left(\begin{pmatrix}z_{1}^{\mathrm{t}}&\cdots&z_{N}^{\mathrm{t}}\\ \vdots&&\vdots\\ z_{1}^{\mathrm{t}}&\cdots&z_{N}^{\mathrm{t}}\end{pmatrix}-\begin{pmatrix}z_{1}^{\mathrm{t}}&\cdots&z_{1}^{\mathrm{t}}\\ \vdots&&\vdots\\ z_{N}^{\mathrm{t}}&\cdots&z_{N}^{\mathrm{t}}\end{pmatrix}\right)，\quad(7)$$

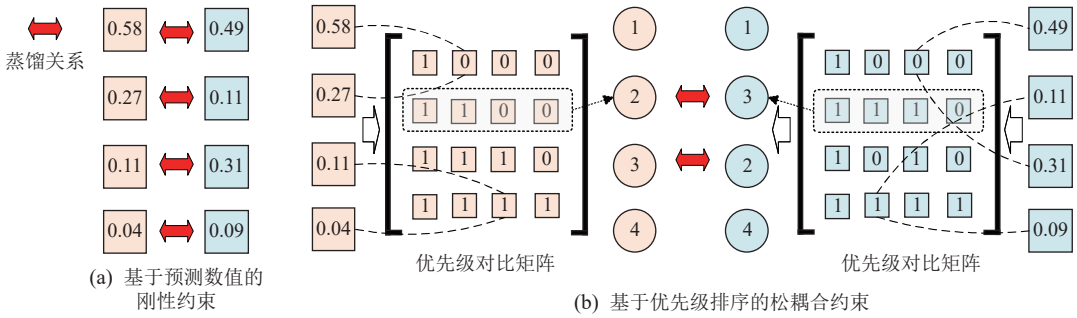


Fig. 1 Illustration of priority ranking constraint mechanism

图 1 优先级排序约束机制示意图

$$I(x)=\begin{cases}1, & \text{if } x<0,\\ 0, & \text{if } x\geqslant 0,\end{cases}$$

其中$I(\cdot)$为二值示性函数，$D_{ij}^{\mathrm{t}}=I(z_i^{\mathrm{t}}-z_j^{\mathrm{t}})$表示在教师模型响应中第$i$个元素与第$j$个元素的优先级比较，根据优先级矩阵$\boldsymbol{D}^{\mathrm{t}}$可定义$z_i^{\mathrm{t}}$在元素队列$(z_1^{\mathrm{t}},z_2^{\mathrm{t}},\cdots,z_N^{\mathrm{t}})\in\mathbb{R}^d$中的优先级排序$L_i^{\mathrm{t}}$为

$$L_i^{\mathrm{t}}=1+\sum_{k=1,k\neq i}^{d} I\left(\boldsymbol{D}_i^{\mathrm{t}}<\boldsymbol{D}_k^{\mathrm{t}}\right), \tag{8}$$

其中k代表元素队列中第k个不同于待排序i的元素。虽然式(8)能够有效对元素排序值进行构建，其所涉及的二值示性函数$I(\cdot)$是非连续不具备可导属性的，并不适用于梯度回传优化模式下的对抗蒸馏任务。针对这一问题，优先级排序约束机制提出使用双曲正切函数变体$\mathbb{A}(\cdot)$来近似示性函数$I(\cdot)$的优先级对比功能

$$I\left(z_i^{\mathrm{t}}-z_j^{\mathrm{t}}\right)\approx\mathbb{A}\left(z_i^{\mathrm{t}}z_j^{\mathrm{t}}\right)=\frac{1}{2}\left(\tanh\left(z_i^{\mathrm{t}}-z_j^{\mathrm{t}}\right)+1\right)。 \tag{9}$$

由此，式(9)可求导为

$$\frac{\partial\mathbb{A}}{\partial z_i^{\mathrm{t}}}=1-\frac{1}{2}\left(\tanh\left(z_i^{\mathrm{t}}-z_j^{\mathrm{t}}\right)\right)^2, \tag{10}$$

则z_i^{t}在元素队列$(z_1^{\mathrm{t}},z_2^{\mathrm{t}},\cdots,z_N^{\mathrm{t}})\in\mathbb{R}^d$中的优先级排序$L_i^{\mathrm{t}}$可改进为

$$L_i^{\mathrm{t}}\approx 1+\sum_{k=1,k\neq i}^{d}\frac{1}{2}\left(\tanh\left(z_i^{\mathrm{t}}-z_k^{\mathrm{t}}\right)+1\right)。 \tag{11}$$

基于式(11)，所提出的优先级排序约束机制可进一步使用绝对值函数$|\cdot|$对教师模型与学生模型间响应排序的一致性进行监督。对于包含N^{b}个样本的训练批次，其训练损失函数$\mathcal{L}_{\mathrm{rp}}$可表达为

$$\mathcal{L}_{\mathrm{rp}}=\frac{1}{N^{\mathrm{b}}}\sum_{i=1}^{N^{\mathrm{b}}}\sum_{j=1}^{d}\left|L_{ij}^{\mathrm{t}}-L_{ij}^{\mathrm{s}}\right|。 \tag{12}$$

与传统的知识蒸馏机制(式(3))进行对比，优先级排序约束机制能够实现教师-学生间对抗鲁棒性的高效传递。其中，优先级对比矩阵实现逐对元素间重要性比较，为对抗鲁棒知识进行高效建模；排序一致性约束则充分利用松耦合监督促进能力受限小模型对大模型鲁棒性行为的模仿。排序约束在训练阶段计算复杂度为$O(N^2)$，高于KL散度的$O(N)$，但在必要情况下可通过Top-K采样策略将开销控制在一定范围内；测试阶段，该方法的推理过程与计算复杂度与常规方法无异。

3.2 多层次排序一致性蒸馏框架

优先级排序约束机制聚焦于教师-学生网络间的对抗样本知识建模与传递的模式，而多层次排序一致性蒸馏框架则注重将不同视角的对抗鲁棒性从大模型充分传递给小模型，具体可分为类感知排序一致性蒸馏、实例感知排序一致性蒸馏，以及差异感知排序一致性蒸馏，方式示意如图 2 所示。

1)类感知排序一致性蒸馏。与主流对抗蒸馏方法视角类似，类感知排序一致性蒸馏$\mathcal{L}_{\mathrm{crd}}$关注给定对抗样本下教师-学生网络预测间的知识蒸馏，此时待排序元素是全部候选预测类别。以学生模型$S(\cdot)$中类感知排序为例，网络对第i个对抗样本的输出逻辑为$S(x_{\mathrm{adv}})_{i:}\in\mathbb{R}^C$，其中$C$为候选预测类别数量，则第$j$个预测类别的优先级排序$L_{ij}^{\mathrm{sc}}$可表示为

$$L_{ij}^{\mathrm{sc}}=1+\sum_{k=1,k\neq j}^{C}\frac{1}{2}\left(\tanh\left(S(x_{\mathrm{adv}})_{ij}-S(x_{\mathrm{adv}})_{ik}\right)+1\right)。 \tag{13}$$

同理可得教师模型$T(\cdot)$下的优先级排序L_{ij}^{tc}。基于3.1节提出排序一致性监督损失(式(12))，类感知排序一致性蒸馏损失$\mathcal{L}_{\mathrm{crd}}$为每个类别加入了额外类感知排序权重$W_{i:}^{\mathrm{c}}=\sigma(T(x_{\mathrm{cle}})_{i:},\tau)$，其中$\tau$为温度系数，可以表示为

$$\mathcal{L}_{\mathrm{crd}}=\frac{1}{N^{\mathrm{b}}}\sum_{i=1}^{N^{\mathrm{b}}}\sum_{j=1}^{C}W_{ij}^{\mathrm{c}}\cdot\left|L_{ij}^{\mathrm{tc}}-L_{ij}^{\mathrm{sc}}\right|, \tag{14}$$

其中类感知排序权重$W_{i:}^{\mathrm{c}}$为教师模型对干净样本i的类别预测，体现了在该标签样本下不同类别的语义关联。相关研究[34]指出，如果错误候选类别与标签

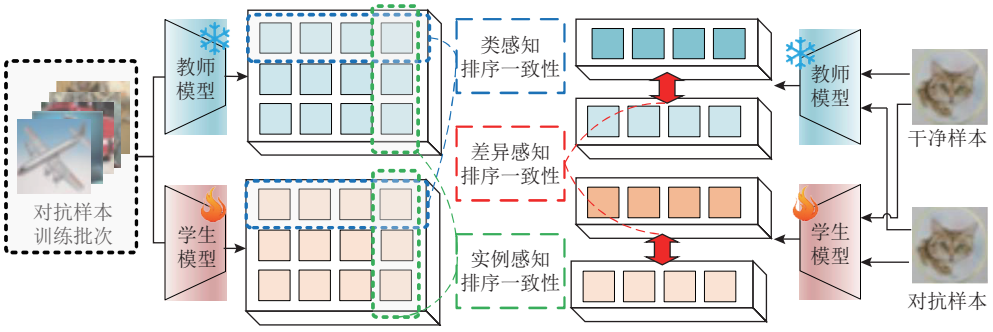


Fig. 2 Illustration of multi-level ranking consistency distillation framework

图 2 多层次排序一致性蒸馏框架示意图

类别具有强语义关系，这些类别普遍将成为对抗样本的扰动方向，而$W_{i:}^{\mathrm{c}}$为类感知排序一致性蒸馏$\mathcal{L}_{\mathrm{crd}}$提供了鲁棒性排序学习的重要类别选择能力。

2)实例感知排序一致性蒸馏。区别于主流对抗蒸馏方法仅关注类别知识，实例感知排序一致性蒸馏$\mathcal{L}_{\mathrm{ird}}$聚焦不同对抗样本实例与特定类别的语义相似性排序，此时待排序元素是一个训练批次N^{b}内全部的对抗样本。以学生模型$S(\cdot)$中实例感知排序为例，模型对N^{b}个训练样本在类别j上的响应可表示为$S(x_{\mathrm{adv}})_{:j}\in\mathbb{R}^{N^{\mathrm{b}}}$，则第$i$个对抗样本的优先级排序$L_{ij}^{\mathrm{sI}}$可表示为

$$L_{ij}^{\mathrm{sI}}=1+\sum_{k=1,k\neq i}^{N^{\mathrm{b}}}\frac{1}{2}\left(\tanh\left(S(x_{\mathrm{adv}})_{ij}-S(x_{\mathrm{adv}})_{kj}\right)+1\right)。\tag{15}$$

计算得到L_{ij}^{sI}和L_{ij}^{tI}，与$\mathcal{L}_{\mathrm{crd}}$类似我们额外加入了干净样本引导的实例感知权重$W_{:j}^{\mathrm{I}}=\sigma(T(x_{\mathrm{cle}})_{:j},\tau)$以区分不同对抗样本在特定类别排序计算中的语义关联性，最后的实例感知排序一致性蒸馏损失$\mathcal{L}_{\mathrm{ird}}$表示为

$$\mathcal{L}_{\mathrm{ird}}=\frac{1}{N^{\mathrm{b}}}\sum_{i=1}^{N^{\mathrm{b}}}\sum_{j=1}^{C}W_{ij}^{\mathrm{I}}\cdot\left|L_{ij}^{\mathrm{tI}}-L_{ij}^{\mathrm{sI}}\right|。\tag{16}$$

3)差异感知排序一致性蒸馏。现有对抗蒸馏方法依赖于对抗样本作为输入进行显示鲁棒知识传递，普遍忽略了对抗和干净样本之间的模型响应差异，如模型面向同一干净和对抗样本之间的类别预测偏移。本文认为这种预测差异能够对对抗样本攻击方向、模型抵御能力等隐式鲁棒知识性知识进行建模。针对这一点，差异感知排序一致性蒸馏$\mathcal{L}_{\mathrm{drd}}$聚焦于不同类别在对抗或干净样本下预测差异的重要性排序，将隐式鲁棒知识从大模型传递给小模型。具体地，先构建学生模型针对样本i的鲁棒性-准确率差异$D^{\mathrm{s}}(x)_{i:}=S(x_{\mathrm{adv}})_{i:}-S(x_{\mathrm{cle}})_{i:}$，同理可得教师模型差异$D^{\mathrm{t}}(x)_{i:}$。差异感知排序$L_{ij}^{\mathrm{sD}}$关注模型从预测干净样本到对抗样本的过程中，不同类别预测变化的程度并进行排序，其排序L_{ij}^{sD}以及一致性蒸馏损失$\mathcal{L}_{\mathrm{drd}}$可表达为

$$\begin{aligned}L_{ij}^{\mathrm{sD}}&=1+\sum_{k=1,k\neq j}^{C}\frac{1}{2}\left(\tanh\left(D^{\mathrm{s}}(x)_{ij}-D^{\mathrm{t}}(x)_{ik}\right)+1\right),\\ \mathcal{L}_{\mathrm{drd}}&=\frac{1}{N^{\mathrm{b}}}\sum_{i=1}^{N^{\mathrm{b}}}\sum_{j=1}^{d}\left|L_{ij}^{\mathrm{tD}}-L_{ij}^{\mathrm{sD}}\right|。\end{aligned}\tag{17}$$

基于上述介绍的约束机制和蒸馏框架，最终的训练损失函数综合考虑 3 种层次的知识蒸馏，可表示为

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{crd}}+\lambda\cdot\mathcal{L}_{\mathrm{ird}}+\mu\cdot\mathcal{L}_{\mathrm{drd}},\tag{18}$$

其中λ，μ为权重因子，用于动态调整不同层次的一致性约束在对抗排序蒸馏中的贡献，通过第 4 节实验证明，最终对抗排序损失函数的简单取值为$\lambda=\mu=1$时效能达到最优。

3.3 数学证明分析

本节从梯度幅值上界以及梯度幅值衰减 2 个方面，对优先级排序约束相较于 KL 散度约束的优越性进行数学证明。其中，显式的上界约束确保参数更新量可控，防止优化过程因梯度幅值过大而震荡或发散；而具备梯度幅值衰减能力可促使模型减弱远离决策边界对抗样本对梯度的贡献，提升模型的鲁棒性。

定理 1. 排序约束的梯度幅值上界。

排序约束的梯度幅值满足以下不等式：

$$\left\|\frac{\partial\mathcal{L}_{\mathrm{crd}}}{\partial z_k^{\mathrm{s}}}\right\|\leqslant\frac{1}{2}(C-1),\ \forall k\in\{1,2,\cdots,C\}。\tag{19}$$

证明.

对任一类别的输出z_k^{s}，比较对贡献计算得到

$$\frac{\partial\mathcal{L}_{\mathrm{crd}}}{\partial z_k^{\mathrm{s}}}=\frac{1}{2}\sum_{j\neq k}\mathrm{sgn}\left(z_k^{\mathrm{t}}-z_j^{\mathrm{t}}-z_k^{\mathrm{s}}+z_j^{\mathrm{s}}\right)\cdot sech^2(z_k^{\mathrm{s}}-z_j^{\mathrm{s}})。\tag{20}$$

由于$sech^2(x)\leqslant 1$对所有实数x成立，梯度幅值满足：

$$\left|\frac{\partial\mathcal{L}_{\mathrm{crd}}}{\partial z_k^{\mathrm{s}}}\right|\leqslant\frac{1}{2}\sum_{j\neq k}1=\frac{1}{2}(C-1)。\tag{21}$$

证毕.

推论 1. 总梯度范数满足

$$\|\nabla\mathcal{L}_{\mathrm{crd}}\|_2\leqslant\frac{1}{2}C\sqrt{C-1}。\tag{22}$$

定理 2. KL 散度的梯度幅值上界。

KL 散度的梯度幅值满足

$$\left\| \frac{\partial L_{\mathrm{KL}}}{\partial z_k^{\mathrm{s}}} \right\| \leqslant 1-\frac{1}{C},\ \forall k \in \{1,2,\cdots,C\} 。 \tag{23}$$

证明.

KL 散度的梯度为

$$\frac{\partial L_{\mathrm{KL}}}{\partial z_k^{\mathrm{s}}} = \mathbb{E}_x \left[\sigma(\boldsymbol{Z}^{\mathrm{s}})_k - \sigma(\boldsymbol{Z}^{\mathrm{t}})_k \right] 。 \tag{24}$$

由于$\sigma(\boldsymbol{Z}^{\mathrm{s}})_k \in (0,1)$且$\sum_{k=1}^{C} \sigma(\boldsymbol{Z}^{\mathrm{s}})_k = 1$，当类别数 C 较大时，$\sigma(\boldsymbol{Z}^{\mathrm{s}})_k \approx \frac{1}{C}$。因此

$$\left| \frac{\partial L_{\mathrm{KL}}}{\partial z_k^{\mathrm{s}}} \right| \leqslant \max\left(\sigma(\boldsymbol{Z}^{\mathrm{s}})_k, \sigma(\boldsymbol{Z}^{\mathrm{t}})_k \right) \leqslant 1-\frac{1}{C} 。 \tag{25}$$

证毕.

梯度幅值上限结论：排序约束的梯度幅值上界严格低于 KL 散度，训练更具稳定性，即

$$\frac{1}{2}(C-1) < C\left(1-\frac{1}{C}\right) 。 \tag{26}$$

定理 3. KL 散度的梯度幅值无衰减性。

当学生模型的某一类别逻辑值z_k^{s}显著高于其他类别（即 $z_k^{\mathrm{s}} \gg z_j^{\mathrm{s}}, \forall j \neq k$）时，KL 散度的梯度幅值不随$z_k^{\mathrm{s}} - z_j^{\mathrm{s}}$的增大而衰减。

证明.

假设$z_k^{\mathrm{s}} \to +\infty$，而其他 z_j^{s}固定或有限，则 Softmax 概率近似为

$$\sigma(\boldsymbol{Z}^{\mathrm{s}})_k \approx 1,\ \sigma(\boldsymbol{Z}^{\mathrm{s}})_j \approx 0 (\forall j \neq k) 。 \tag{27}$$

此时，梯度表达式简化为

$$\frac{\partial L_{\mathrm{KL}}}{\partial z_k^{\mathrm{s}}} \approx \mathbb{E}_x \left[1-\sigma\left(\boldsymbol{Z}^{\mathrm{t}}(x)\right)_k \right] , \tag{28}$$

梯度幅值为

$$\left| \frac{\partial L_{\mathrm{KL}}}{\partial z_k^{\mathrm{s}}} \right| \approx \left| 1-\mathbb{E}_x \left[\sigma\left(\boldsymbol{Z}^{\mathrm{t}}(x)\right)_k \right] \right| , \tag{29}$$

该幅值仅取决于教师模型的概率分布$\sigma(\boldsymbol{Z}^{t})_k$，与学生模型输出逻辑值的绝对差异$z_k^{\mathrm{s}} - z_j^{\mathrm{s}}$无关。因此，KL 散度的梯度幅值在$z_k^{\mathrm{s}} \gg z_j^{\mathrm{s}}$时不会自动衰减。

证毕.

定理 4. 排序约束的梯度指数衰减性。

当学生模型的某一类别逻辑值z_k^{s}显著高于其他类别（即$z_k^{\mathrm{s}} \gg z_j^{\mathrm{s}}$）时，排序约束的梯度幅值随$z_k^{\mathrm{s}} - z_j^{\mathrm{s}}$的增大而指数衰减：

$$\left| \frac{\partial \mathcal{L}_{\mathrm{crd}}}{\partial z_k^{\mathrm{s}}} \right| \leqslant \frac{\gamma}{2} \sum_{j \neq k} 4\mathrm{e}^{-2\gamma(z_k^{\mathrm{s}} - z_j^{\mathrm{s}})} 。 \tag{30}$$

证明.

由双曲正割函数的性质$sech^2(x) \leqslant 4\mathrm{e}^{-2|x|}$，当$z_k^{\mathrm{s}} \gg z_j^{\mathrm{s}}$时

$$sech^2(\gamma(z_k^{\mathrm{s}} - z_j^{\mathrm{s}})) \leqslant 4\mathrm{e}^{-2\gamma(z_k^{\mathrm{s}} - z_j^{\mathrm{s}})} 。 \tag{31}$$

因此，梯度幅值满足

$$\left| \frac{\partial \mathcal{L}_{\mathrm{crd}}}{\partial z_k^{\mathrm{s}}} \right| \leqslant \frac{\gamma}{2} \sum_{j \neq k} 4\mathrm{e}^{-2\gamma(z_k^{\mathrm{s}} - z_j^{\mathrm{s}})} = 2\gamma \sum_{j \neq k} \mathrm{e}^{-2\gamma(z_k^{\mathrm{s}} - z_j^{\mathrm{s}})} , \tag{32}$$

当$z_k^{\mathrm{s}} - z_j^{\mathrm{s}} \to +\infty$时，指数项$\mathrm{e}^{-2\gamma(z_k^{\mathrm{s}} - z_j^{\mathrm{s}})} \to 0$，梯度幅值指数衰减至 0。

证毕.

梯度衰减性结论：梯度幅值上界低是训练稳定性的基础，但梯度衰减性直接决定收敛率，对训练的影响程度更高。常规 KL 散度约束梯度幅值不随学生模型输出差异增大而衰减，导致优化过程易受远离决策边界样本的干扰；而排序化约束通过双曲正割函数的指数衰减特性，梯度幅值在输出差异显著时自动收缩，提升训练稳定性与效率。

4 实验分析

本节主要研究 6 个问题：

1）对抗排序蒸馏对小模型鲁棒性的提升效果如何。

2）对抗排序蒸馏的每个主要组件是否对小模型鲁棒性提升做出贡献。

3）对抗排序蒸馏的主要参数对小模型鲁棒性产生了怎样的影响。

4）对抗排序蒸馏在教师-学生模型一致的对抗自蒸馏场景下能否保持提升效果。

5）对抗排序蒸馏在目标模型未知的白盒攻击场景下能否保持提升效果。

6）对抗排序蒸馏在训练数据受限场景下能否保持提升效果。

为讨论以上 6 个研究问题，本节首先在 4.1 节分别介绍数据集、对比方法、网络架构、训练配置、评价指标等相关实验设置，并在 4.2~4.6 节分别对这些研究问题展开针对性实验分析。

4.1 实验设置

实验数据选择为在对抗防御研究中常用的 CIFAR10[35], CIFAR100[35], Tiny ImageNet[36] 数据集，对应候选类别分别是 10, 100, 200, 这 3 个数据集中图像的像素范围均被归一化至 [0,1] 的区间内。在此基础上我们选取了 9 种广泛使用且高效的小模型对抗防御方法，包括对抗训练代表方法 PGD-AT[36] 和 TRADES[37], 以及对抗蒸馏代表方法 ARD[13], IAD[20], RSLAD[16], AdaAD[21], MTARD[19], Fair-ARD[18], DGAD[22]。所有对比方法的实验配置均参照其原文推荐的最优

参数, 并在其公开代码基础上实现。

1)基础网络架构。根据现有对抗蒸馏工作[21-22]中的设置, 在 CIFAR10 数据集中选择 WideResNet-34-10(参数量为 66.5×10^6)、WideResNet-34-20(参数量为 125×10^6)以及 WideResNet-28-10(36.5×10^6)作为教师模型, 而在 CIFAR100, Tiny ImageNet 中教师模型选择为 WideResNet-34-10, 所有教师模型事先通过 TRADES 进行预先训练来构建对抗鲁棒能力。学生模型方面, 3 个数据集中均使用 ResNet-18(11.7×10^6)和 MobileNet-v2(3.4×10^6)作为轻量级模型进行对抗训练。

2)基本训练配置。我们使用 SGD 动量优化器训练模型, 其中初始学习率设置为 0.1, 动量为 0.9, 权重衰减为 5E−4。遵照主流对抗蒸馏工作[21-22], 对抗训练的数据批次大小设置为 128。对于 PGD-AT, 我们采用 110 个训练周期, 并使用提前停止策略[31,35]; 而对于 TRADES 和其他自适应对抗方法, 我们使用 200 个训练周期, 并且在第 100 和第 150 个周期时将学习率缩小为原来的 1/10。除非另有说明, 内部优化的迭代次数设置为 10, 步长为 2/255, 总扰动界限在 $L\infty$约束下为 8/255。用于对比的对抗蒸馏方法, 实验结果依据相关公开代码, 并遵从其文章中所推荐的最优参数进行设置。所有实验通过配置有 2 张 NVIDIA RTX4090 显卡的服务器实现, 并基于 Pytorch 框架和 advertorch 库。

3)评价指标。除了识别准确率(Clean)外, 实验还使用了 3 种对抗防御研究中普遍使用的典型对抗攻击方法 FGSM, PGD, CW 作为鲁棒性评估指标。同时, 我们还采用了一种 RobustBench[38] 中的评估工具, 一种强攻击策略 AutoAttack(AA)深入评估对抗鲁棒性。对于 FGSM, PGD, AA, 最大扰动大小设置为 8/255, 而 PGD 采用 10 个步骤, 步长为 2/255。CW 受到 L2 范数约束, 其平衡常数设置为 0.1。除非另有说明, 最终模型参数选择基于训练中在 PGD-10 准确率最高的检查点上的结果。

4.2 方法对比实验

本节针对在同一教师模型下面向不同对抗防御方法的准确率-鲁棒性效能对比分析。具体地, 以 WideResNet-34-10 为教师模型, ResNet-18 和 MobileNet-v2 为学生模型, 所提出的对抗排序蒸馏方法和 9 种典型对抗防御方法的实验结果如表 2 所示。

Table 2 Method Recognition Accuracy and Adversarial Robustness Based on WideResNet-34-10 Teacher Model in CIFAR-10 Dataset

表 2 CIFAR-10 数据集下基于 WideResNet-34-10 教师模型的方法识别准确率与对抗鲁棒性 %

学生模型	方法	识别准确率		对抗鲁棒性							
		Clean(best)	Clean(last)	FGSM(best)	FGSM(last)	PGD(best)	PGD(last)	CW(best)	CW(last)	AA(best)	AA(last)
ResNet-18	PGD-AT	82.95		57.17		52.87		77.56		47.69	
	TRADES	83.00		58.42		53.18		76.92		49.21	
	ARD	83.13	82.97	57.94	57.32	53.55	52.91	78.63	78.55	48.77	48.32
	IAD	83.50	83.18	60.97	60.85	55.71	55.32	79.44	79.31	51.01	50.72
	RSLAD	83.78	83.51	61.13	60.78	56.45	56.32	79.56	79.39	51.71	51.34
	MTARD	84.51	84.32	59.83	59.31	54.52	54.39	79.21	79.19	50.53	50.11
	FAIR-ARD	83.51	83.42	57.98	57.71	53.85	53.77	78.50	78.31	48.51	48.50
	AdaAD	84.54	84.15	60.14	59.83	56.12	55.96	80.38	80.18	51.50	51.34
	DGAD	84.73	84.51	60.98	60.77	57.35	57.31	80.51	80.58	51.83	51.92
	本文	**84.82**	**84.71**	**62.33**	**62.14**	**58.59**	**58.49**	**80.85**	**80.83**	**53.78**	**53.72**
MobileNet-v2	PGD-AT	77.54		53.58		49.90		72.54		44.56	
	TRADES	79.80		54.84		50.51		75.31		45.67	
	ARD	80.54	80.15	55.16	54.91	50.97	50.81	76.93	76.61	46.57	46.51
	IAD	81.34	81.00	56.01	55.71	52.21	51.98	77.33	77.21	47.93	47.85
	RSLAD	82.44	81.97	56.13	55.98	52.91	52.77	77.51	77.50	48.31	48.32
	MTARD	83.13	82.87	55.83	55.51	51.15	50.97	77.29	77.16	47.81	47.70
	FAIR-ARD	81.33	81.27	55.83	55.71	51.45	51.39	76.81	76.85	46.88	46.81
	AdaAD	82.80	82.56	56.81	56.75	53.16	52.88	78.31	78.29	48.04	47.51
	DGAD	83.87	83.58	58.33	58.21	55.29	55.28	79.35	79.03	50.17	50.06
	本文	**84.45**	**84.22**	**61.20**	**60.88**	**56.94**	**56.87**	**80.39**	**80.09**	**52.05**	**51.84**

注: 黑体数值表示最优结果, 下划线数值表示次优结果。best 为最高检查点结果, last 为 200 训练周期结束时的检查点结果。

首先，8 种对抗蒸馏策略的准确率-鲁棒性效能显著优于 PGD-AT 和 TRADES 两种典型对抗训练方法，这说明教师-学生模型知识传递策略是提升小模型鲁棒性的高效范式。其次，我们所提出的对抗排序蒸馏方法在准确率和全部对抗指标达到最优，其中典型鲁棒性指标 PGD 和 AA 相比于当前先进的 AdaAD 和 DGAD 方法能够取得压倒性优势，这种性能优势可归因于对抗排序蒸馏在鲁棒性知识建模以及传递的高效设计；最后，相比于其他对抗蒸馏策略，本文方法在 PGD-10 最高检查点 PGD(best)和训练结束检查点 PGD(last)间测试结果差异较小，这说明对抗排序蒸馏中松耦合约束机制能够有效提升小模型对抗训练过程的稳定性，缓解潜在的对抗过拟合问题。

图 3 与图 4 体现了在不同教师-学生模型组合下对抗蒸馏方法的性能比较，其中横坐标为学生-教师模型参数比例(数值越小说明大小模型能力差距越大)，而纵坐标为准确率和鲁棒性指标(PGD, AA)。总体而言，对抗排序蒸馏策略在 6 种教师-学生模型组合下均实现了最优结果，体现了其性能高效以及场景泛用性。同时，相比于其他对抗蒸馏策略，对抗排序蒸馏方法随教师模型参数增大呈现明显能力上升趋势，说明该方法能够有效缓解因大小模型能力差异导致的学生网络学习难度提高问题，充分发挥教师模型能力优势。最后，表 3 与表 4 从不同实验数据集视角对本文提出的第 1 个研究问题进行分析，本文方法仍有显著的鲁棒性能优势，彰显了该方法的泛用性和性能稳定性。

4.3 算法有效性测试

本节聚焦研究问题 2 和问题 3，即验证所提出对抗排序蒸馏方法中组件的有效性，以及超参数的影响。首先针对问题 2，基于 CIFAR10 数据集以及 WideResNet-34-10 教师模型，分别对 2 种不同学生模

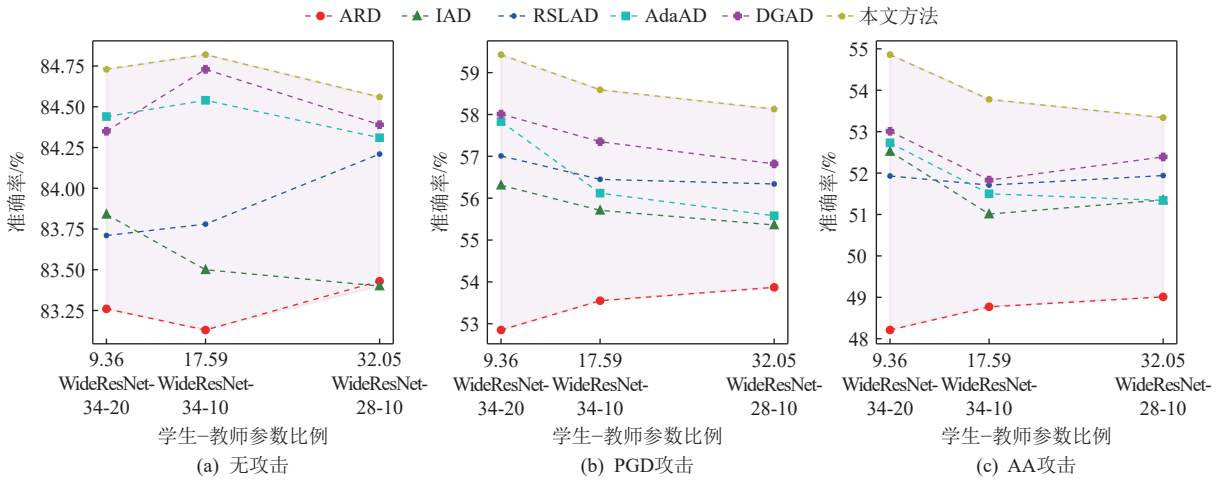


Fig. 3 Accuracy without attack of ResNet-18 student model guided by different teacher models

图 3 ResNet-18 学生模型在不同教师模型引导下的无攻击准确率

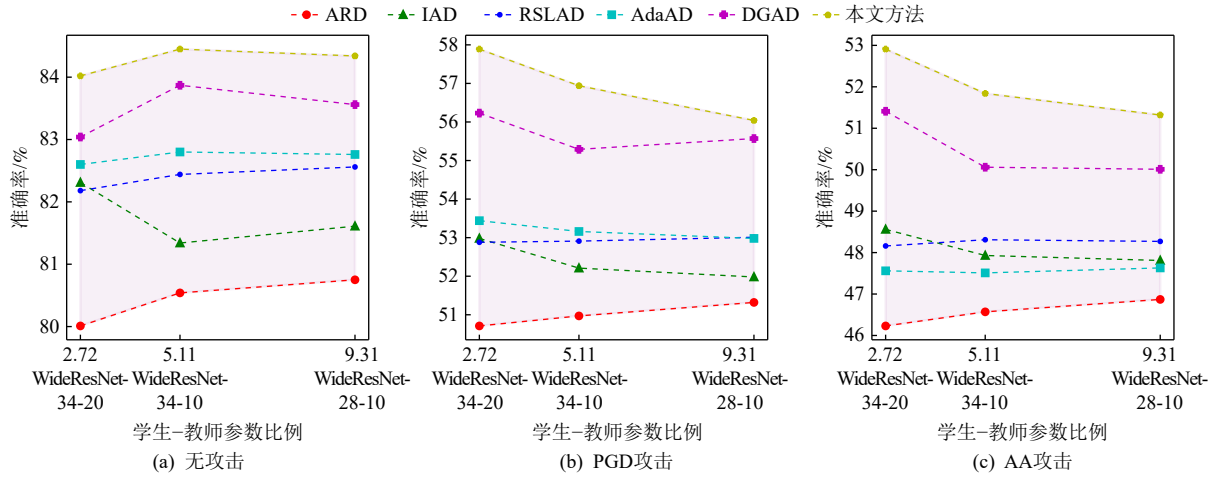


Fig. 4 Accuracy without attack of MobileNet-v2 student model guided by different teacher models

图 4 MobileNet-v2 学生模型在不同教师模型引导下的无攻击准确率

Table 3 Method Recognition Accuracy and Adversarial Robustness Based on WideResNet-34-10 Teacher Model in CIFAR-100 Dataset

表 3 CIFAR-100 数据集下基于 WideResNet-34-10 教师模型的方法识别准确率与对抗鲁棒性

方法	ResNet-18					MobileNet-v2				
	Clean	FGSM	PGD	CW	AA	Clean	FGSM	PGD	CW	AA
PGD-AT	56.27	32.08	29.84	49.05	24.99	51.55	29.34	27.26	45.73	22.07
TRADES	57.82	32.52	30.38	51.30	25.02	53.05	29.07	27.44	47.62	21.82
ARD	60.94	35.31	32.72	53.67	26.04	57.18	33.13	30.91	51.50	24.20
IAD	60.43	35.75	32.80	52.71	26.84	56.33	32.88	30.18	49.00	24.07
RSLAD	59.55	35.68	33.35	52.89	27.77	56.04	32.76	30.29	50.14	24.56
AdaAD	62.19	35.33	32.52	54.67	26.74	61.44	34.75	31.97	54.21	25.91
DGAD	63.24	36.09	33.68	55.47	27.66	**62.25**	34.90	32.64	54.54	26.56
本文	**63.58**	**39.21**	**35.58**	**55.63**	**28.32**	61.98	**36.45**	**33.71**	**54.46**	**26.83**

注：黑体数值表示最优结果。

Table 4 Method Recognition Accuracy and Adversarial Robustness Based on WideResNet-34-10 Teacher Models in TinyImageNet Dataset

表 4 TinyImageNet 数据集下基于 WideResNet-34-10 教师模型的方法识别准确率与对抗鲁棒性

方法	ResNet-18					MobileNet-v2				
	Clean	FGSM	PGD	CW	AA	Clean	FGSM	PGD	CW	AA
ARD	41.66	24.47	23.30	37.76	17.23	38.79	21.51	22.17	34.78	14.74
IAD	44.57	23.97	22.15	37.54	17.58	40.15	21.98	23.18	35.88	15.98
RSLAD	43.62	23.45	22.58	37.12	17.05	41.37	22.12	22.95	35.53	16.17
AdaAD	47.54	24.22	22.82	42.79	17.41	43.60	22.45	23.13	36.57	16.44
DGAD	47.92	24.42	23.05	42.81	17.18	**45.27**	23.19	23.25	36.91	16.87
本文	**48.15**	**24.98**	**24.03**	**42.83**	**18.57**	45.19	**23.77**	**23.53**	**37.71**	**17.11**

注：黑体数值表示最优结果。

型下方法的重要组件有效性进行探索。实验结果如表 5 所示，其中方法组件 KL 为常规对抗蒸馏方法使用的 KL 散度约束，SP 为典型样本关系感知的知识蒸馏策略，CRD, IRD, DRD 分别为本文方法中所涉及的 3 个核心组件模块，即类感知、实例感知、差异感知排序一致性蒸馏。

首先，由表 5 中案例 1, 2, 3 与案例 5 的实验结果进行对比可知，相比于主流知识蒸馏的刚性 KL 散度约束，对抗排序蒸馏中基于优先级排序约束机制下的类感知一致性蒸馏在准确率，以及 3 种不同对抗攻击下的鲁棒性表现上均呈现压倒性性能优势。尽管 KL 散度可以通过调整温度系数 T 进行松弛调整，但各方面效果仍与本文提出的排序约束机制有明显差距。其次，通过案例 4 和案例 7 的实验结果可知，引入典型的实例关系蒸馏方法 SP 对轻量级模型鲁棒性提升十分有限。相反，同样是对训练批次中样本关系进行知识传递，本文中所提出的实例感知排序一致性蒸馏（案例 6）充分发挥排序约束的松耦合优势，在类感知的基础上全面提升了小模型准确率和鲁棒性。最后，案例 9 引入差异感知排序一致性蒸馏后，完整的对抗排序蒸馏方法能够在维持准确率的基础上，充分挖掘教师模型在干净/对抗样本响应间的差异性知识，全面提升学生模型鲁棒性效能。

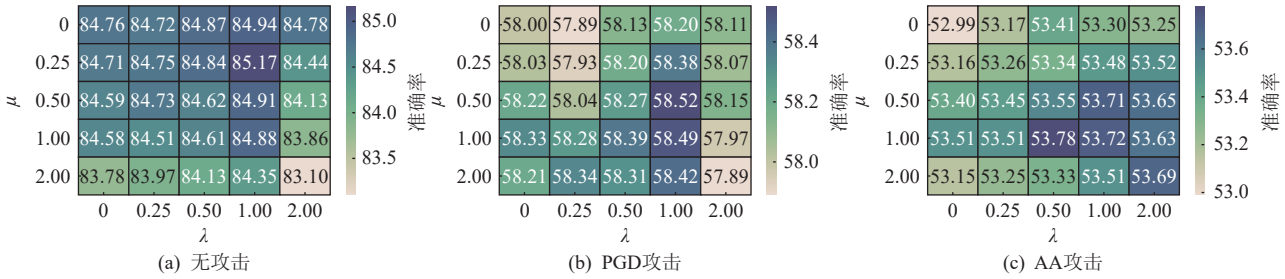
针对问题 3，图 5（ResNet-18）和图 6（MobileNet-v2）共同讨论了对抗排序蒸馏总训练损失（式（18））中权重调整因子 λ, μ 的不同取值对于该方法性能的影响，其中将 λ, μ 分别在 {0, 0.25, 0.5, 1, 2, 4} 的合理区间内进行动态调整取值。由实验结果可得，当差异感知排序蒸馏参数 μ 取值在 [0, 1] 区间时能够维持准确率的前提下大幅提高 PGD, AA 这 2 种鲁棒性指标性能，

Table 5 Ablation Study Based on Adversarial Ranking Distillation

表 5 基于对抗排序蒸馏的消融实验

案例	方法组件					ResNet-18				MobileNet-v2			
	KL	SP	CRD	IRD	DRD	Clean	FGSM	PGD	AA	Clean	FGSM	PGD	AA
1	✔($T=1$)					84.44	59.94	55.97	51.12	82.80	56.81	53.16	48.04
2	✔($T=2$)					84.20	60.10	56.13	51.17	82.65	56.95	53.45	48.22
3	✔($T=3$)					84.01	59.75	55.63	50.98	82.62	56.64	52.99	47.93
4	✔($T=1$)	✔				84.55	59.87	55.79	51.05	82.54	56.49	52.78	47.82
5			✔			84.76	62.09	58.00	52.99	83.93	60.03	55.57	50.81
6			✔	✔		84.94	62.42	58.20	53.30	84.68	60.88	56.36	51.09
7		✔	✔			84.88	61.97	57.92	53.01	84.00	60.12	55.61	50.83
8			✔		✔	84.58	62.17	58.33	53.51	83.96	61.08	56.58	51.53
9			✔	✔	✔	84.90	62.53	58.49	53.72	84.45	61.20	56.95	52.05

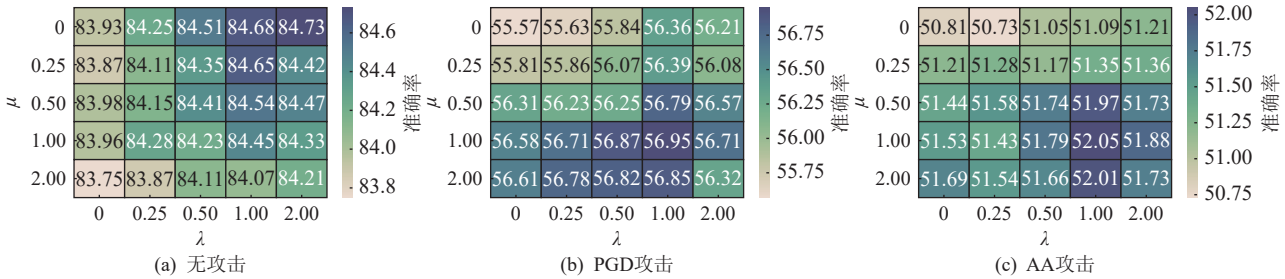
注：T 表示温度变量。



(a) 无攻击 (b) PGD攻击 (c) AA攻击

Fig. 5 Accuracy of ResNet-18 student model under different values of λ and μ

图 5 ResNet-18 学生模型在不同λ，μ取值下的准确率



(a) 无攻击 (b) PGD攻击 (c) AA攻击

Fig. 6 Accuracy of MobileNet-v2 student model under different values of λ and μ

图 6 MobileNet-v2 学生模型在不同λ, μ取值下的准确率

而取值为 2 时容易降低干净样本准确率，而实例感知蒸馏参数同样在 λ=1 取得最好实验表现。由此，在进行简单超参数探索并取值 λ=μ=1 时，本文所提出方法得到最佳效果。

4.4 对抗自蒸馏场景对比实验

除了轻量级模型鲁棒性，知识蒸馏在教师-学生模型网络结构一致的对抗自蒸馏场景下普遍也被认为是提升对抗训练效果的关键技术，因此本节讨论所提出对抗排序蒸馏方法在对抗自蒸馏任务下是否能保持其优越性能(问题 6)。此过程首先用对抗训练得到教师模型，并使用同模型结构的学生网络进行对抗鲁棒知识迁移。本节采用 ResNet-18 和 MobileNet-v2 的学生网络框架在 CIFAR10 数据集上进行自蒸馏对比实验，结果如表 6 所示。

由于对抗自蒸馏弱化了教师-学生模型间网络参数以及表征能力差距，因此相较于表 2 实验结果，本文方法在性能方面有所减弱，但准确率和鲁棒性表现仍明显强于 AdaAD 和 DGAD 等先进方法，其中 PGD，AA 指标稳定提升 1.5 个百分点和 1 个百分点以上。

4.5 黑盒攻击场景对比实验

由于现实应用中攻击目标模型的网络结构普遍未知其特点，本节从黑盒攻击场景深入探索本文方法的性能优越性和实用性(问题 5)。具体地，针对在教师模型 WideResNet-34-10 上进行对抗蒸馏得到的学生模型 ResNet-18 以及 MobileNet-v2，我们以

Table 6 Method Recognition Accuracy and Adversarial Robustness in Adversarial Self-Distillation Scenario

表 6 对抗自蒸馏场景下的方法识别准确率与对抗鲁棒性

方法	ResNet-18					MobileNet-v2				
	Clean	FGSM	PGD	CW	AA	Clean	FGSM	PGD	CW	AA
ARD	80.66	55.68	50.90	74.87	46.61	79.42	53.18	50.90	73.14	44.76
IAD	81.32	57.54	52.91	75.69	48.20	79.81	54.83	51.28	73.58	44.98
RSLAD	81.92	57.94	53.29	76.26	49.06	79.78	54.67	51.48	73.49	45.51
AdaAD	83.13	57.54	53.30	77.62	49.61	81.90	55.86	51.83	75.83	46.92
DGAD	83.26	58.91	54.37	77.90	50.54	82.75	56.23	52.15	76.17	46.73
本文	**84.01**	**59.33**	**55.83**	**78.12**	**51.68**	**83.24**	**58.07**	**53.87**	**77.23**	**48.59**

注：黑体数值表示最优结果。

ResNet34 和 VGG16 网络作为元模型对本文方法和相关策略进行迁移攻击，实验结果如表 7 所示。总体而言，与白盒攻击策略下的实验结果趋势一致，所提出的对抗排序蒸馏方法在 3 种不同鲁棒性评估指标下均取得了优越性能，PGD 性能稳定优于 DGAD 方法 1 个百分点以上。

4.6 低资源场景对比实验

在边缘计算等实际应用中，除了设备模型参数轻量化需求外，训练数据资源受限也是方法设计需考虑的现实问题。为探索对抗排序蒸馏方法在低训练资源场景下的实用性(问题 6)，本节在学生网络训练数据集仅有原数据集 5%，10%，25% 样本比例下，以 WideResNet-34-10 为教师模型，对学生模型 ResNet-18

Table 7 Method Recognition Accuracy and Adversarial Robustness in Black-Box Scenario

表 7 黑盒场景下的方法识别准确率与对抗鲁棒

黑箱模型	性能指标	ResNet-18					MobileNet-v2				
		ARD	IAD	AdaAD	DGAD	本文	ARD	IAD	AdaAD	DGAD	本文
ResNet34	FGSM	66.63	66.97	67.31	68.19	**69.06**	64.13	63.87	65.44	66.01	**67.30**
	PGD	64.75	64.57	65.33	65.74	**66.48**	63.75	63.81	64.54	65.21	**66.17**
	JSMA	86.29	86.13	87.35	88.54	**89.33**	84.23	85.13	86.97	87.15	**87.92**
VGG16	FGSM	67.24	67.31	68.94	69.33	**70.45**	65.37	65.41	66.01	67.54	**68.37**
	PGD	66.07	66.58	68.13	68.54	**70.08**	64.31	64.34	65.19	66.11	**67.23**
	JSMA	87.15	86.97	88.22	89.57	**89.93**	84.99	85.14	86.58	87.29	**88.10**

注：黑体数值表示最优结果。

以及 MobileNet-v2 继续训练，实现结果分别如图 7 与图 8 所示。

总体而言，尽管训练数据规模受限，相比其他先进对抗蒸馏策略，本文方法在准确率和鲁棒性效能上仍呈现压倒性优势。特别地，本文方法在数据比例 5% 场景下 FGSM, PGD, AA 的对抗鲁棒性性能，甚至优于 ARD, IAD, RSLAD 等方法在 10% 数据量上的表现。这种显著的鲁棒性优势可归因于 2 个方面：一是本文方法采用了优先级排序的松耦合且高效的约束机制，能够在模型参数/数据规模差异明显的条

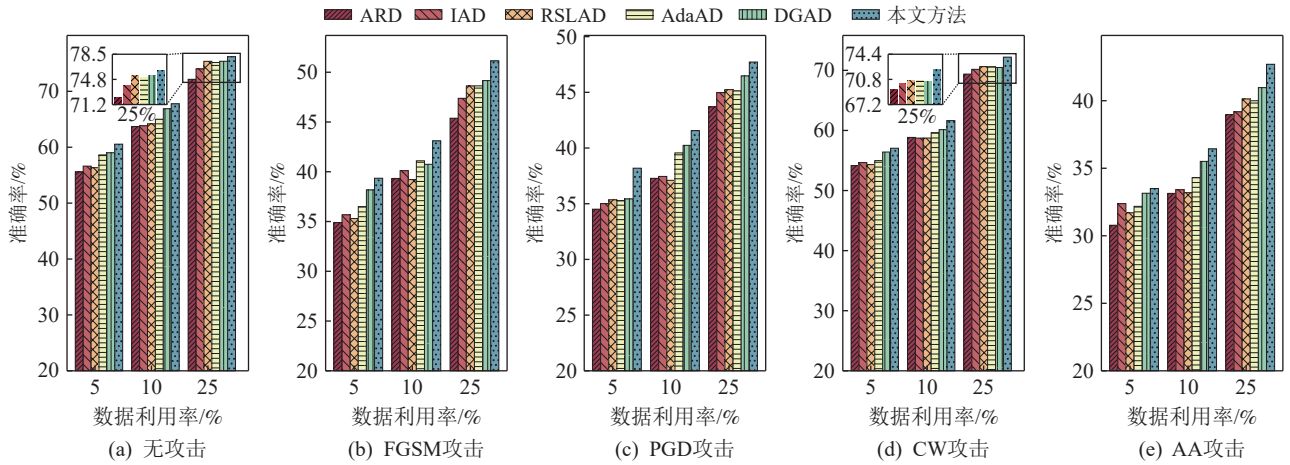


Fig. 7 Method recognition accuracy and adversarial robustness based on ResNet-18 student model in low resource scenarios

图 7 低资源场景下基于 ResNet-18 学生模型的方法识别准确率与对抗鲁棒性

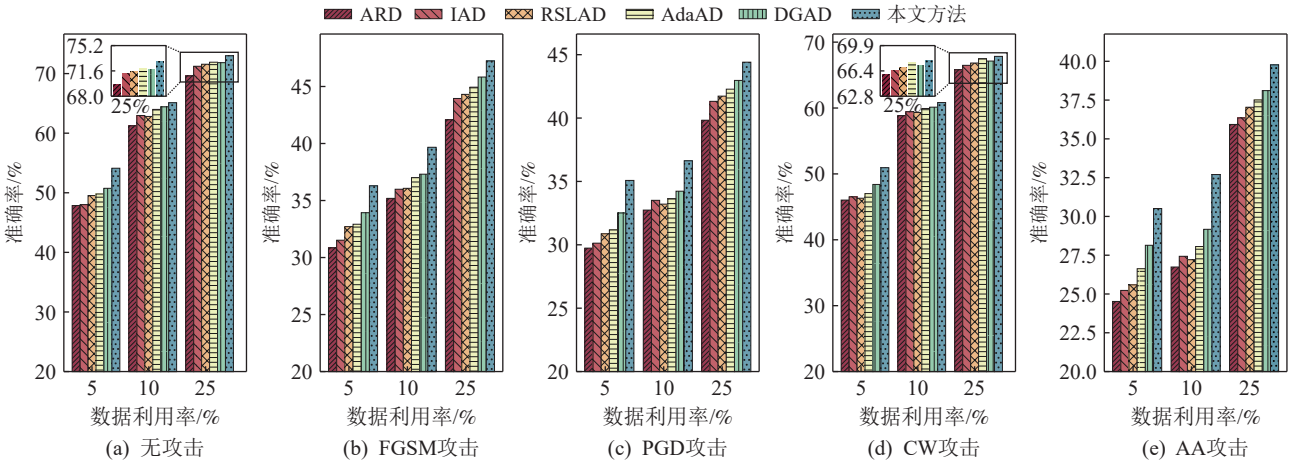


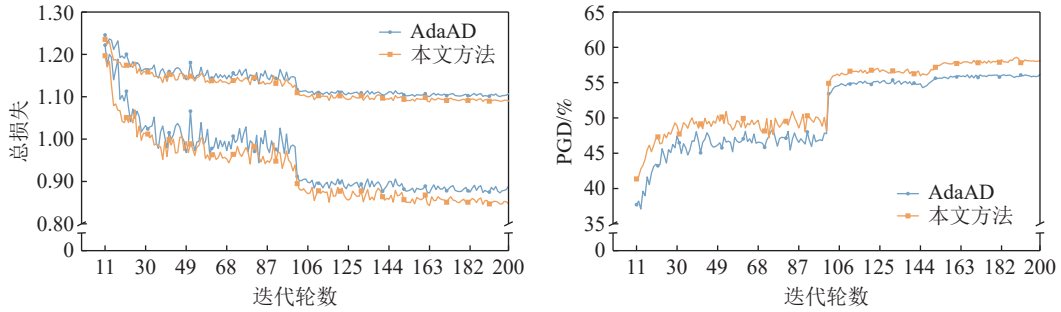
Fig. 8 Method recognition accuracy and adversarial robustness based on MobileNet-v2 student model in low resource scenarios

图 8 低资源场景下基于 MobileNet-v2 学生模型的方法识别准确率与对抗鲁棒性

件下实现高效的鲁棒知识传递；二是本文方法在优先级排序约束机制的基础上，从类别、实例关系、干净/对抗样本预测差异 3 个维度协同进行样本鲁棒性知识挖掘，充分提升小模型应对恶意攻击的韧性。

4.7 案例分析

为映证 3.3 节数学理论分析中对抗排序蒸馏方法在梯度幅值上界和梯度幅值衰减上训练可控的优势，本节以在 CIFAR-10 数据集上训练 ResNet-18 学生模型为案例，记录了训练过程中基于常规 KL 散度的 AdaAD 和本文方法在训练过程中损失函数衰减曲线，以及 PGD 对抗鲁棒性上升曲线。如图 9 所示，本文方法梯度幅值下降更快，总体趋势更平稳，且在后期能够维持较低水平；而比较之下，基于 KL 散度的 AdaAD 方法整体训练损失较高，且对抗鲁棒性存在明显劣势，映证了本文方法在数学证明上训练收敛的优越性。



(a) 训练过程中损失函数变化曲线　　(b) PGD对抗鲁棒性变化曲线

Fig. 9 Case study: Loss function and adversarial robustness variation curves during training for AdaAD and our method

图 9 案例分析：AdaAD 与本文方法在训练过程中损失函数和对抗鲁棒性变化曲线

5 结　论

对抗蒸馏方法作为提升轻量级模型对抗防御能力的关键技术，普遍忽视了对抗样本预测类别间关系的挖掘以及大小模型间鲁棒性学习能力的差异。为此，本文提出了一种新颖且高效的轻量级模型鲁棒性提升方法——对抗排序蒸馏，具体涉及 2 个重要组件：优先级排序约束机制提出了一种基于元素间对比排序实现大小模型间松耦合约束的通用策略；多层次排序一致性蒸馏框架则在此基础上，分别从类别语义关联、样本语义关联以及对抗差异关联 3 个方面进行高效对抗鲁棒性建模与知识迁移。实验结果表明，相比于现有先进的对抗蒸馏策略，本文方法能够在稳定维持轻量级模型干净样本准确率的同时显著提升对抗鲁棒性能。同时在训练样本受限、内部优化函数变化以及对抗自蒸馏的实验场景中，对抗排序蒸馏方法具备更强的适应力与泛用性。

下一步工作集中于理论创新方面，所提出的多层次排序一致性蒸馏框架采用动态搜索的方法对类别语义关联、样本语义关联以及对抗差异关联 3 种不同对抗鲁棒性知识传递进行动态赋权，尽管探索过程简单但仍缺乏部署灵活性，在接下来的研究工作中将进一步考虑设计自适应权重计算策略协同调整 3 种异质知识蒸馏的贡献度；应用实践方面，当前轻量级模型鲁棒性研究仅聚焦在视觉与图像识别应用领域，接下来会进一步将本文的工作扩展到轻量化本文识别以及语音分析模型的对抗鲁棒性研究中。

作者贡献声明：老明瑞负责方案整体设计和实验方法，并撰写论文；李正、张雪毅负责部分算法思路和实验；丁兆云、岳祥虎、郭延明提出指导意见并修改论文。岳祥虎和郭延明同为通信作者。

参 考 文 献

[1] Dou Hui, Zhang Lingming, Han Feng, et al. Survey on convolutional neural network interpretability[J] Journal of Software, 2024, 35(1): 159–184 (in Chinese)
(窦慧, 张凌茗, 韩峰, 等. 卷积神经网络的可解释性研究综述[J] 软件学报, 2024, 35(1): 159−184)

[2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436−444

[3] Rawat W, Wang Zenghui. Deep convolutional neural networks for image classification: A comprehensive review[J]. Neural Computation, 2017, 29(9): 2352−2449

[4] Zou Zhengxia, Chen Keyan, Shi Zhenwei, et al. Object detection in 20 years: A Survey[J]. Journal of the IEEE, 2023, 111(3): 257−276

[5] Chang Yupeng, Wang Xu, Wang Jindong, et al. A survey on evaluation of large language models[J]. ACM Transactions on

Intelligent Systems and Technology, 2024, 15(3): 1−45

[6] Alabdulmohsin I, Neyshabur B, Zhai Xiaohua. Revisiting neural scaling laws in language and vision[J]. Advances in Neural Information Processing Systems, 2022, 35: 22300−22312

[7] Chakraborty A, Alam M, Dey V, et al. A survey on adversarial attacks and defences[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 25−45

[8] Ye Shaokai, Xu Kaidi, Liu Sijia, et al. Adversarial robustness vs. model compression, or both?[C/OL]//Proc of the 14th IEEE/CVF Int Conf on Computer Vision. 2019[2025-09-06]. https://openaccess.thecvf.com/content_ICCV_2019/html/Ye_Adversarial_Robustness_vs._Model_Compression_or_Both_ICCV_2019_paper.html

[9] Qian Zhuang, Huang Kaizhu, Wang Qiufeng, et al. A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies[J]. Pattern Recognition, 2022, 113: 1−11

[10] Pang Tianyu, Yang Xiao, Dong Yinpeng, et al. Bag of tricks for adversarial training[J]. arXiv preprint, arXiv: 2010.00467, 2020

[11] Wu Boxi, Chen Jinghui, Cai Deng, et al. Do wider neural networks really help adversarial robustness?[C/OL]//Proc of the 35th Neural Information ProcessingSystems. 2021[2025-09-06]. https://proceedings.neurips.cc/paper/2021/file/3937230de3c8041e4da6ac3246a888e8-Paper.pdf

[12] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[13] Micah G, Liam F, Soheil F, et al. Adversarially robust distillation[C/OL]//Proc of the 20th AAAI Conf on Artificial Intelligence. 2020[2025-09-06]. https://ojs.aaai.org/index.php/AAAI/issue/view/252

[14] Gou Jianping, Yu Baosheng, Stephen J M, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789−1819

[15] Abbasi S, Hajabdollahi M, Karimi N, et al. Modeling teacher-student techniques in deep neural networks for knowledge distillation[C]// Proc of the 2nd Int Conf on Machine Vision and Image Processing (MVIP). Piscataway, NJ: IEEE, 2020: 1−6

[16] Zi Bojia, Zhao Shihao, Ma Xingjun, et al. Revisiting adversarial robustness distillation: Robust soft labels make student better[C]// Proc of the 16th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 16443−16452

[17] Maroto J, Ortiz-Jiménez G, Frossard G. On the benefits of knowledge distillation for adversarial robustness[J]. arXiv preprint, arXiv: 2203.07159, 2022

[18] Zhao Shiji, Duan Ranjie, Wang Xizhe, et al. Improving adversarial robust fairness via anti-bias soft label distillation[C/OL]//Proc of the 38th Int Conf on Neural Information Processing Systems. 2024[2025-09-06]. https://papers.nips.cc/paper_files/paper/2024/file/a2370db7c99791ad5d9f3ef48ad6d464-Paper-Conference.pdf

[19] Zhao Shiji, Yu Jie, Sun Zhenlong, et al. Enhanced accuracy and robustness via multi-teacher adversarial distillation[C/OL]//Proc of the 14th European Conf on Computer Vision. 2022[2025-09-06]. https://lin k.springer.com/chapter/10.1007/978-3-031-19772-7_34

[20] Zhu Jianing, Yao Jiangchao, Han Bom, et al. Reliable adversarial distillation with unreliable teachers[C/OL]//Proc of the 10th Int Conf on Learning Representations. 2022[2025-09-06]. https://openreview.net/ forum?id=u6TRGdzhfip&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DICLR.cc%2F2022%2FConference%2FAuthors%23your-submissions)

[21] Huang Bo, Chen Mingyang, Wang Yi, et al. Boosting accuracy and robustness of student models via adaptive adversarial distillation[C]// Proc of the 40th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE 2023: 24668−24677

[22] Park Hyejin, Min Dongbo. Dynamic guidance adversarial distillation with enhanced teacher knowledge[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2025: 204−219

[23] Erven V, Harremos T. Rényi divergence and Kullback-Leibler divergence[J]. IEEE Transactions on Information Theory, 2014(7): 3797−3820

[24] Guo Chuan, Gardner Jacob, You Yurong, et al. Simple black-box adversarial attacks[C]//Proc of the 12th Int Conf on Machine Learning. New York: PMLR, 2019: 2484-2493

[25] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C/OL]//Proc of the 3rd In Conf on Learning Representations. 2015[2025-09-06]. https://www.researchgate.net/publication/269935591_Explaining_and_Harnessing_Adversarial_Examples

[26] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C/OL]//Proc of the 6th In Conf on Learning Representations. 2018[2025-09-06]. https://www.padl.ws/papers/Paper%2019.pdf

[27] Carlini L, Wagner D. Towards evaluating the robustness of neural networks[C]//Proc of the IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39−57

[28] Croce F, Hein F. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C/OL]//Proc of the 13th Int Conf on Machine Learning. New York: PMLR, 2020: 2206−2216

[29] Sui Chenhong, Wang Ao, Zhou Shengwen, et al. A survey on adversarial training for robust learning[J]. Journal of Image and Graphics, 2023, 28(12): 3629−3650 (in Chinese)
(隋晨红, 王奥, 周圣文, 等. 面向鲁棒学习的对抗训练技术综述[J]. 中国图象图形学报, 2023, 28(12): 3629−3650)

[30] Chen Sihong, Shen Haojing, Wang Ran, et al. Relationship between prediction uncertainty and adversarial robustness[J]. Journal of Software, 2022, 33(2): 524−538 (in Chinese)
(陈思宏, 沈浩靖, 王冉, 等. 预测不确定性与对抗鲁棒性的关系研究[J]. 软件学报, 2022, 33(2): 524−538)

[31] Zhao Shiji, Wang Xizhe, Wei Xingxing. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 1−14

[32] Yin Shenglin, Xiao Zhen, Song Mingxuan, et al. Adversarial distillation based on slack matching and attribution region alignment[C]//Proc of the 41st IEEE/CVF Conf on Computer Vision

and Pattern Recognition. Piscataway, NJ: IEEE 2024: 24605−24614

[33] Wu Yulun, Lao Mingrui, Guo Yanming, et al. Boosting adversarial robustness distillation via hybrid decomposed knowledge[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 4715−471

[34] Long Teng, Gao Qi, Xu Lili, et al. A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions[J]. Computers & Security, 2022, 121: 102847

[35] Krizhevsky A, Hinton G. Learning multiple layers of features from Tiny Images[J/OL]. Handbook of Systemic Autoimmune Diseases. 2009[2025-09-06]. https://www.semanticscholar.org/paper/ Learning-Multiple- Lay ers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62 413346235c02b1aa086

[36] Benbrahim H, Behloul A. Fine-tuned Xception for image classification on Tiny ImageNet[C/OL]//Proc of the 1st Int Conf on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP). 2021 [2025-09-06]. https://www.researchgate.net/publication/3577550 47_Fine-tuned_Xception_for_Image_Classification_on_Tiny_ImageNet

[37] Zhang Hongyang, Yu Yaodong, Jiao Jiantao, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proc of the 12th Int Conf on Machine Learning. New York: PMLR, 2019: 7472−7482

[38] Croce F, Andriushchenko M, Sehwag V, et al. RobustBench: A standardized adversarial robustness benchmark[J]. arXiv preprint, arXiv: 2010.09670, 2020



Lao Mingrui, born in 1995. PhD, associate professor. His main research interests include trustworthy AI and multimodal learning.

老明瑞, 1995 年生。博士，副教授。主要研究方向为可信人工智能、多模态学习。



Li Zheng, born in 2000. PhD candidate. His main research interests include adversarial training and multimodal learning.

李　正, 2000 年生。博士研究生。主要研究方向为对抗训练、多模态学习。



Ding Zhaoyun, born in 1982. PhD, professor. His main research interests include data mining and trustworthy AI.

丁兆云, 1982 年生。博士，教授。 主要研究方向为数据挖掘、可信人工智能。



Zhang Xueyi, born in 1998. PhD candidate. His main research interests include adversarial training and continual learning.

张雪毅, 1998 年生。博士研究生。 主要研究方向为对抗训练、持续学习。



Yue Xianghu, born in 1994. PhD, associate professor. His main research interests include multimodal learning and trustworthy AI.

岳祥虎, 1994 年生。 博士，副教授。主要研究方向为多模态学习、可信人工智能。



Guo Yanming, born in 1989. PhD, associate professor. His main research interests include multimodal learning and trustworthy AI.

郭延明, 1989 年生。 博士，副教授。 主要研究方向为多模态学习、可信人工智能。