基于对抗混合专家后训练机制的鲁棒AI生成图像检测方法

张睿萱[1],刁云峰[1*],陆智远[1],夏海峰[2],郭治卿[3],郝孝帅[4],汪 萌[1]
(1. 合肥工业大学计算机与信息学院,安徽合肥 230601;2. 中山大学网络空间安全学院,广东深圳 518107;
3. 新疆大学计算机科学与技术学院,新疆乌鲁木齐 830017;4. 小米汽车,北京 100085)

摘 要: AI生成图像技术(AI-Generated Images,AIGI)技术实现了高质量视觉内容的自动化生产,在艺术创作、数字娱乐及虚拟现实等领域展现出巨大的应用潜力。然而,该技术在赋能内容生产的同时,也带来了严峻的安全与伦理挑战。生成模型可能被恶意用于伪造真实人物或事件,进而制造虚假信息、传播深度伪造内容,甚至干扰网络舆论。因此,如何有效识别AI生成图像(AIGI检测),已成为保障数字内容可信性和维护网络空间安全的重要研究课题。然而,现有的AIGI检测器在面对对抗攻击时普遍表现出鲁棒性不足的问题,攻击者仅需向合成图像中添加人眼难以察觉的细微对抗扰动,即可使其绕过检测,导致合成内容被误判为真实图像,且对于此类攻击的防御机制仍鲜有研究。针对该问题,本文首先系统评估了对抗训练在AIGI检测任务上的有效性。理论分析与实验结果表明,其在训练过程中易诱发特征纠缠现象,进而导致检测性能严重退化甚至崩塌。鉴于此,亟需发展一种针对AIGI检测任务有效的专用对抗防御方法。与对抗训练中出现的特征纠缠不同,本文发现在标准训练的检测器中,对抗扰动会致使对抗样本在特征空间中的表示明显偏离于干净样本,从而形成显著的可分离性。基于该观察,本文提出将对抗样本视作独立类别进行建模的策略,并构建了一种后训练防御框架:在保持预训练特征提取器固定的前提下,仅通过学习新的分类边界以拟合对抗样本的特征分布。为增强模型对未知攻击的泛化能力,本文进一步提出一种对抗混合专家后训练机制。该机制利用多个专家模块分别学习特定攻击类型的特征模式,并引入共享专家以捕捉不同攻击间的共性表征,从而实现对多类对抗样本的高效建模与鲁棒识别。实验结果表明,本文方法在ProGAN和Stable Diffusion等主流AIGI数据集上,面对多种典型对抗攻击方式,在不牺牲良性样本检测精度的前提下,其平均对抗准确率相较现有主流防御方法分别提升了18.92%与12.56%,展现出良好的实用性与在实际安全场景中的应用潜力。

关键词: AI生成图像检测;对抗样本;对抗攻击;对抗防御;混合专家;后训练策略

基金项目: 国家自然科学基金(No.62302139,No.62406068,No.62302427);中央高校基本科研业务费专项资金项目(No.JZ2025HGTB0227)

中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2026)03-1178-16

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251196

Adversarial Mixture of Experts Post-Training for Robust AI-Generated Image Detection

ZHANG Ruixuan[1], DIAO Yunfeng[1*], LU Zhiyuan[1], XIA Haifeng[2],
GUO Zhiqing[3], HAO Xiaoshuai[4], WANG Meng[1]
(1. *Department of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui* 230601, *China*;
2. *School of Cybersecurity, Sun Yat-sen University, Shenzhen, Guangdong* 518107, *China*;
3. *Department of Computer Science and Technology, Xinjiang University, Urumqi, Xinjiang* 830017, *China*;
4. *Xiaomi EV, Beijing* 100085, *China*)

Abstract: AI-generated imagery (AIGI) technology has enabled the automated production of high-quality visual content, demonstrating enormous application potential in fields such as artistic creation, digital entertainment, and virtual reality. However, while empowering content production, this technology also brings serious security and ethical challenges. Generative models can be maliciously used to forge real people or events, thereby creating false information, spreading deepfake content, and even interfering with online public opinion. Therefore, how to effectively identify AI-generated images (AIGI detection) has become an important research topic for ensuring the credibility of digital content and maintaining cyberspace security. However, existing AIGI detectors generally exhibit insufficient robustness against adversarial attacks. At-

收稿日期:2026-02-27;录用日期:2026-03-12;责任编委:彭勃
*通信作者:刁云峰

tackers only need to add subtle adversarial perturbations imperceptible to the human eye to the synthesized image to bypass detection, causing the synthesized content to be misclassified as a real image, and defense mechanisms against such attacks are still scarce. To address this issue, this paper first systematically evaluates the effectiveness of adversarial training in AIGI detection tasks. Theoretical analysis and experimental results show that it is prone to inducing feature entanglement during training, leading to severe degradation or even collapse of detection performance. Therefore, there is an urgent need to develop a dedicated adversarial defense method effective for AIGI detection tasks. Unlike feature entanglement that occurs in adversarial training, this paper finds that adversarial perturbations in standard-trained detectors cause adversarial examples to deviate significantly from clean examples in the feature space, resulting in significant separability. Based on this observation, this paper proposes a strategy of modeling adversarial examples as independent categories and constructs a post-training defense framework: while keeping the pre-trained feature extractor fixed, it only learns new classification boundaries to fit the feature distribution of adversarial examples. To enhance the model's generalization ability to unknown attacks, this paper further proposes an adversarial hybrid expert post-training mechanism. This mechanism utilizes multiple expert modules to learn feature patterns for specific attack types and introduces shared experts to capture common representations among different attacks, thereby achieving efficient modeling and robust identification of multiple classes of adversarial examples. Experimental results show that on mainstream AIGI datasets such as ProGAN and Stable Diffusion, facing various typical adversarial attack methods, the average adversarial accuracy is improved by 18.92% and 12.56% respectively compared to existing mainstream defense methods without sacrificing the detection accuracy of benign examples, demonstrating good practicality and application potential in real-world security scenarios.

Keywords: AI-generated image detection; adversarial example; adversarial attacks; adversarial defense; mixture of experts; post-training strategy

Foundation Item(s): National Natural Science Foundation of China (No.62302139, No.62406068, No.62302427); Fundamental Research Funds for the Central Universities (No.JZ2025HGTB0227)

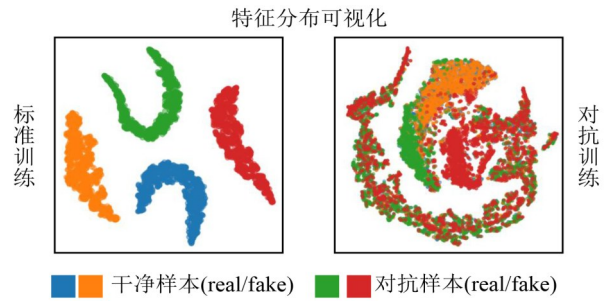
0 引言

随着深度学习技术的发展,AI生成图像技术(AI-Generated Images,AIGI)取得了突破性进展,能够生成与真实照片几乎难以区分的高质量图像。这些技术在艺术创作、虚拟现实等领域展现出显著的应用潜力,但与此同时也伴生了多重安全隐患。不法分子可能利用相关技术生成深度伪造内容,用于虚假新闻传播或舆论操控,从而对社会信任体系、舆论安全与监管机制构成严峻挑战。近年来,多家新闻媒体的调查报道不断揭示深度伪造在商业诈骗、虚假信息生产等场景中的滥用趋势。据国家反诈中心统计,2025年AI诈骗案件同比增长超1 900%[1],凸显其风险外溢的速度与规模。由此可见,AIGI技术滥用所引发的安全风险已突破技术范畴,正在演变为威胁国家安全和公民合法权益的复合型挑战。

为应对上述风险,AI合成图像检测(AIGI检测)已成为当前重要的研究方向。国家网信办等部门联合发布的《生成式人工智能服务管理暂行办法》[2]明确提出,应提升对生成式AI中虚假有害信息的识别与抵御能力,确保技术应用不危害国家安全、公共利益及他人合法权益。然而,已有研究表明,即便当前最先进的AIGI检测器,在面对对抗攻击时仍存在鲁棒性不足的问题[3],攻击者仅需向合成图像中添加人眼难以察觉的细微对抗扰动,即可使其绕过检测,导致合成内容被误判为真实图像。AIGI检测器的对抗脆弱性削弱了现有检测器的防护效能,为恶意生成内容规避监管提供了可行路径,因此,如何构建面向复杂对抗场景的鲁棒检测机制,已成为AIGI检测研究中亟待解决的重要难题。

尽管对抗攻击的威胁日益凸显,相关防御研究仍处于初步阶段。由于对抗训练是最有效且广泛采用的防御方法之一,本文首先尝试应用对抗训练来提高模型鲁棒性。然而,这在AIGI检测任务中却导致性能崩溃,即损失函数无法收敛,检测器丧失判别能力。为调查模型崩塌的原因,本文对标准训练和对抗训练场景下的样本进行特征可视化分析,如图1所示。对模型输出的置信度进行分析,如图2所示。从图1和图2可以观察到对抗训练导致的两种退化现象。

(1)特征纠缠:干净样本与对抗样本的隐藏层特征高度混杂,如图1(b)所示。(2)置信度可分性降低:各类别输出置信度得分的区分度显著下降,如图2(b)所示。这些现象从信息论视角可得出统一解释:它们反映了特征与目标标签之间的互信息显著下降。为深入理解这一退化机制,本文从信息论的角度将对抗训练的优化目标分解为干净样本与对抗扰动的互信息组件,并通过互信息追踪发现了明显的内部互信息权衡现象:最大化对抗样本的互信息(提高鲁棒性)往往以干净样本互信息急剧下降(牺牲良性准确率)为代价(如图3所示),这揭示了对抗训练AIGI检测任

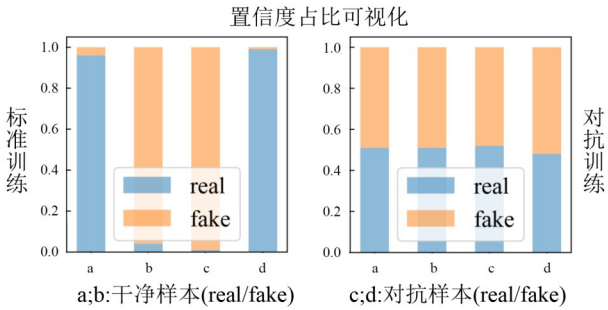


(a) 标准训练特征可视化 (b) 对抗训练特征可视化

(a) Standard feature visualization (b)Adversarial feature visualization

图1 标准训练和对抗训练场景下干净样本与对抗样本的特征t-SNE可视化

Figure 1 t-SNE visualization of clean versus adversarial samples under standard and adversarial training



(a) 标准训练置信度 (b) 对抗训练置信度

(a) Standard confidence (b) Adversarial confidence

图2 标准训练和对抗训练场景下干净样本与对抗样本的输出置信度分布

Figure 2 Confidence score distributions of clean versus adversarial samples under standard and adversarial training

务中的根本性局限。

尽管在AIGI检测任务中,对抗训练模型的特征分布高度纠缠,但本文观察到,在标准训练(训练过程中未使用对抗样本)条件下,对抗扰动导致对抗样本的隐藏特征相对于干净样本发生显著偏移,使二者在特征空间中表现出显著的分离(如图1(a)所示)。基于上述现象,对抗样本在AIGI检测任务中可以视为一种独立类别。由于在特征空间中,真实样本、伪造样本及其对应的对抗样本之间呈现明显的可分性,本文提出一种后训练方法,在固定标准预训练检测器的特征提取部分的基础上,仅学习新的决策边界,以精细刻画不同类型样本之间的区分结构。与重新训练整个模型相比,该方法避免了对大规模参数的更新,因而计算开销显著降低,同时还能在保持原有检测器对干净样本判别能力的前提下,有效建模对抗样本的特征分布,从而进一步增强检测器应对对抗扰动的鲁棒性。

虽然基于后训练的决策边界调整可以提升模型对训练过程中使用的对抗样本的鲁棒性,但在面对未见过的攻击类型时,其鲁棒性提升依然有限。为有效建模多种对抗样本,本文提出一种对抗混合专家后训练方法,通过采样多种不同类型的对抗样本,利用混合专家(Mixture of Experts,MoE)[4]结构中"专家-门控机制"的动态选择能力,使各个专家能够分别捕获不同对抗模式下的特征偏移,从而实现对多样化对抗样本的灵活适应。具体而言,本文首先采样多种典型的对抗攻击方法生成对抗样本,将这些样本用于对抗混合专家后训练:各独立专家负责学习特定攻击类型的伪影特征,而共享专家则专注于捕捉不同攻击间的共性特征和全局模式。实验结果表明,这种灵活的协作机制使得模型能够根据输入数据的特征,动态激活最合适的专家组合。最终实现了对多类对抗样本的统一且鲁棒的判别,从而显著增强了模型对未知对抗扰动的泛化识别能力。

总体来看,本文的主要工作与创新点如下:

(1)本文在信息论视角下,从理论分析和实验验证两个维度深入阐明了对抗训练在AIGI检测中导致性能崩溃的深层机制,揭示了对抗训练在AIGI检测任务中存在的根本性局限,为后续防御方法提出了潜在的改进方向。

(2)本文提出了一种基于对抗混合专家后训练机制的鲁棒AIGI检测方法。在固定预训练检测器特征提取部分的基础上,仅学习新的决策边界,通过采样多种典型对抗样本并利用"专家-门控机制",提升鲁棒泛化能力。

(3)大量实验结果表明,本文所提出的方法在不牺牲良性样本检测精度的前提下,面对多种对抗攻击场景均展现出显著的有效性,验证了其在提升AIGI检测器鲁棒性、增强对抗样本识别能力方面的优越性能。

1 相关工作

1.1 AI生成图像检测技术

早期的AIGI检测方法通常基于卷积神经网络架构,这类方法利用生成模型在上采样和特征重建过程中往往会产生特有的高频伪影或纹理不连续性,例如空间域和频域的周期性伪影[5]、纹理分布异常[6]以及相邻像素异常分布[7]等。另一类方法基于重建误差原理,研究者利用自编码器或扩散重建模型,将输入图像还原至潜在空间后再重构原图,通过比较输入与重建图像之间的差异来判断其真实性[8-9]。随着视觉语言模型的发展,基于CLIP(Contrastive Language-Image Pre-training)的AIGI检测器逐渐成为新的研究趋势。此类检测器利用图像-文本对齐的语义表示能

力,从语义层面对图像内容进行一致性分析,不仅关注像素层面的伪影,更能捕捉AI生成图像中存在的语义细微差别与概念组合异常[10-11]。总体来看,现有AIGI检测器在多种生成模型上取得了较高的准确率,并展现出逐步提升的跨域泛化能力。

1.2　针对AIGI检测器的对抗攻击

De等人[12]系统评估了基于CNN(Convolutional Neural Network)和CLIP的AIGI检测器在多种对抗攻击(如PGD[13]和UA[14])下的鲁棒性,结果表明两类模型在白盒攻击下均表现出明显脆弱性。Mavali等人[15]进一步指出,即使攻击者无法访问目标模型,且对抗样本经过压缩等后处理操作,仍能够在真实场景中有效攻击当前最先进的检测器,显著削弱其检测能力并增加误判风险。在伪造内容检测领域,一些针对AIGI检测任务的对抗攻击方法通过攻击检测器对频域特征的依赖性,对AIGI检测器构成了重大挑战。Dong等人[16]从频谱对齐的角度出发,通过调整伪造图像在频域中的功率分布,使其与真实图像的频谱特征更加一致,从而削弱检测器对高频伪影的敏感性;Hou等人[17]则基于统计量对齐思想,通过匹配图像在像素空间或特征空间中的统计分布来隐藏生成伪迹。近年来,也有研究将频域与空间域的联合优化引入对抗样本生成过程中,利用多尺度特征重构与梯度约束策略设计出扰动不可见、强度较高的对抗噪声[18-19],使添加噪声后的AI合成图像能够有效躲避检测器识别。这些研究充分揭示了AIGI检测模型在对抗样本面前的脆弱性,也为后续鲁棒性分析提供了重要技术参考。

1.3　对抗防御技术

为应对神经网络对抗样本的脆弱性,研究者们提出了各种对抗防御方法[20-22]。其中,对抗训练逐渐成为最具有效性的防御策略之一。PGD-AT[13]采用基于PGD的最小-最大优化框架,通过在训练过程中加入对抗样本显著提升模型在白盒攻击下的鲁棒性,但通常会导致干净样本准确率明显下降。为兼顾鲁棒性与良性准确率,TRADES[23]将对抗风险分解为自然风险与预测分布间的KL散度,从而在两者之间取得平衡。RAT[24]通过在参数中注入随机噪声使模型更易收敛到更“平坦”的最优解。LAS-AT[25]则通过在对抗训练使用可学习的攻击策略来提升对抗鲁棒性。然而,Diao等人[26]发现,现有的对抗训练方法在AIGI检测任务中会发生表现崩塌现象,与此同时,当前针对AIGI检测的防御方法仍然缺失。为解决这个问题,本文从对抗样本与干净样本的特征分布规律出发,提出一种基于对抗混合专家后训练的防御方法,以提升检测器在对抗场景下的鲁棒性。

近年来,MoE因其独特的架构优势在鲁棒学习领域展现出巨大潜力。一方面,MoE架构在对抗鲁棒性方面具有天然优势。研究表明,利用稀疏激活机制和多专家并行结构,模型可以在结构层面上有效缓解对抗扰动对整体性能的影响[27-28]。另一方面,针对MoE架构的防御增强策略也逐渐得到关注。通过引入特定优化机制,可以进一步提升MoE模型在复杂对抗环境下的预测稳定性[29-30]。

2　AIGI检测场景下的对抗训练分析

由于对抗训练被广泛认为是对抗攻击最有效的防御方法之一,本文首先考察了对抗训练在AIGI检测任务中的表现。实验结果(如表1和表2所示)表明,无论采用何种对抗训练方法,其在AIGI数据集上均出现了性能崩溃的现象,表现为训练过程不收敛,模型的准确率和鲁棒性停滞在50%左右,甚至低于50%,这一现象在多种检测器上均有体现,打破了对抗训练能够提升鲁棒性的认知。尽管已有研究注意到AIGI检测中对抗训练存在性能崩溃[26],但根本原因尚未得到分析。

为探究其根本原因,本文首先可视化了标准训练和对抗训练场景下干净样本与对抗样本的潜在特征分布,并分析了检测器在各类别上的输出置信度。图1(b)显示,在对抗训练的检测器中,干净样本与对抗样本的隐藏层特征高度纠缠。同时,图2(b)表示二者在输出端被赋予几乎相同的置信度。从信息论的角度来看,这表示模型学习到的特征表示与目标标签之间的互信息下降,削弱了模型区分真伪样本的能力。Qin等人[31]从信息论的角度指出,有监督分类任务的优化目标可以表述为最大化特征表示与目标标签的互信息:

$$\max_{\theta} I(z; y) \tag{1}$$

其中,$I(\cdot;\cdot)$表示两个变量之间的互信息;$z \in \mathbb{R}^d$表示输入图像x的压缩特征;y是其对应的标签。$I(z; y)$量化了z中保留了多少关于y的信息,$I(z; y)$可以分解为$H(y)-H(y|z)$,其中$H(y)$描述了数据集中y的分布,$H(y|z)$描述了在给定z的条件下预测y的不确定性。受此启发,对抗训练的优化目标可以重新解释为

$$\max_{\theta} \min_{\tilde{x}} I(\tilde{z}; y) \tag{2}$$

其中,$\tilde{z}$为对抗样本$\tilde{x}$的特征,通过分析对抗训练过程中互信息的组成部分与变化规律,本文提供命题1来解耦$I(\tilde{z}; y)$,并阐明四个变量之间的互信息转换关系。

命题1

设z、$\tilde{z}$、Δz、y表示四个随机变量,其中z、$\tilde{z}$、Δz存在关系$\tilde{z}=z+\Delta z$。以下关系成立:

$$I(\tilde{z};y)\approx I(z;y)+I(\Delta z;y|\tilde{z}) \tag{3}$$

在对抗环境中，Δz 表示对抗样本特征相对干净样本特征的偏移。相应的证明在附录中。命题1体现出 $I(\tilde{z};y)$ 可以分解为两个相互竞争的目标：(1) $I(z;y)$，其负责衡量模型在干净输入上的预测性能；(2) $I(\Delta z;y|\tilde{z})$，其负责捕获对抗扰动信息与给定其原始标签的特征之间的条件依赖性。

通过追踪和对比上述两个优化目标在对抗训练期间的动态变化（如图3所示），本文观察到关键现象：在对抗训练过程中，$I(\Delta z;y|\tilde{z})$ 始终保持在较低的水平，这表明对抗扰动在对抗训练中被视为与 y 无关的随机噪声。此外，不同于语义分类数据集（猫狗数据集），在AIGI数据集（ProGAN）上，$I(z;y)$ 并没有明显的上升趋势，而是保持在接近于0的最小值附近。这一现象表明，在AIGI检测任务中，对抗训练限制了干净样本特征与真实标签之间的互信息增益，从而削弱了模型的判别能力，最终导致了特征纠缠和性能崩塌。该发现突显了开发替代对抗训练的AIGI检测新防御策略的迫切性。

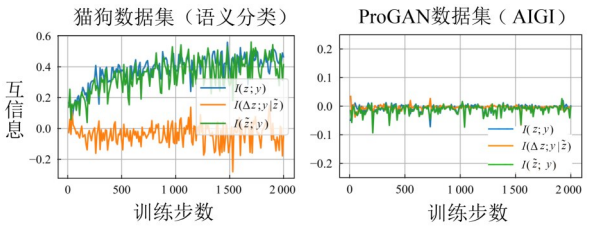


图3 在语义分类数据集(猫狗数据集)与AIGI数据集(ProGAN)上执行对抗训练时，互信息 $I(\tilde{z};y)$、$I(z;y)$、$I(\Delta z;y|\tilde{z})$ 随训练步数的变化趋势

Figure 3 Trend of mutual information $I(\tilde{z};y)$, $I(z;y)$, $I(\Delta z;y|\tilde{z})$ across training steps during adversarial training on semantic classification (Cats vs. Dogs) and AIGI (ProGAN) dataset

上述分析从信息论视角将对抗样本拆解为对抗扰动与干净样本，并观察二者在对抗训练过程中的作用，为探究对抗训练的内在机制提供了一个独特视角。此外，为了直观阐明训练崩塌的根源，本文对内部损失最大化（互信息最小化）与外部经验风险最小化（互信息最大化）的耦合过程进行了分析，首先将对抗扰动替换为等幅度的随机高斯噪声进行对比实验。

由于对抗训练的目标可以表示为 $\max_{\theta}\min_{\tilde{x}} I(\tilde{z};y)$，因此本文用 $I(\tilde{z};y)$ 作为评估指标，将对抗扰动替换为等幅度的随机高斯噪声，使训练退化为常规的噪声鲁棒性训练。图4表明，在不同幅度的高斯噪声下，$I(\tilde{z};y)$ 能够正常上升，且模型收敛稳定；而在不同幅度的对抗噪声下，$I(\tilde{z};y)$ 始终维持在0附近。这表明内部损失最大化加噪方式阻碍了 $I(\tilde{z};y)$ 的正常上升。

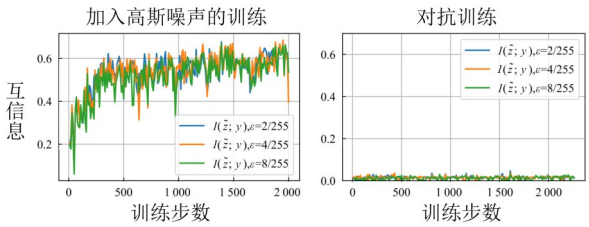


图4 加入高斯噪声的训练与对抗训练过程中互信息 $I(\tilde{z};y)$ 随训练步数的变化

Figure 4 Mutual information $I(\tilde{z};y)$ trends during training with Gaussian versus adversarial noise.

3 基于对抗混合专家后训练机制的鲁棒AI生成图像检测方法

尽管在AIGI检测任务中，对抗训练模型的特征分布高度纠缠，但本文观察到，在标准训练（未使用对抗样本）条件下，对抗扰动导致对抗样本的隐藏特征相对于干净样本发生显著偏移，即干净样本及其对应的对抗样本在特征空间中表现出显著的分离。受此启发，本文提出了一种基于对抗专家的后训练机制，旨在保证干净样本高精度的前提下，有效抵抗多源、多强度对抗扰动的攻击，算法整体流程如图5所示。模型由冻结特征提取器 F、后训练对抗混合专家网络 H、动态对抗生成器 $\mathcal{G}$ 组成，模型分为两个主要阶段：动态样本注入与鲁棒性边界优化。

在本文提出的基于对抗专家的后训练方法中，模型训练分为两个主要阶段。

阶段1：动态对抗样本注入与标签扩展。首先，将AIGI检测任务的标签空间从二分类扩展至四细粒度分类 $\tilde{\mathcal{Y}}$。预训练AIGI检测器特征提取器 F 的参数 ϕ 固定不变，通过动态对抗生成器 $\mathcal{G}$ 对攻击算法 α 和超参数 h 进行双层随机采样，在线生成对抗样本 x_{adv}，并扩展标注 $\tilde{\mathcal{Y}}$。该阶段为模型学习提供了高变异性的对抗训练数据，从而捕获多样化的对抗特征。

阶段2：基于对抗专家后训练的决策边界调整。在第二阶段，仍然固定特征提取器 F 的参数 ϕ 不变，更新对抗混合专家网络 H 的参数 θ，在四分类交叉熵损失 $\mathcal{L}_{\text{cls}}$ 的基础上，引入Logit一致性正则化损失 $\mathcal{L}_{\text{reg}}$，约束干净样本与对抗样本的Logit差异，确保决策边界平滑并减少对特定对抗扰动的过拟合。阶段内，动态样本生成与MoE优化交替循环进行：一方面，$\mathcal{G}$ 利用当前 H_θ 的状态反馈，得到梯度等信息，生成针对性更强的对抗样本；另一方面，基于动态对抗样本注入提供的高多样性数据流不断更新 H_θ 的决策边界，从

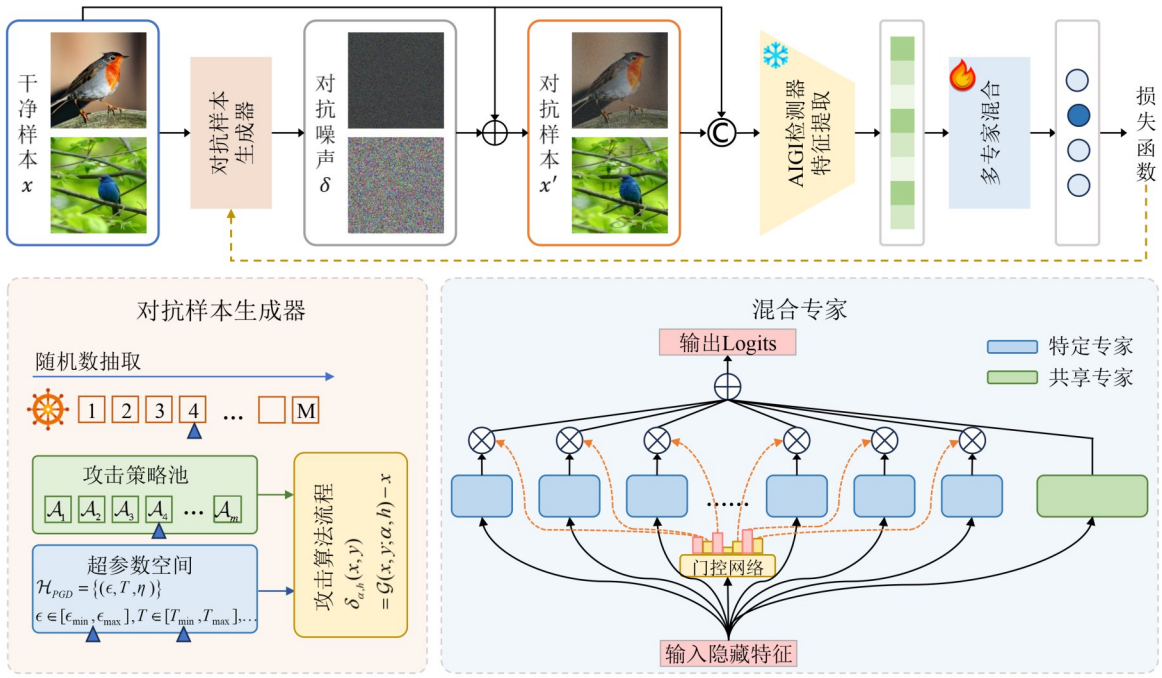


图5　基于对抗混合专家后训练机制的对抗鲁棒AI生成图像检测方法整体框架

Figure 5　The proposed framework for adversarial robust AIGI detection using an adversarial mixture-of-experts post-training mechanism

而实现对多源对抗攻击的鲁棒泛化。

3.1　问题设定

考虑一个K分类的任务。给定数据分布$\mathcal{D}$,其中输入空间$\mathcal{X}\subseteq\mathbb{R}^d$,标签空间$\mathcal{Y}=\{1,2,\cdots,K\}$。定义一个分类器$f_\Phi:\mathcal{X}\to\mathbb{R}^K$,其中$\Phi$为可学习的参数。对于输入$x$,模型输出Logits向量或概率分布,其对应的离散预测类别(决策函数)$C_\Phi(x)$定义为

$$C_\Phi(x)=\arg\max_{k\in\mathcal{Y}} f_\Phi(x)_k \tag{4}$$

对抗攻击旨在为给定的干净样本(x,y)寻找一个微小的扰动,使得模型产生错误的预测。该扰动必须受到约束集合$\mathcal{S}$的限制(例如L_p范数球:$\mathcal{S}=\{\delta:\|\delta\|_p\leqslant\epsilon\}$),以保证其不可察觉性。形式上,对抗攻击的目标是寻找满足以下条件的扰动δ:

$$C_\Phi(x+\delta)\neq y,\quad \text{s.t.}\|\delta\|_p\leqslant\epsilon \tag{5}$$

为了生成有效的对抗扰动,攻击者通常求解一个约束优化问题,即在扰动限制内最大化模型损失函数以诱导分类错误。

鲁棒学习的目标是训练出具有防御能力的模型,确保其在面对约束范围内的恶意扰动时,依然能够做出正确的决策,而不仅仅局限于对干净样本的拟合。在数学定义上,这等价于在一个极小极大(Min-Max)博弈框架下,最小化模型在最坏扰动情况下的期望风险:

$$\min_{\Phi}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\delta\|_p\leqslant\epsilon}\mathbb{I}\left(C_\Phi(x+\delta)\neq y\right)\right] \tag{6}$$

相应地,为了确保模型的训练过程可解且可微,研究者将内部的离散0-1风险松弛为连续可导的损失函数,即经典的对抗训练范式:

$$\min_{\Phi}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\delta\|_p\leqslant\epsilon}\mathcal{L}\left(f_\Phi(x+\delta),y\right)\right] \tag{7}$$

其中,内部最大化寻找最强攻击样本,外部最小化则更新模型参数以降低对抗样本带来的损失。

与标准对抗训练重新训练模型的所有参数不同,本文提出的方法冻结了预训练AIGI检测器特征提取部分F的参数ϕ,仅优化后训练对抗专家网络H的参数θ。整体优化目标定义为

$$\begin{aligned}\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}&\left[\max_{\|\delta\|_p\leqslant\epsilon}\mathcal{L}\left(H_\theta\left(F_\phi(x+\delta)\right),\tilde{y}\right)\right]\\&+\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}\left(H_\theta\left(F_\phi(x)\right),y\right)\right]\end{aligned} \tag{8}$$

其中,$\tilde{y}$是为对抗样本扩展的新标签。该博弈过程包含两个交替进行的阶段。内部最大化:寻找能够最大化模型损失的扰动δ,即生成有效的对抗样本。外部最小化:更新参数θ,以降低模型在干净和对抗样本上的混合损失。

3.2　模型解耦与特征冻结

设原始的检测器f_θ由特征提取器$F_\phi:\mathcal{X}\to\mathcal{Z}$和二分类头$C_\psi:\mathcal{Z}\to\mathbb{R}^2$组成,其中$\phi$和$\psi$为各自的参数。本方法在后训练阶段冻结特征提取器的参数F_ϕ,令ϕ^*表示预训练好的固定参数,对于任意输入x,其特征表示z为

$$z = F_{\phi^*}(x), \quad \text{where } \phi^* \text{ is fixed} \tag{9}$$

随后，将 C_ψ 替换为待训练的后训练网络 H_θ。新的检测模型由 F_{ϕ^*} 和后训练网络 H_θ 构成，有如下形式：

$$f_{\text{new}}(x) = H_\theta\left(F_{\phi^*}(x)\right) \tag{10}$$

优化过程仅更新参数 θ，而不改变特征空间 $\mathcal{Z}$ 的结构。此外，为了引导 H_θ 显式地学习区分干净样本与对抗样本，将原始的二元标签空间 $\mathcal{Y}=\{\text{real}, \text{fake}\}$ 扩展为四元标签空间 $\tilde{\mathcal{Y}}$。新的类别定义如下：

$$\tilde{y} = \begin{cases} 0, & \text{if } x \text{ is Clean Real} \\ 1, & \text{if } x \text{ is Clean Fake} \\ 2, & \text{if } x \text{ is Adversarial (Real} + \delta) \\ 3, & \text{if } x \text{ is Adversarial (Fake} + \delta) \end{cases} \tag{11}$$

通过这种映射，原本被视为单一类别的“攻击样本”被细分为独立的类别。这迫使模型在特征空间中刻画出更精细的决策边界。

3.3 动态对抗样本生成

传统静态对抗样本集由于分布同质性及特征空间覆盖不足，容易导致模型对特定攻击类型产生过拟合。为克服这一局限并提升模型在未知扰动下的泛化鲁棒性，本文设计并采用了动态对抗生成器（Dynamic Adversarial Generator，DAG）。DAG 的核心目标在于替代预先构建的静态数据集，实现训练过程中对对抗样本流形的全面覆盖。通过这种动态注入机制，模型能够持续适应多样化扰动，从而学习到更具判别力的本征特征。

具体而言，本方法构建了一个包含 M 种攻击算法的策略池 $\mathcal{A}=\{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_M\}$。为在攻击强度与分布多样性之间取得平衡，策略池中特意纳入了目标机制存在显著差异的多类攻击方式。

（1）基于梯度的最大损失攻击。以 PGD 为代表，通过最大化替代损失函数来寻找满足范数约束的局部最坏扰动，从而生成高强度对抗样本。

（2）基于决策边界的最小距离攻击。以 FAB 为典型，致力于在决策边界附近搜索最小范数扰动，生成更接近分类边界的精细对抗样本。

除此之外，每种攻击 $\mathcal{A}_k$ 都关联着一个超参数空间 $\mathcal{H}_k$，定义了扰动预算（ϵ）、迭代步数（T）或步长（η）等关键参数范围的集合。例如对于 $\mathcal{A}_{PGD}$，$\mathcal{H}_{PGD}$ 定义为

$$\mathcal{H}_{\text{PGD}} = \left\{(\epsilon, T, \eta) \middle| \epsilon \in \left[\epsilon_{\min}, \epsilon_{\max}\right], \cdots\right\} \tag{12}$$

在每个训练迭代步中，对于给定的干净样本批次 $\mathcal{B}=(x_i, y_i)_{i=1}^{B}$，生成器 $\mathcal{G}$ 执行双层随机采样。

（1）算法采样。根据预设的采样概率分布 π_α，从策略池 $\mathcal{A}$ 中随机选择一种攻击算法 α。

（2）超参数采样。根据所选攻击方法 α 对应的超参数空间 $\mathcal{H}\alpha$ 上的分布 π_h，随机采样一组超参数配置 $h \in \mathcal{H}\alpha$。

随后，生成器 $\mathcal{G}$ 基于选定的算法 α 和超参数配置 h 施加对抗扰动，得到对应的动态对抗样本：

$$x_{\text{adv}} = \mathcal{G}(x, y;\ \alpha, h) = x + \delta_{\alpha,h}(x, y) \tag{13}$$

3.4 基于对抗混合专家的鲁棒建模

为有效提升在未知攻击场景中的对抗鲁棒性，后训练网络 H_θ 被设计为混合专家（MoE）架构。通过将复杂的分类任务分解为多个子任务，并利用“专家-门控机制”的动态选择能力，使各个专家能够分别捕获不同对抗模式下的特征偏移，从而实现对多样化对抗样本的灵活适应。

以冻结骨干网络所提取的特征 $z \in \mathbb{R}^d$ 为输入，H_θ 的目标是针对扩展的四分类标签空间 $\tilde{\mathcal{Y}}$ 的预测 Logits。H_θ 由以下三个核心组件构成。

（1）共享专家（Shared Expert，E_s）。一个独立且持续激活的模块，负责提取所有样本的通用特征及全局模式，保障模型的基础判别能力。

（2）特定专家组（Specific Experts，$\{E_i\}_{i=1}^{N}$）。该专家组由 N 个独立的神经网络模块组成。各专家通过协同学习分别捕获异构特征分布中的细微差异，从而能够拟合特定攻击类型的判别性特征。

（3）门控网络（Gating Network，G）。一个轻量级的决策模块，根据输入特征 z 动态地计算每个特定专家的激活权重。

模型的最终输出 $H_\theta(z)$ 采用残差连接的形式，由共享专家的输出与加权后的特定专家输出叠加而成。其数学表达如下：

$$H_\theta(z) = E_s(z) + \sum_{i=1}^{N} G(z)_i \cdot E_i(z) \tag{14}$$

共享专家 $E_s(z)$ 和特定专家组 $E_i(z)$ 均由多层感知机（MultiLayer Perceptron，MLP）实现，其输出维度与分类数目（即 4 类）一致。该设计确保即便门控网络在处理某些困难样本时判断不够精确，共享专家仍能够提供稳健的特征基底，从而增强整体训练过程的稳定性。门控网络 $G: \mathbb{R}^d \to \mathbb{R}^N$ 充当“路由器”的角色，它根据输入特征 z 来确定各特定专家的参与权重。为此，门控模块使用线性层结合 Softmax 函数生成归一化的权重向量：

$$G(z) = \text{Softmax}(W_g z + b_g) \tag{15}$$

其中，$W_g \in \mathbb{R}^{N \times d}$ 和 $b_g \in \mathbb{R}^N$ 是门控网络的可学习参数。$G(z)_i$ 表示第 i 个专家处理当前输入样本 x 的权重。

通过联合优化共享专家与经门控选择的特定专家，模型能够在“全局通用模式”与“攻击特定伪影”

之间实现自适应权衡,从而更准确地区分标准 干净样本与各类对抗样本。

3.5 训练策略

综上,优化目标可以表示为

$$\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{E}_{(\alpha,h)\sim\mathcal{A}}\left[\mathcal{L}\left(H_\theta\left(F_{\phi^*}(x+\delta_{\alpha,h})\right),\tilde{y}\right)\right]\right]+\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}\left(H_\theta\left(F_{\phi^*}(x),y\right)\right)\right],\delta_{\alpha,h}\sim\mathcal{G}_{\alpha,h}(x) \tag{16}$$

其中,$\tilde{y}$为扩展后的四分类标签;$\mathcal{A}$表示对攻击算法类型α与超参数配置h的采样分布。具体来说,在每个训练迭代步(Step)中,首先从训练集中采样一个干净样本批次$\mathcal{B}=(x_i,y_i)_{i=1}^{B}$,其中$y_i\in 0,1$为原始真伪标签。随后,利用前述动态对抗生成器(DAG),为每个干净样本x_i在线生成对应的对抗样本x_i^{adv}。根据3.1节定义的扩展标签规则,为这两类样本分配四分类标签:干净样本x_i对应标签$\tilde{y}_i=y_i$,对抗样本x_i^{adv}对应标签$\tilde{y}_i^{\text{adv}}=y_i+2$(即映射至类别2或3)。

为了最大化计算效率,将干净样本与对抗样本Batch 维度上拼接,形成联合输入批次:

$$X_{\text{joint}}=[x_1,\cdots,x_B,x_1^{\text{adv}},\cdots,x_B^{\text{adv}}] \tag{17}$$

将X_{joint}输入参数冻结的特征提取器F_{ϕ^*}中,得到联合特征表示Z_{joint},再送入可学习的混合专家网络H_θ得到预测 Logits:

$$\text{Logit}_{\text{joint}}=H_\theta\left(F_{\phi^*}(X_{\text{joint}})\right) \tag{18}$$

其中,$\text{Logit}_{\text{joint}}\in\mathbb{R}^{2B\times 4}$包含了干净样本的Logits$l_i$与对应对抗样本的Logits$l_i^{\text{adv}}$。

为引导模型学习准确的决策边界,同时防止在拟合对抗扰动时出现过拟合现象,本文提出了包含分类损失与一致性正则化的混合目标函数。(1)四分类交叉熵损失。主损失函数采用标准的交叉熵损失,旨在监督 MoE正确区分四种类别:

$$\mathcal{L}_{\text{cls}}=\frac{1}{2B}\sum_{i=1}^{B}\left(\mathcal{L}_{CE}(l_i,\tilde{y}_i)+\mathcal{L}_{CE}(l_i^{\text{adv}},\tilde{y}_i^{\text{adv}})\right) \tag{19}$$

(2)Logit一致性正则化。尽管干净样本与对抗样本被划分为不同类别,但其底层的图像信息高度一致。为抑制模型对对抗扰动的过度响应并防止对特定噪声模式的过拟合,本方法引入了Logit层面的一致性正则化,通过最小化干净样本与对应对抗样本Logits之间的均方误差(Mean Squared Error,MSE),保持模型输出的平滑性:

$$\mathcal{L}_{\text{reg}}=\frac{1}{B}\sum_{i=1}^{B}\left\| l_i-l_i^{\text{adv}}\right\|_2^2 \tag{20}$$

最终的损失函数是上述两项的加权和:

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{cls}}+\lambda\cdot\mathcal{L}_{\text{reg}} \tag{21}$$

其中,λ是平衡系数,用于控制正则化强度。通过最小化$\mathcal{L}_{\text{total}}$,模型不仅能够精确地将对抗样本隔离到独立类别中,还能在面对未知扰动时保持输出稳定,从而在保持良性准确率的基础上提升对抗鲁棒性。整体算法流程如算法1所示。

算法1　基于对抗混合专家后训练机制的鲁棒AI生成图像检测方法

输入:训练数据集$\mathcal{D}_{\text{train}}$;冻结特征提取器$F_{\phi^*}$;混合专家网络$H_\theta$(参数$\theta$);攻击算法池$\mathcal{A}$及超参数空间$\mathcal{H}$;正则化权重$\lambda$

输出:鲁棒混合专家分类网络H_θ

初始化θ随机或使用预训练权重
While 训练未收敛 **do**
　从$\mathcal{D}_{\text{train}}$中采样干净样本批次$\mathcal{B}=\{(x_i,y_i)\}_{i=1}^{B}$
　初始化空的对抗样本批次$\mathcal{B}_{\text{adv}}$
　For $i=1$到B **do**
　　采样攻击算法$\alpha\sim\pi_\alpha(\mathcal{A})$
　　采样超参数配置$h\sim\pi_h(\mathcal{H}_\alpha)$
　　生成对抗样本:$x_i^{\text{adv}}=\mathcal{G}(x_i,y_i;\alpha,h)$
　　$\tilde{y}_i=y_i;\tilde{y}_i^{\text{adv}}=y_i+2$
　　将$(x_i^{\text{adv}},\tilde{y}_i^{\text{adv}})$加入$\mathcal{B}_{\text{adv}}$
　End For
　构建联合输入:$X_{\text{joint}}=\text{Concat}(\mathcal{B},\mathcal{B}_{\text{adv}})$
　$Z_{\text{joint}}=F_{\phi^*}(X_{\text{joint}})$
　获得 Logits:$\text{Logit}_{\text{joint}}=H_\theta(Z_{\text{joint}})$
　分割 Logits:$L_{\text{clean}}=L_{\text{joint}}[0{:}B]$, $L_{\text{adv}}=L_{\text{joint}}[B{:}2B]$
　计算分类损失$(\mathcal{L}_{\text{cls}})$
　计算$\mathcal{L}_{\text{reg}}=\frac{1}{B}\sum_{i=1}^{B}\left\| L_{\text{clean},i}-L_{\text{adv},i}\right\|_2$ (Logit 一致性 MSE)
　$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{cls}}+\lambda\cdot\mathcal{L}_{\text{reg}}$
　$\theta\leftarrow\text{Optimizer}(\mathcal{L}_{\text{total}},\theta)$
End While
Return H_θ

4 实验与分析

4.1 实验设置

数据集介绍。本文使用ProGAN[5]和SDv1.4[32]这两个涵盖GAN和扩散模型生成范式的AIGI数据集评估本文方法。ProGAN数据集包含约3.60×10^5真实图像和约3.60×10^5假图像,其中真实图像选自LSUN[33]数据集的20个不同对象类别,假图像使用20个ProGAN模型各自对应LSUN数据集中不同对象类别训练生成。SDv1.4数据集是GenImage[34]数据集的一个子集,由约1.68×10^5假图像和约1.68×10^5真图像组成。其中假图像由Stable Diffusion V1.4模型生成,生成过程使用“photo of class”形式的文本提示模板,“class”对应ImageNet[35]中的1 000个类别标签,每个类别包含162张训练图像和6张测试图像。真图像来源于ImageNet数据集。

攻击方法介绍。本文采用6种不同的攻击方法

以全面评估本文方法的对抗鲁棒性，针对白盒攻击的PGD[13]、APGD[36]、C&W[37]和FAB[38]，针对黑盒攻击的Square[39]和集成一系列白盒和黑盒攻击的AutoAttack[36]。PGD利用模型梯度信息在多步迭代中最大化损失函数。APGD在PGD的基础上引入自适应步长和动量机制，自动调节超参数以克服梯度掩蔽问题并提高攻击效率。C&W将对抗样本生成转化为无约束优化问题，利用基于Logit的间隔损失函数在最小化扰动距离的同时最大化攻击成功率。FAB使用线性近似将样本迭代投影至决策边界以寻找距离原始样本最近的误分类点，实现最小范数攻击。Square是一种基于分数的黑盒攻击，使用随机搜索策略在图像边界位置选择方形区域进行更新以降低正确类别的置信度。AutoAttack集成APGD、FAB和Square等多种参数无关的攻击策略，通过自适应参数调整提供更可靠的鲁棒性性能评估。

AIGI检测器介绍。为验证所提方法的通用性，本文选取6种基于不同架构的先进AIGI检测器作为防御对象。CNN生成图像检测器（CNN-generated Image Detector，CNNSpot）[5]利用ResNet-50作为骨干网络，通过在训练过程中引入高斯模糊和JPEG压缩等数据增强策略挖掘不同生成模型产生的通用伪影。GramNet[6]通过利用全局图像纹理表示，捕捉伪造面部图像的纹理差异，增强了对图像编辑和不同GAN模型生成伪造图像的鲁棒性和泛化能力。邻域像素关系网络（Neighboring Pixel Relationships，NPR）[7]利用邻域像素间的相关性捕捉生成图像中存在的空间结构异常，提取泛化性伪造特征。通用伪造图像检测器（Universal Fake image Detector，UnivFD）[10]利用预训练视觉语言模型CLIP的冻结特征空间，保留对未知生成模型的泛化能力。中间编码块表示网络（Representations from Intermediate Encoder-blocks，RINE）[40]利用CLIP图像编码器中间层Transformer块所含的低级视觉信息，使用可训练的重要性估计器动态融合不同层级特征以提升检测泛化性。高效正交建模（Efficient orthogonal modeling，Effort）[41]利用奇异值分解将特征空间分解为正交子空间，通过冻结主成分保留预训练知识并仅微调剩余成分以学习伪造模式。

基准方法介绍。为验证本文所提方法优越性，在2个数据集上分别选取基准方法进行比较。投影梯度下降对抗训练（PGD Adversarial Training，PGD-AT）[13]采用极小-极大优化博弈框架，在训练过程中动态注入基于投影梯度下降生成的对抗样本以提高模型的对抗鲁棒性。鲁棒性权衡防御（TRADES）[23]将对抗风险分解为自然分类误差与预测分布间的KL散度正则项，通过优化二者的加权和在自然准确率与对抗鲁棒性之间取得平衡。扩散净化（Diffusion Purification，DiffPure）[42]利用预训练扩散模型的正向随机加噪过程破坏对抗扰动结构，并通过逆向去噪过程重构出干净图像以实现无需重训的测试时防御。

实验参数与细节。本文所提方法使用NVIDIA GeForce RTX 3090进行训练，显存为24.0 GB，Pytorch版本2.1.0，Python版本3.11.7，学习率为0.001，MoE架构采用了8个专家模块，对于不同的基础模型，输入维度有所不同，例如CNNSpot使用2 048，NPR使用512等，隐藏层维度为64，输出维度为4。专家网络由两层全连接层和ReLU激活函数构成，并且所有专家共享一个共享专家模块，用于提取输入的全局特征。门控机制使用两层全连接网络来计算每个专家的权重，最终通过加权求和结合共享专家的输出生成最终结果。

4.2 对抗鲁棒性评估

基于CNN架构的AIGI检测器鲁棒性分析。表1展示了在三类典型CNN架构检测器（CNNSpot、GramNet、NPR）上，不同防御方法在干净样本与多种对抗攻击下的检测性能。可以观察到，其中传统对抗训练方法和对抗净化方法在AIGI检测任务中均出现性能退化或表现崩溃，而本文方法表现出稳定且显著的优势。

首先，Vanilla模型在干净样本上具有接近100%的准确率，但在各类攻击尤其是AutoAttack下几乎完全失效（准确率接近0%）。这表明AIGI检测器对对抗扰动高度敏感。其次，对抗训练方法（PGD-AT，TRADES）在AIGI检测上出现严重的性能崩塌。虽然PGD-AT在某些局部攻击（如PGD）上能维持中等水平的鲁棒性，但其干净样本准确率大幅下降至仅约50%，TRADES的情况更加明显。相比之下，对抗净化方法DiffPure在三种架构上均表现出高度一致但有限的鲁棒性（平均准确率约53%）。其鲁棒性波动较小，但干净样本准确率也仅维持在53%左右。DiffPure在AIGI检测任务中需要对输入图像进行生成重建，而GAN生成模型本身难以保持微小伪造特征（如ProGAN痕迹、频域伪影等），在重建过程中往往会丢失这些关键检测线索，使净化后的图像无法保留判别信息。相比之下，本文提出的方法在三种CNN架构上均取得了最优且稳定的性能。以CNNSpot为例，本文方法在干净样本上达到88.21%，在强攻击APGD和APGD-DLR下仍保持78.92%和77.61%，其在AutoAttack上准确率达到60.89%，平均准确率达到76.49%。类似性能提升在GramNet与NPR上也同样稳定。

基于CLIP架构的AIGI检测器对抗鲁棒性分析。表2展示了近年来广泛采用的CLIP架构（如UnivFD、RINE、Effort）的AIGI检测器在干净样本与多种强攻

表1　基于CNN架构的AIGI检测器在ProGAN上的良性准确率与对抗鲁棒性评估　　单位:%

Table 1　Evaluation of CNN-based AIGI detectors on ProGAN: benign accuracy and adversarial robustness　　unit:%

检测器	方法	Clean	PGD	APGD	APGD-DLR	C&W	FAB	Square	AA	AVG
CNNSpot	Vanilla	99.80	21.05	21.31	21.54	50.29	31.24	0.36	0.21	30.73
	PGD-AT	51.26	49.90	50.58	50.73	50.86	50.60	49.39	40.02	49.17
	TRADES	50.12	6.54	33.03	38.69	50.10	50.12	50.09	6.40	35.64
	DiffPure	53.31	52.74	53.28	52.94	54.70	53.54	55.36	52.18	53.51
	Ours	88.21	**83.47**	**78.92**	**77.61**	**71.88**	**76.94**	**74.02**	**60.89**	**76.49**
GramNet	Vanilla	99.80	21.15	21.10	21.78	50.72	31.55	0.40	0.25	30.84
	PGD-AT	54.64	50.15	49.93	55.51	51.78	51.45	49.15	48.25	51.36
	TRADES	51.32	13.12	30.67	25.59	51.92	50.05	42.47	23.48	36.08
	DiffPure	53.73	54.34	53.98	54.67	53.75	53.62	53.29	52.70	53.76
	Ours	84.67	**82.36**	**80.91**	**69.88**	**66.27**	**82.41**	**68.15**	**58.66**	**74.16**
NPR	Vanilla	99.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.49
	PGD-AT	50.00	50.00	54.89	55.27	51.50	51.50	49.43	48.48	51.38
	TRADES	51.45	16.16	23.45	25.33	1.89	42.60	49.87	23.22	29.25
	DiffPure	53.57	54.17	53.84	54.56	53.58	53.51	53.17	52.62	53.63
	Ours	92.51	**89.35**	**87.47**	**68.65**	**69.43**	**71.67**	**70.70**	**59.39**	**76.15**

注:AA代表AutoAttack,AVG代表平均值。

表2　基于CLIP架构的AIGI检测器在ProGAN上的良性准确率与对抗鲁棒性评估　　单位:%

Table 2　Evaluation of CLIP-based AIGI detectors on ProGAN: benign accuracy and adversarial robustness　　unit:%

检测器	方法	Clean	PGD	APGD	APGD-DLR	C&W	FAB	Square	AA	AVG
UnivFD	Vanilla	99.60	0.00	0.00	0.00	55.34	16.62	0.00	0.00	21.45
	PGD-AT	69.26	0.00	0.18	1.03	41.58	36.15	0.48	0.00	18.59
	TRADES	95.49	0.00	0.00	0.00	72.27	51.32	0.39	0.06	27.44
	DiffPure	62.50	60.71	64.57	62.40	62.49	62.48	61.50	60.30	62.12
	Ours	95.24	**96.63**	**96.92**	**69.84**	**76.88**	**82.54**	**71.08**	**61.34**	**81.31**
RINE	Vanilla	99.01	5.23	1.84	1.92	50.56	25.13	6.38	1.45	23.94
	PGD-AT	72.79	2.68	2.13	2.73	48.84	27.67	42.90	1.92	25.21
	TRADES	88.90	3.97	2.59	3.56	49.72	31.61	45.29	2.52	28.52
	DiffPure	61.45	59.37	62.61	61.08	59.89	60.48	61.85	60.43	60.90
	Ours	96.55	**79.14**	**77.49**	**66.55**	**72.84**	**79.57**	**74.92**	**62.13**	**76.15**
Effort	Vanilla	99.01	0.00	0.00	0.00	50.32	24.89	6.00	0.00	22.53
	PGD-AT	50.68	1.21	1.28	2.09	48.36	27.41	42.63	1.08	21.84
	TRADES	49.90	3.32	2.02	2.87	48.59	26.23	44.99	1.42	22.42
	DiffPure	61.21	58.16	61.85	62.34	59.75	61.18	59.52	60.71	60.59
	Ours	91.16	**76.04**	**66.87**	**72.60**	**71.71**	**75.67**	**79.73**	**64.75**	**74.82**

注:AA代表AutoAttack,AVG代表平均值。

下的性能表现。相比传统CNN架构,基于CLIP的检测器具备更强的跨模态特征表达能力和对不同生成图像的良性泛化能力,代表了该领域的SOTA架构。尽管如此,从表2中仍可观察到若干值得关注的趋势。和CNN检测器相似,Vanilla CLIP检测器在干净样本上通常保持极高的准确率,然而在面对AutoAttack系列攻击时仍表现脆弱,尤其在PGD、APGD、APGD-DLR等基于梯度的攻击下准确率最低跌落至0%,仅在C&W与FAB攻击下保持中等水平。这说明CLIP的决策边界仍易被对抗扰动破坏,在CLIP架构上,PGD-AT与TRADES同样出现了对抗训练崩塌。以UnivFD为例,PGD-AT在PGD/APGD攻击下依旧接近0%,TRADES也未能获得明显提升。而本文所提出的方法依然能够维持与Vanilla接近的高良性准确率,并在各类攻击下超越其他方法。

4.3　泛化能力评估

表3展示了在ProGAN数据集上完成训练,在Stable Diffusion v1.4生成数据集上评估的实验结果,

这意味着表中性能反映的是跨数据分布的泛化能力。可以观察到，与 Vanilla 方法在对抗样本上几乎完全失效（PGD/APGD 攻击下准确率接近 0%）形成鲜明对比，本文方法在 UnivFD、RINE 与 Effort 三个 CLIP 架构检测器上均取得了显著的鲁棒性提升，对抗鲁棒性普遍达到 75% 以上，并远高于基础模型的表现。更为突出的现象是，本文方法在多个检测器上都出现了“对抗鲁棒性超过良性准确率”的结果，例如 UnivFD 和 RINE 在 PGD/APGD 下的准确率明显高于其 Clean 性能，这表明模型已经成功学习到跨模型、跨分布的对抗伪造模式，使得对抗扰动在特征空间中反而呈现出更容易区分的异常结构。由于训练完全在 ProGAN 上进行，而测试分布切换到与其风格差异显著的 Stable Diffusion v1.4，本实验结果充分说明本文方法不仅能够抵御多种强攻击，而且能够在未见过的新型生成数据上保持稳定的泛化鲁棒性，体现了强跨分布迁移能力与对伪造模式的显式建模优势。

表 3 本文所提方法在 stable diffusion v1.4 上的良性准确率与对抗鲁棒性 单位:%

Table 3 Evaluation of the proposed method on Stable Diffusion v1.4: benign accuracy and adversarial robustness unit:%

检测器	方法	Clean	PGD	APGD	AVG
UnivFD	Vanilla	73.60	0.00	0.00	24.53
	Ours	68.14	80.12	77.34	75.20
RINE	Vanilla	88.20	43.15	40.26	57.20
	Ours	85.33	82.03	81.39	80.93
Effort	Vanilla	84.74	0.00	0.00	28.25
	Ours	80.59	76.86	75.82	77.76

注：AVG 表示平均值。

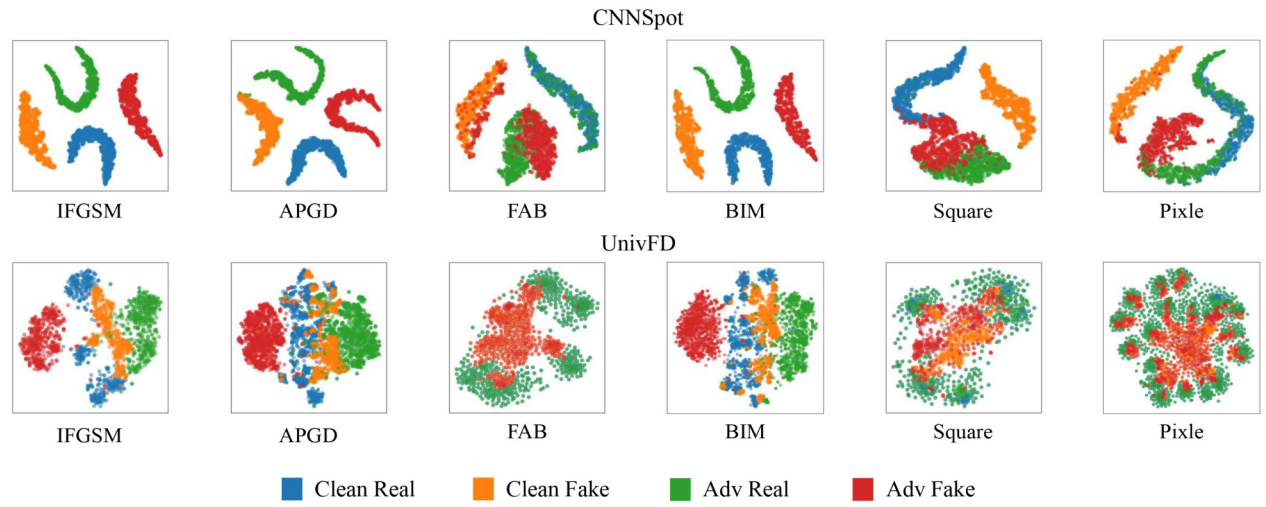
4.4 特征可视化

本节针对干净样本和对抗样本在 CNNSpot 和 UnivFD 检测器上的特征分布进行可视化分析。如图 6 所示，无论采用传统的 CNN 架构还是基于 CLIP 的多模态架构，AIGI 检测器的特征提取器所映射的特征空间中普遍存在一个固有的分布规律：干净样本（包括 Real 和 Fake）在特征空间中形成高度内聚且稳定的聚类，体现了模型对图像内容本身的良好判别能力和对微小自然变化的鲁棒性。然而，对抗样本尽管语义内容与干净样本高度一致（保证了不可察觉性），其特征向量却能被对抗扰动沿特定方向推动，使其漂移并偏离良性样本的特征流形。这种有目的的侧向漂移使得对抗样本的特征聚类脱离了原始聚类的中心，并被推向模型的决策边界附近或错误分类的一侧。实验结果显示，在白盒攻击下，对抗的真实样本和伪造样本在特征空间中形成清晰分离的簇；而在黑盒攻击下，对抗特征分布呈高度重叠。造成这一差异的原因可能在于攻击机制本身：白盒攻击能够直接利用模型梯度，有针对性地压制真实类别的预测概率，因而更易产生低熵输出；相比之下，黑盒攻击由于无法获取梯度，只能将输入推向决策边界附近，从而导致更高的不确定性。这种特征层面的可分性差异不仅揭示了不同攻击方式在优化目标与扰动方向上的本质区别，也进一步说明 AIGI 检测器内部的表示空间对对抗扰动具有结构化响应模式。值得注意的是，无论检测器基于哪种架构，其特征空间均未对这类“特征漂移型”攻击形成有效的吸收或抵消机制，导致对抗样本能够稳定地脱离干净样本流形并沿着特定的脆弱方向扩散。此外，我们观察到这种特征偏移在不同生成模型、不同训练数据以及不同攻击强度下均具有一致性，表明这种分布规律并非偶然噪声，而是 AIGI 检测模型在高维表示空间中固有的脆弱性表达。

4.5 计算开销评估

图 7 展示了本文方法与两种常见的对抗训练方法（PGD-AT 和 TRADES）在训练单个 epoch 所需时长（h）上的对比。由于本文方法采用了后训练机制，而不是从头开始训练，因此在时间效率上相比传统对抗训练方法具有明显的优势。从图 7 可以看出，在所有列出的模型中（包括 CNNSpot、GramNet、NPR、UnivFD、RINE、C2p-CLIP 和 Effort），本文方法所需的训练时长普遍低于 PGD-AT 和 TRADES。这种差距在一些模型上尤为显著，例如 RINE 和 Effort，这两个模型在 PGD-AT 和 TRADES 下的训练时间明显高于本文方法。具体来说，本文方法在 RINE 和 Effort 模型上减少了大约 10~15 h 的训练时间，而 PGD-AT 和 TRADES 的训练时长则分别保持在较高的水平。此外，PGD-AT 和 TRADES 的方法显示了较为均衡的训练时长，与之相比，本文方法表现出了更为灵活和高效的训练特点，尤其是在处理复杂模型时，其时长优势更为明显。总体来看，这一结果充分证明了本文方法在训练效率方面的显著优势，尤其适合需要高效训练场景。

尽管本文方法通过冻结特征提取器降低了部分训练成本，但引入动态对抗生成与多专家结构后，整体训练时间、显存占用及推理阶段存在额外计算开销，为验证本文所提方法的实用性，本文对比分析了原始 AIGI 检测器与引入多专家结构后的模型在参数量和推理开销方面的变化。具体而言，表 4 和表 5 给出了相应的统计结果。可以看到，引入的多专家结构仅带来了极小的参数量与推理开销增幅，与原始检测器基本保持一致，表明本文方法在保证性能提升的同时具有良好的实用性。



注:clean real表示干净的真实样本,adv real表示对抗的真实样本,clean fake表示伪造的真实样本,adv fake表示对抗的伪造样本。

图6　干净样本和对抗样本在CNNSpot和UnivFD的特征空间的聚类可视化

Figure 6　Feature clustering of clean and adversarial samples on CNNSpot and UnivFD labels represent

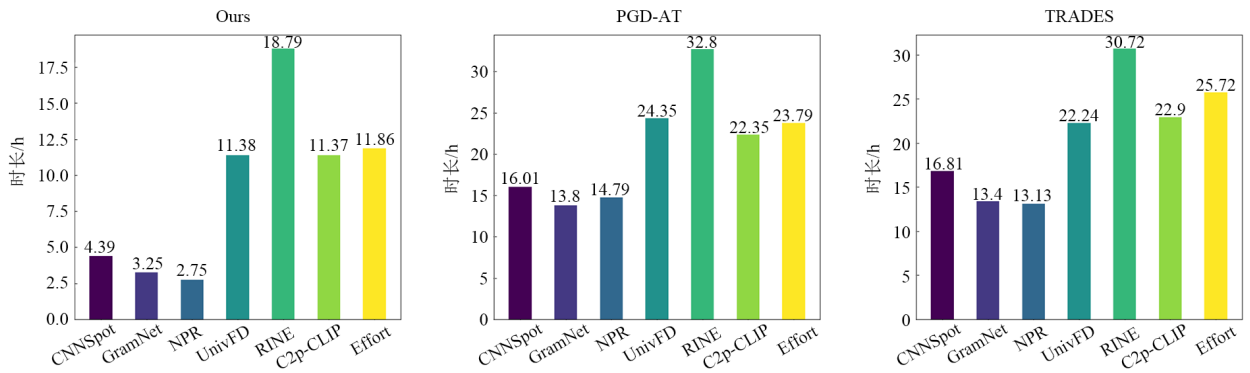


图7　本文的方法与对抗训练(PGD-AT,TRADES)相比训练单个epoch所需的时长(h)

Figure 7　Training time per epoch (hours) for the proposed method versus PGD-AT and TRADES baselines

表4　引入多专家结构前后检测器的参数量(10^6)对比

Table 4　Parameter comparison between the baseline and the proposed MoE-based detector

	CNNSpot	GramNet	NPR	UnivFD	RINE	Effort
Vanilla	23.51	11.73	1.44	427.62	433.94	427.81
Ours	28.94	12.35	2.04	428.70	435.63	428.89

表5　引入多专家结构前后检测器的推理开销(GFLOPs)对比

Table 5　Inference overhead (GFLOPs) comparison between Vanilla and proposed MoE-based models

	CNNSpot	GramNet	NPR	UnivFD	RINE	Effort
Vanilla	4.13	2.56	1.51	51.89	51.95	51.63
Ours	4.30	2.59	1.64	51.93	52.01	51.68

4.6　攻击采样策略分析

标准对抗训练要求对抗样本和对应的干净样本保持标签一致性,而本工作为对抗样本引入了额外的监督信号,显式引导模型在特征空间中学习与对抗扰动相关的决策边界结构。该设计旨在攻击先验未知的现实场景下,提升模型在多样化扰动分布下的整体鲁棒性。在此设定中,对策略池中的攻击算法采用均匀采样可以作为对未知攻击分布的一种保守近似,从而最大限度地避免模型过拟合于特定攻击模式相关的特征。

为进一步验证攻击策略池中各种攻击的比例对鲁棒性的影响,本文在保持策略池成分不变的前提下,系统性地调整不同攻击类型在训练阶段的采样比例(表6)。如表7和表8所示,当训练过程中引入大量决策边界附近的对抗样本时,模型的良性样本准确率明显下降,同时对远离决策边界的对抗样本的识别能力也会受到削弱。若训练中过度依赖以损失为导向的白盒攻击,模型在识别决策边界附近的对抗样本时同样会出现性能退化。上述现象表明,不同攻击类型在特征空间中诱导的偏移具有显著差异,对某一类攻击的过度拟合往往会以牺牲鲁棒泛化性为代价。

表6 攻击策略采样策略

Table 6 Adversarial attack strategies

采样方法	Loss-Attack	Boundary-Attack
A	0.5	0.5
B	0.2	0.8
C	0.0	1.0
D	0.8	0.2
E	1.0	0.0

注:Loss-Attack表示以损失为目标的攻击,Boundary-Attack表示以跨越边界为目标的攻击。

表7 攻击采样比例对良性准确率与泛化性能的影响(基于CNNSpot)

单位:%

Table 7 Benign accuracy and generalization performance under various attack sampling ratios (schemes A~E) using CNNSpot unit:%

采样方法	Clean	PGD	APGD	APGD-DLR	C&W	FAB	Square	AA	AVG
A	88.21	83.47	78.92	77.61	71.88	76.94	74.02	60.89	76.49
B	83.12	72.47	70.89	68.34	83.22	88.56	86.41	52.97	75.75
C	77.68	35.51	32.29	31.72	87.89	89.11	86.54	21.83	57.82
D	92.78	90.33	89.45	85.61	66.29	70.12	69.74	56.58	77.61
E	96.42	95.87	90.17	88.93	58.71	64.38	61.20	53.99	76.21

注:AVG代表平均值。

相比之下,均匀采样策略表现出显著的鲁棒泛化优势,能够有效兼顾良性样本的分类精度与各类对抗攻击下的防御能力。

4.7 消融实验

为了进行细致的消融分析,本文从以下三个方面对模型进行拆解分析。

(1)专家数量的影响。在保持其余设置不变的情况下,系统比较了不同专家数量(如K=2,4,6,8)对良性准确率和对抗鲁棒性的影响。表9表明,随着专家数量的增加,模型性能先显著提升,随后逐渐趋于饱和,说明适度增加专家数量有助于提升模型对不同对抗特征的建模能力,但过多专家带来的收益有限。

表8 攻击采样比例对良性准确率与泛化性能的影响(基于UnivFD)

单位:%

Table 8 Benign accuracy and generalization performance under various attack sampling ratio for UnivFD unit:%

采样方法	Clean	PGD	APGD	APGD-DLR	C&W	FAB	Square	AA	AVG
A	95.24	96.63	96.92	69.84	76.88	82.54	71.08	61.34	81.31
B	89.36	78.92	76.15	75.68	85.41	87.73	80.27	54.88	78.55
C	80.54	53.19	51.72	49.06	88.95	91.34	88.61	42.79	68.28
D	94.83	92.47	91.28	90.64	65.17	71.59	67.42	56.90	78.79
E	98.21	96.88	93.53	90.12	62.74	68.36	65.05	51.67	78.32

注:AVG代表平均值。

表9 专家数量K对模型平均对抗鲁棒性(AVG)的影响 单位:%

Table 9 Average adversarial robustness (AVG) versus the number of experts (K) unit:%

K	CNNSpot	GramNet	NPR	UnivFD	RINE	Effort
K=2	52.37	48.62	55.43	65.19	60.28	58.72
K=4	68.14	65.47	70.82	76.53	71.39	69.46
K=6	74.93	72.81	75.36	80.24	75.62	73.50
K=8	76.49	74.16	76.15	81.31	76.15	74.82

(2)共享专家的贡献分析。为验证共享专家的实际作用,本文对比了完整模型与移除共享专家、仅保留独立专家的变体在训练时使用的攻击以及训练时未见过的攻击上的对抗鲁棒性。表10显示,共享专家能够有效捕获跨攻击与跨生成模型的共性特征,在提升整体鲁棒性的同时显著改善模型的稳定性,验证了其设计的必要性。

表10 共享专家模块对检测性能影响的消融实验结果 单位:%

Table 10 Impact of the shared expert module on robustness unit:%

Variant	PGD	APGD	C&W	JSMA	BIM	Square	UA	Pixle	AVG
w/o Shared	70.57	64.12	68.28	63.74	68.13	72.35	57.81	67.92	66.62
w/ Shared	76.04	66.87	71.71	65.32	74.89	79.73	64.02	75.56	74.82

注:AVG代表平均值。

(3)Logit一致性正则化的作用。本文进一步分析了引入Logit一致性正则化前后的鲁棒泛化能力。表11结果表明,该正则项能够有效约束不同专家预测的一致性,减少过拟合个别攻击模式的现象,从而在面对训练时未见过的攻击方法时能够有效提升鲁棒性。

表11 Logit一致性正则化(LCR)对模型鲁棒泛化性能的影响 单位:%

Table 11 Robust generalization performance with and without Logit Consistency Regularization (LCR) unit:%

Variant	PGD	APGD	C&W	JSMA	BIM	Square	UA	Pixle	AVG
w/o LCR	74.12	63.58	68.94	62.17	71.45	76.82	61.34	72.48	68.86
w/ LCR	76.04	66.87	71.71	65.32	74.89	79.73	64.02	75.56	74.82

注:AVG代表平均值。

5 总结与展望

本文深入分析了现有对抗训练方法在AIGI(AI生成图像)检测中的局限性,并提出了基于后训练的对抗混合专家架构作为一种新型的鲁棒防御机制。通过信息论视角的分析,本文揭示了对抗训练在AIGI检测任务中的根本性问题,尤其是训练过程中出现的特征纠缠与置信度可分性下降等现象。基于这些观察,本文提出的后训练方法有效避免了对大规模模型参数的更新,在提升对抗样本识别能力的同时,保持了对干净样本的判别能力。而对抗混合专家架构进一步通过专家协作与共享专家机制,在应对多样化对抗扰动方面表现出了显著的优势。然而,在高强度的自适应白盒攻击场景下,对抗样本在特征空间中产生定向的特征偏移,并与目标类别的良性样本分布发生流形重叠。在这种情况下,对抗特征与良性特征的不可区分性导致了本文方法防御性能局限性。针对这一挑战,未来的研究方向应集中在设计更为智能化、自动化的防御策略。随着AIGI技术和对抗攻击手段的不断演化,如何提高防御机制的适应性和鲁棒性,是未来研究的重要方向。

参考文献

[1] 新京报.【防范网络诈骗】如何防范AI诈骗[EB/OL].(2025-11-25)[2026-02-27]. https://xinwen.bjd.com.cn/content/s69251de8d5de1e4309a10ee8.html.

[2] 国家互联网信息办公室,国家发展和改革委员会,教育部,等.生成式人工智能服务管理暂行办法[EB/OL].(2023-07-10)[2026-02-27]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[3] Carlini N, Farid H. Evading deepfake-image detectors with white- and black-box attacks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2020: 2804-2813.

[4] Pavlitska S, Hubschneider C, Struppek L, et al. Sparsely-gated mixture-of-expert layers for CNN interpretability[C]//Proceedings of 2023 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2023: 1-10.

[5] Wang Shengyu, Wang O, Zhang R, et al. CNN-generated images are surprisingly easy to spot… for now[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 8692-8701.

[6] Liu Zhengzhe, Qi Xiaojuan, Torr P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 8057-8066.

[7] Tan Chuangchuang, Liu Huan, Zhao Yao, et al. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 28130-28139.

[8] Ricker J, Lukovnikov D, Fischer A. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 9130-9140.

[9] Wang Zhendong, Bao Jianmin, Zhou Wengang, et al. DIRE for diffusion-generated image detection[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 22388-22398.

[10] Ojha U, Li Yuheng, Lee Y J. Towards universal fake image detectors that generalize across generative models[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 24480-24489.

[11] Liu Huan, Tan Zichang, Tan Chuangchuang, et al. Forgery-aware adaptive transformer for generalizable synthetic image detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 10770-10780.

[12] De Rosa V, Guillaro F, Poggi G, et al. Exploring the adversarial robustness of CLIP for AI-generated image detection[C]//Proceedings of 2024 IEEE International Workshop on Information Forensics and Security. Piscataway: IEEE, 2024: 10810719.

[13] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[PP/OL]. V4.arVix (2019-09-04)[2026-02-27]. https://arxiv.org/abs/1706.06083.

[14] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 86-94.

[15] Mavali S, Ricker J, Pape D, et al. Adversarial robustness of AI-generated image detectors in the real world[PP/OL]. V3. arXiv (2024-10-02) [2026-02-27]. https://arxiv. org/abs/2410.01574.

[16] Dong Chengdong, Kumar A, Liu Eryun. Think twice before detecting GAN-generated fake images from their spectral domain imprints[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recog-

nition (CVPR). Piscataway: IEEE, 2022: 7855-7864.

[17] Hou Yang, Guo Qing, Huang Yihao, et al. Evading DeepFake detectors via adversarial statistical consistency[C]// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 12271-12280.

[18] Jia Shuai, Ma Chao, Yao Taiping, et al. Exploring frequency adversarial attacks for face forgery detection[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4093-4102.

[19] Zhou Ziyin, Sun Ke, Chen Zhongxi, et al. StealthDiffusion: Towards evading diffusion forensic detection through diffusion model[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 3627-3636.

[20] 张世辉, 张晓微, 宋丹丹, 等. 基于逆扰动融合生成对抗网络的对抗样本防御方法[J]. 电子学报, 2023, 51(4): 879-884.
Zhang Shihui, Zhang Xiaowei, Song Dandan, et al. Adversarial example defense method based on inverse perturbation fusing generative adversarial network[J]. Acta Electronica Sinica, 2023, 51(4): 879-884. (in Chinese)

[21] 潘杰, 刘波, 邹筱瑜. 基于特征异常检测与伪标签回归的无监督对抗域适应[J]. 电子学报, 2025, 53(1): 128-140.
Pan Jie, Liu Bo, Zou Xiaoyu. Feature anomaly detection and pseudo-label regression for adversarial domain adaptation[J]. Acta Electronica Sinica, 2025, 53(1): 128-140. (in Chinese)

[22] 刁云峰, 姜凯超, 郭丹, 等. 基于贝叶斯能量对抗后训练的黑盒对抗防御方法[J]. 中国科学: 信息科学, 2025, 55(8): 1986-2001.
Diao Yunfeng, Jiang Kaichao, Guo Dan, et al. Post-train black-box defense through energy-based Bayesian adversarial training[J]. SCIENTIA SINICA Informationis, 2025, 55(8): 1986-2001. (in Chinese)

[23] Zhang Hongyang, Yu Yaodong, Jiao Jiantao, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proceedings of the 36th International Conference on Machine Learning (ICML). Vienna: PMLR, 2019: 7472-7482.

[24] Jin Gaojie, Yi Xinping, Wu Dengyu, et al. Randomized adversarial training via Taylor expansion[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 16447-16457.

[25] Jia Xiaojun, Zhang Yong, Wu Baoyuan, et al. LAS-AT: Adversarial training with learnable attack strategy[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 13388-13398.

[26] Diao Yunfeng, Zhai Naixin, Miao Changtao, et al. Vulnerabilities in AI-generated image detection: The challenge of adversarial attacks[PP/OL]. V6. arXiv (2024-07-30)[2026-02-27]. https://arxiv.org/abs/2407.20836.

[27] Pavlitska S, Eisen E, Zöllner J M. Towards adversarial robustness of model-level mixture-of-experts architectures for semantic segmentation[C]//Proceedings of 2024 International Conference on Machine Learning and Applications (ICMLA). Piscataway: IEEE, 2024: 1460-1465.

[28] Pavlitska S, Fan Haixi, Ditschuneit K, et al. Robust experts: The effect of adversarial training on CNNs with sparse mixture-of-experts layers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2025: 251-260.

[29] Zhang Yihua, Cai Ruisi, Chen Tianlong, et al. Robust mixture-of-expert training for convolutional neural networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 90-101.

[30] Zhang Xu, Xu Kaidi, Hu Ziqing, et al. Optimizing robustness and accuracy in mixture of experts: A dual-model approach[PP/OL]. V3. arVix (2025-05-27) [2026-02-27]. https://arxiv.org/abs/2502.06832.

[31] Qin Zhenyue, Kim D, Gedeon T. Rethinking softmax with cross-entropy: Neural network classifier as mutual information estimator[PP/OL]. V4. arXiv (2020-09-17) [2026-02-27]. https://arxiv.org/abs/1911.10688v4.

[32] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10674-10685.

[33] Yu F, Seff A, Zhang Y, et al. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015.

[34] Zhu Mingjian, Chen Hanting, Yan Qiangyu, et al. GenImage: A million-scale benchmark for detecting AI-generated image[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2023: 3398.

[35] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.

[36] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[PP/OL]. V2. arVix (2020-08-04)[2026-02-27]. https://arxiv.org/abs/2003.01690.

[37] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2017: 39-57.

[38] Croce F, Hein M. Minimally distorted adversarial examples with a fast adaptive boundary attack[PP/OL]. V2. arVix (2020-07-20)[2026-02-27]. https://arxiv.org/abs/1907.02044.

[39] Andriushchenko M, Croce F, Flammarion N, et al. Square attack: A query-efficient black-box adversarial attack via random search[M]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 484-501.

[40] Koutlis C, Papadopoulos S. Leveraging representations from intermediate encoder-blocks for synthetic image detection[C]//Proceedings of the 18th European Conference on Computer Vision (ECCV). Heidelberg: Springer, 2024: 394-411.

[41] Yan Zhiyuan, Wang Jiangming, Jin Peng, et al. Orthogonal subspace decomposition for generalizable AI-generated image detection[PP/OL]. V4. arVix (2025-05-20)[2026-02-27]. https://arxiv.org/abs/2411.15633.

[42] Nie Weili, Guo B, Huang Yujia, et al. Diffusion models for adversarial purification[PP/OL]. V1. arVix (2022-05-16)[2026-02-27]. https://arxiv.org/abs/2205.07460.

作者简介

张睿萱　女,2002年7月出生于河南省郑州市。现为合肥工业大学硕士研究生。主要研究方向为AI生成内容检测与对抗安全。
E-mail: ruixuanzhangr@gmail.com

刁云峰　男,1993年7月出生于山东省烟台市。现为合肥工业大学计算机与信息学院副教授。在国内外知名期刊/会议发表学术论文40余篇。主要研究方向为媒体内容安全、人工智能安全。中国电子学会会员编号:E190201650M。
E-mail: diaoyunfeng@hfut.edu.cn

陆智远　男,2004年12月出生于福建省南平市。现为合肥工业大学计算本科生。主要研究方向为对抗样本。
E-mail: 18063718180@163.com

夏海峰　男,1993年7月出生于山东省枣庄市。2023年博士毕业于美国杜兰大学计算机科学系。现为中山大学副教授。主要研究方向为计算机视觉、多模态学习和迁移学习。
E-mail: xiahf5@mail.sysu.edu.cn

郭治卿　男,1991年9月出生于新疆维吾尔自治区霍城县。现为新疆大学计算机科学与技术学院副教授、博士生导师。在国内外发表学术论文50余篇。
E-mail: guozhiqing@xju.edu.cn

郝孝帅　男,1994年10月出生山东省烟台市。现为小米汽车自动驾驶与具身智能算法专家。在国内外知名期刊/会议发表学术论文50余篇。主要研究方向为自动驾驶鲁棒性和具身基座大模型。
E-mail: haoxiaoshuai@xiaomi.com

汪　萌　男,1984年12月出生于湖北省监利市。现任合肥工业大学党委副书记、校长、教授、博士生导师。主要研究方向为模式识别与多媒体信息处理等。获国家杰出青年科学基金资助,入选国际电气与电子工程师协会会士、国际模式识别协会会士。中国电子学会会员编号:E190011561M。
E-mail: eric.mengwang@gmail.com