中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-13

论文引用格式： Zhu Jindong, Zhang Yujin, Zhang Tao, Wang Yongqi, Wu Fei. A fast adversarial training method with prior structure guidance[J/OL]. Journal of Image and Graphics, XXXX:1-13. DOI: 10.11834/jig.250501. (朱进东, 张玉金, 张涛, 王永琦, 吴飞. 先验结构引导的快速对抗训练方法[J/OL]. 中国图象图形学报, XXXX:1-13. DOI: 10.11834/jig.250501.)[DOI:10.11834/jig.250501]

先验结构引导的快速对抗训练方法

朱进东[1],张玉金[1],张涛[2],王永琦[1],吴飞[1]

1. 上海工程技术大学 电子电气工程学院, 上海 201620; 2. 苏州工学院 计算机科学与工程学院,江苏 常熟 215500

摘 要：目的 对抗训练(adversarial training,AT)是防御对抗攻击的主要方法,能够有效提升深度神经网络(deep neural networks,DNN)的鲁棒性。快速对抗训练(fast adversarial training,FAT)在降低计算开销的同时,易发生灾难性过拟合,导致模型鲁棒性下降。为此,本文提出了一种先验结构引导的快速对抗训练方法。**方法** 首先,设计了基于图像梯度与结构先验的扰动引导机制,指导多样化对抗样本的生成;然后,通过在连续批次之间共享对抗扰动信息,有效缓解了单步对抗训练中梯度方向收敛过快的问题;最后,构建正则化损失函数,将结构引导与分类损失联合优化,进一步提升模型的鲁棒性与收敛稳定性。**结果** 在CIFAR-10与CIFAR-100数据集上,以ResNet-18为目标网络,面对PGD-10攻击时,所提算法在CIFAR-10上的鲁棒精度比现有FAT方法提升了约2%~12%,在CIFAR-100上提升了约2%~8%;同时在干净精度保持率方面表现优异。实验结果表明所提方法不仅可以有效避免灾难性过拟合,而且可以提高模型的鲁棒性和泛化能力,能够更好地应对不同的对抗攻击。**结论** 本文方法有效结合了先验结构引导机制与快速对抗训练框架,既保持了FAT的高效性,又改善了对抗训练的稳定性,显著提升了深度神经分类网络的防御性能。

关键词： 对抗训练;对抗样本;扰动初始化;鲁棒性;结构特征

A fast adversarial training method with prior structure guidance

Zhu Jindong[1], Zhang Yujin[1], Zhang Tao[2], Wang Yongqi[1], Wu Fei[1]

1. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; 2. School of Computer Science and Engineering, Suzhou University of Technology, Changshu 215500, China

Abstract: Objective Adversarial training (AT) is one of the most effective and widely adopted defense strategies against adversarial attacks, and it has been extensively used to improve the robustness of deep neural networks (DNNs). As an efficient variant of AT, fast adversarial training (FAT) significantly reduces computational cost by employing single-step adversarial attacks to guide model optimization, making it more suitable for practical applications where efficiency is critical. However, despite its efficiency advantage, FAT is prone to catastrophic overfitting, a phenomenon in which the model rapidly overfits to adversarial samples generated during training, leading to a sharp degradation in robustness at later stages. This limitation restricts the effectiveness of FAT in scenarios requiring stable and reliable robustness. To address these issues, this paper proposes a fast adversarial training method guided by prior structural information. Specifically, the proposed method leverages inherent structural characteristics in image data to guide the generation of adversarial perturbations, thereby improving the diversity and effectiveness of adversarial samples. In addition, the proposed strategy enhances

收稿日期：2025-10-13;**修回日期：**2026-05-08

基金项目：国家自然科学基金资助项目(62072057);上海市自然科学基金资助项目(17ZR1411900);中国高校产学研创新基金资助项目(2021ZYB1003)

Supported by: National Natural Science Foundation of China(62072057); Shanghai Natural ScienceFoundation(17ZR1411900);

the stability of the training process and promotes better generalization, ultimately achieving improved robustness while maintaining high computational efficiency. **Method** The proposed method introduces a structure-guided adversarial training paradigm that integrates the intrinsic geometric and texture properties of images into both the perturbation generation and model optimization processes. Specifically, a structure-aware perturbation guidance mechanism is developed, which combines conventional gradient-based optimization with prior structural cues extracted from image edges, gradient magnitudes, and local texture features. Instead of relying solely on pixel-wise gradients, the perturbation direction is adjusted to align with the dominant structural features of the image, improving perturbation diversity and stability. This mechanism effectively reduces random high-frequency noise within the adversarial perturbations, thereby stabilizing the optimization trajectory and enhancing the reliability of the training process. In addition, to further mitigate catastrophic overfitting, an inter-batch perturbation sharing strategy is introduced. In this design, part of the adversarial perturbations generated in one batch are adaptively propagated to the next batch, establishing continuity between consecutive training iterations. This mechanism increases the temporal coherence of perturbation distributions and prevents the model from converging too quickly to a narrow gradient subspace, which often leads to overfitting. As a result, the model is exposed to a broader range of adversarial patterns, encouraging more generalizable feature representations and a smoother optimization landscape. Finally, a regularized joint loss function is constructed to improve both convergence stability and robustness. The overall loss integrates a structural regularization term, derived from spatial smoothness and gradient constraints, into the standard cross-entropy objective. This joint optimization formulation penalizes deviations from the natural image structure and promotes consistency between the learned representations and underlying visual priors. Together, these components form an efficient and unified fast adversarial training framework that effectively leverages structural guidance, inter-batch continuity, and regularized optimization to improve model robustness and stability against adversarial perturbations. **Result** To verify the effectiveness of the proposed method, experiments were conducted on the widely used CIFAR-10 and CIFAR-100 benchmark datasets, using ResNet-18 as the target network architecture. The trained models were evaluated under multiple adversarial attack settings, including FGSM, PGD-10, PGD-20, and AutoAttack. Experimental results demonstrate that, under the PGD-10 attack scenario, the proposed method achieves a robust accuracy improvement of approximately 2%−12% on CIFAR-10 and 2%−8% on CIFAR-100 compared with existing FAT approaches, while maintaining excellent clean accuracy. Moreover, the learning curves show that the proposed method effectively suppresses catastrophic overfitting, maintaining stable robustness throughout training and ensuring smooth convergence. Qualitative analysis of the adversarial examples further shows that the generated perturbations exhibit greater structural coherence and fewer random artifacts compared to those produced by conventional FAT. Ablation experiments verify that each component of the method contributes to performance improvement: removing the structure-aware perturbation mechanism leads to a noticeable drop in robustness, while disabling the inter-batch perturbation sharing mechanism accelerates overfitting and reduces model stability. These findings consistently demonstrate that the proposed structural guidance strategy provides a more reliable and interpretable form of regularization, improving both the robustness and generalization of the trained models under adversarial conditions. **Conclusion** In summary, this work presents a fast adversarial training approach guided by prior structural information, effectively combining the efficiency of single-step training with enhanced robustness and stability. By leveraging the inherent geometric and texture features of image data to guide perturbation generation, the method prevents catastrophic overfitting and facilitates more stable optimization. The introduction of inter-batch perturbation sharing further diversifies adversarial examples and strengthens generalization, and ensures structural consistency between model representations and natural images. Experimental results on benchmark datasets confirm that the proposed method significantly improves the robustness and defense capability of deep neural networks compared with existing fast adversarial training techniques, without additional computational burden. The approach thus provides an efficient and practical framework for robust deep learning, offering a promising direction for developing low-cost yet high-robustness models applicable to a wide range of visual recognition tasks.

Key words: adversarial training; adversarial example; perturbation initialization; robustness; structural feature

论文引用格式: Zhu Jindong, Zhang Yujin, Zhang Tao, Wang Yongqi, Wu Fei. A fast adversarial

training method with prior structure guidance[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.250501.(朱进东,张玉金,张涛,王永琦,吴飞. 先验结构引导的快速对抗训练方法[J/OL]. 中国图象图形学报. DOI:10.11834/jig.250501.)

0 引 言

随着深度学习模型在图像识别(Carion 等,2020)、目标检测(Zhao 等,2019;潘晓英 等,2023)、自动驾驶(陈妍妍 等,2024)等领域取得了出色的发展,其安全问题逐渐暴露,并引起国内外学者的广泛关注和系统研究(隋晨宏 等,2023)。在图像分类任务中,Szegedy 等(2014)发现向原始数据添加难以察觉的对抗性扰动生成对抗样本,导致深度学习模型误分类,揭示了深度神经网络的脆弱性。为了解决对抗样本对深度学习模型的攻击,研究人员提出了众多防御算法,其中,对抗训练(adversarial training,AT)(Madry 等,2018)是防御对抗攻击的有效方法之一。标准对抗训练采用投影梯度下降法(projected gradient descent,PGD)(Madry 等,2018)生成用于训练的对抗样本。但是该方法采用多步迭代,需要反复计算梯度和执行前后向传播,大幅增加了训练负担。相较之下,Goodfellow 等人(2015)采用的快速梯度符号方法(fast gradient sign method,FGSM)在训练时只需进行一次梯度运算,有效降低了对抗训练的计算负担,但是对于模型鲁棒性的提升有限,同时引发灾难性过拟合现象(catastrophic overfitting,CO)(Kim 等,2021),该现象通常在训练初期表现为鲁棒精度快速提升,而在训练后期迅速崩塌,使模型仅对单步攻击保持过拟合的抗性,面对 PGD 等复杂对抗攻击时鲁棒性大幅度下降。

对此,研究人员致力于解决快速对抗训练中的灾难性过拟合问题,同时探索提升模型鲁棒性的方法。Wong 等人(2020)提出将随机扰动注入原始样本中构建更多样化的对抗样本,减轻了训练过程中的过拟合现象,该方法被称为基于随机初始化的快速梯度符号法对抗训练(FGSM AT with random start,FGSM-RS)。Kim 等(2021)提出了检查点式的步长搜索策略,在训练过程中动态选择合适的步长缓解过拟合问题,该方法称为 FGSM-CKPT(FGSM with checkpoint)。Jorge 等(2022)发现将更强的随机噪声作为数据增强手段且不进行裁剪,可以有效避免 CO 的产生,并构建了 N-FGSM(noise-FGSM)方法。Huang 等(2023)从对抗样本的梯度范数角度出发,提出 ATAS(AT with adaptive step size),根据每个实例样本的梯度范数自适应调节攻击步长避免 CO。Pan 等(2024)通过在不同样本间共享扰动方向取代随机初始化,提出 FGSM-UAP(FGSM with universal adversarial perturbation)来提升模型鲁棒性与训练稳定性。Jiang 等(2025)根据数据集各类别的训练状态动态分配正则化权重与标签松弛因子,从而缓解类别间鲁棒性差异和训练过拟合问题,称为 SKG-FAT(self-knowledge guided fast adversarial training)。

上述方法在缓解快速对抗训练不稳定问题方面取得了一定进展,但仍存在两个方面的不足。首先,现有研究主要从优化策略或梯度几何角度分析灾难性过拟合现象,较少关注扰动在输入空间中的结构分布特性;其次,通常将对抗扰动建模为均匀分布的噪声,未充分考虑图像自身结构与纹理特征对扰动有效性的差异性。近年来已有研究尝试利用图像边缘或纹理信息约束对抗扰动的空间分布,以增强扰动与图像语义结构的一致性,例如 Abdukhamidov 等(2021)提出基于边缘感知的对抗扰动生成方法 AdvEdge,这类方法多聚焦于对抗样本质量或攻击特性的提升,并未针对快速对抗训练中的训练稳定性机制进行专门设计。

为此,本文提出了一种数据增强融合策略,即先验结构引导的快速对抗训练,从图像自身结构角度分析,可以有效指导对抗样本的生成,提升对抗样本的针对性和多样性。引入的跨批次对抗样本初始化可以缓解训练早期的优化震荡,并通过正则化策略可以防止当前扰动偏离先验扰动太多,有效提升模型的鲁棒性。

1 基本原理

1.1 对抗攻击

Szegedy 等(2014)首次引入对抗样本的概念,即向干净图像样本中添加人类难以察觉的扰动,使得深度神经网络识别错误,达到攻击者所预期的目标。Goodfellow 等(2015)提出的 FGSM 可以简单而高效地生成对抗样本,公式如下:

$$\delta = \varepsilon \cdot sign(\nabla_x L(f_w(x), y)) \tag{1}$$

式中，x代表干净图像样本，δ为生成的对抗扰动，ε为扰动强度，$sign$为符号函数，$f_w(\cdot)$是参数为w的目标网络，L为损失函数，∇_x表示x的梯度。FGSM的优点是计算开销极小，适合用于大规模数据集，但其单步线性近似有时无法逼近最强对抗方向。为了提高攻击强度与可靠性，Madry等(2018)使用多步PGD来生成对抗样本，与FGSM相比，该方法的精度更高，生成的对抗样本攻击更强，但计算成本较高，公式如下：

$$\delta_{t+1}=\prod_{[-\varepsilon,\varepsilon]}[\delta_t+\alpha\cdot sign(\nabla_x L(f_w(x+\delta_t),y))] \tag{2}$$

式中，δ_{t+1}表示第$t+1$次迭代时的对抗扰动，δ_t表示第t次迭代时的对抗扰动，α表示攻击步长，$\prod_{[-\varepsilon,\varepsilon]}$表示将扰动映射到区间$[-\varepsilon,\varepsilon]$的投影。迭代的步数越多，攻击越强，计算成本越高。

不同于单纯依赖梯度的攻击方式，Carlini和Wagner(2017)提出一种基于优化求解的对抗样本生成策略，被称为C&W攻击(Carlini and Wagner attack)。Croce等(2020)提出将多种无参数攻击(如APGD、FAB等)组合成一个有序攻击集，对目标模型进行自动化攻击，用以提供更准确定量的鲁棒性评估，被称为自动攻击(AutoAttack，AA)。

1.2 对抗防御

对抗训练是防御对抗样本的最有效方法之一，对抗训练框架由Madry等提出，通常可以看作是极大极小优化问题(Wang等，2019)，其公式如下：

$$\min_w E_{(x,y)\sim D}[\max_{\delta\in\Omega} L(f_w(x+\delta),y)] \tag{3}$$

式中，D表示训练数据，Ω表示扰动空间，x表示输入样本，y表示对应的标签，w表示模型权重参数，δ表示生成的对抗扰动，L表示损失函数，$f_w(\cdot)$表示参数为w的目标网络。其目标是通过对抗攻击找到一个扰动使得模型的损失最大化，然后该扰动生成的对抗样本训练模型以最小化损失来提升模型的鲁棒性。

2 本文方法

2.1 模型架构

该方法的详细流程如图1所示，在整个流程中，采用FGSM算法生成对抗扰动，由干净图像得到结构二值化掩膜M和纹理二值化掩膜$\overline{\mathrm{M}}$(1 − M)对该对抗扰动进行分解，将分解后的扰动和干净图像结合生成两种对抗样本，随机混合这两种对抗样本生成最终的对抗样本，将对抗样本输入目标网络进行训练，更新网络模型参数提高模型的鲁棒性。在训练过程中，将上一批次的历史扰动作为下一批次训练的初始扰动，既能加速对抗样本生成，又能让训练更平滑(扰动的连续性)。

通过以上架构，本文方法利用图像的结构信息和对抗扰动的连续性，生成具有针对性和多样性的对抗样本，有效地提高目标模型的对抗鲁棒性。

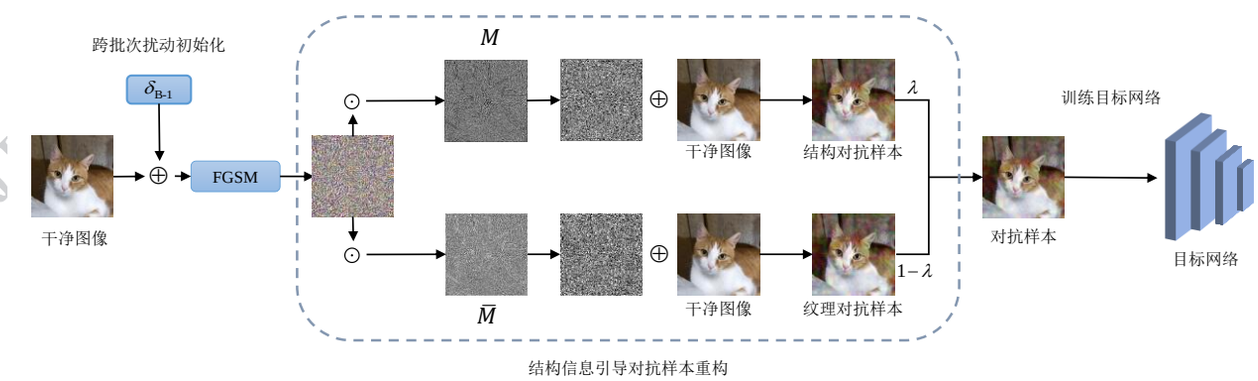


图1 先验结构引导的快速对抗训练方法流程图

Fig. 1 Flowchart of fast adversarial training method with prior structure guidance

2.2 先验结构指导的对抗样本生成

现有研究工作已经证实单步对抗训练中对抗样本初始化方法的重要性，适当的初始化可以有效缓解灾难性过拟合。Jia等人(2022)利用先验扰动信息和动量改善初始化来解决CO问题，但是该方法需要大量的存储器资源来加载整个数据集，还需要维

护扰动和动量的大型张量，对于较大规模的数据集是十分不利的。本文利用批次先验扰动作为扰动初始化实现连续对抗训练，其计算公式如下：

$$\delta_{\mathrm{B}}=\prod_{[-\varepsilon,\varepsilon]}[\delta_{\mathrm{B}-1}+\alpha\cdot sign(\nabla_x L(f_w(x+\delta_{\mathrm{B}-1}),y))] \tag{4}$$

式中，δ_{B}表示当前批次扰动，$\delta_{\mathrm{B}-1}$表示上一批次的扰动作为当前批次的初始化扰动，不仅加速模型的收敛速度，缓解对抗训练的梯度波动，且大大减少了大规模存储和管理成本。该方法在有限资源条件下可以实现高效且稳定的连续性对抗训练。

Adachi等(2023)通过引入一种随机掩膜机制构造两种对抗样本，其公式如下：

$$\begin{aligned}\rho_i&=x_i+\delta_i\odot M,\\ \bar{\rho}_i&=x_i+\delta_i\odot(1-M),\\ \xi_i&=\lambda^* y_i+(1-\lambda^*)\bar{y}_i s,\\ \bar{\xi}_i&=\lambda^*\bar{y}_i s+(1-\lambda^*)y_i\end{aligned} \tag{5}$$

式中，$M\in\{0,1\}^{H\times W}$为二值化掩膜，$\odot$表示哈达玛积，利用随机二值化掩膜M对公式(4)计算得到的δ_i进行分解后与干净样本x_i相加，得到两种对抗样本ρ_i和$\bar{\rho}_i$，ξ_i和$\bar{\xi}_i$为ρ_i和$\bar{\rho}_i$对应的平滑标签，y_i表示真实标签(one-hot向量)，s表示非真实类别平均分配概率，即$1/(K-1)$，其中K为类别数量，$\bar{y}_i$代表非真实类别的指示向量，定义为$\bar{y}_i=1-y_i$(1为全1向量)，将其与s相乘后构成非真实类别的均匀概率分布，λ^*为平滑因子，其计算公式如下：

$$\lambda^*=\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}M(i,j)}{H\times W} \tag{6}$$

式中，λ^*表示随机掩膜区域面积与图像整体面积之比。

接下来，将这两种样本进行随机混合，增加对抗样本的多样性，混合的计算公式如下：

$$\begin{aligned}x_{adv}&=\lambda\rho_i+(1-\lambda)\bar{\rho}_i,\\ y_{adv}&=\lambda\xi_i+(1-\lambda)\bar{\xi}_i\end{aligned} \tag{7}$$

式中，x_{adv}表示混合之后的对抗样本，y_{adv}表示混合之后对抗样本对应的平滑标签，λ表示随机混合系数，是一个从区间[0,1]均匀分布中采样的随机变量，用于控制两种对抗样本的混合比例。

此随机掩膜和混合策略旨在提升对抗样本的多样性，避免模型过拟合至某一种特定的扰动模式，鼓励模型学习到更稳定的对抗鲁棒特征。

随机掩膜虽然带来了一定的多样性，但忽视了图像中容易误导模型的关键位置。提出基于结构二值化掩膜引导的对抗扰动分解与融合策略，使对抗攻击更加高效且具有针对性。

本文采用Sobel算子提取图像的结构信息，Sobel的一阶梯度响应能够有效刻画图像中的轮廓与边缘区域，这些区域通常对应视觉语义变化较为敏感的位置。相比于Canny等复杂边缘检测方法，Sobel计算开销低，无额外参数调节，适合嵌入快速对抗训练流程，保持FAT训练效率的同时提供稳定的结构先验信息，基于Sobel算子生成的梯度幅值图可表示为：

$$\begin{aligned}G_a(i,j)&\approx(I(i+1,j)-I(i-1,j))/2,\\ G_b(i,j)&\approx(I(i,j+1)-I(i,j-1))/2,\\ G(i,j)&=\sqrt{G_a(i,j)^2+G_b(i,j)^2}\end{aligned} \tag{8}$$

式中：$G_a(i,j)$表示像素点(i,j)在水平方向上的导数，$G_b(i,j)$表示像素点(i,j)在垂直方向上的导数，$G(i,j)$表示像素点(i,j)的梯度幅值。

为了抑制梯度幅值图中的噪声并突出关键结构特征，设计了一种基于自适应阈值的二值化掩膜生成方法，通过比较目标像素梯度与其局部周围像素的平均梯度，来决策该点是否属于显著结构特征点，从而生成二值化掩膜M，其公式如下：

$$\begin{aligned}C(i,j)&=\frac{1}{\left|L_k(i,j)\right|}\sum_{(p,q)\in L(i,j)}G(p,q),\\ M(i,j)&=\begin{cases}1 & G(i,j)\geqslant C(i,j)\\ 0 & otherwise\end{cases}\end{aligned} \tag{9}$$

式中，$L_k(i,j)$表示以像素(i,j)为中心，大小为$k\times k$的邻域窗口，$C(i,j)$表示局部均值滤波图，在每个掩膜生成的过程中引入一种多尺度随机感受野策略，随机从尺度集合{3,5,7}中采样一个值作为当前窗口大小k，二值化掩膜生成的过程如图2所示。

从图2中可以看出，对干净图像进行Sobel边缘检测得到梯度幅值图，采用随机尺寸的均值滤波核与梯度幅值图进行卷积操作得到局部均值滤波图，

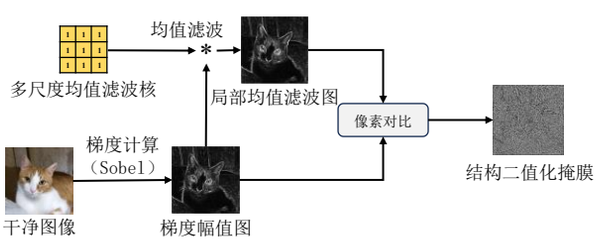


图2　二值化掩膜的生成过程

Fig. 2　Generation process of the binarized mask

将局部均值滤波图的像素值和梯度幅值图的像素值进行比较，根据公式(9)生成二值化掩膜。利用局部二值化编码能够准确地保留细微边缘，提取图像的关键结构信息，结构二值化掩膜能够指导对抗样本的生成，加速对抗训练的收敛。随机选取的卷积核可使掩膜在不同尺度上覆盖多样化的结构特征，增强对抗样本的多样性，提高模型的鲁棒性。

2.3 正则化损失

为了更好地指导训练过程，提升鲁棒性，引入新的正则化损失函数，在标准交叉熵损失的基础上，加入正则项，公式如下：

$$L_{adv}=L_{CE}+\gamma\left|\left\|f(x+\delta_B)-f(x+\delta_{B-1})\right\|_2-\left\|f(x_{adv})-f(x+\delta_B)\right\|_2\right| \tag{10}$$

式中，L_{adv}表示最终用于训练的对抗损失，L_{CE}表示标准交叉熵损失，γ表示正则项权重系数，控制正则化项对总损失的影响程度。δ_{B}表示当前批次生成的对抗扰动，$\delta_{\mathrm{B}-1}$表示上一批次保存的历史扰动，x表示输入样本，x_{adv}表示最终构造的对抗样本，$f(\cdot)$表示模型的输出，$\|\cdot\|_2$表示L_2范数。该正则项的设计通过约束当前扰动与历史扰动之间的模型输出差异，并结合当前最终对抗样本的输出，防止扰动在训练中剧烈震荡，从而提升训练的稳定性与最终的鲁棒性。

2.4 算法描述

本文提出的先验结构的连续快速对抗训练方法的具体步骤如下：

首先，获取随机均匀噪声作为对抗扰动初始化。训练开始后采用FGSM对抗攻击方法得到对抗扰动δ，接着利用结构二值化掩膜和标签平滑因子将对抗扰动δ进行分解，得到两种对抗样本ρ_i和$\bar{\rho}_i$以及其对应的平滑标签ξ_i和$\bar{\xi}_i$。

接着，将两种对抗样本及其对应的平滑标签进行随机混合生成最终的对抗样本x_{adv}及其对应的y_{adv}，使用对抗样本进行对抗训练，更新网络模型参数θ，最后将对抗扰动δ作为下一批次的初始对抗扰动进行传递。具体算法流程如算法1所示。

算法1 先验结构引导的快速对抗训练算法

输入：训练数据集D，批次大小n，训练周期数T，学习率η，最大扰动因子ε，步长α，前一批次扰动δ_{B-1}，下一批次扰动δ_B，由θ参数化的目标网络$f(\cdot)$，随机变量U(服从均匀分布)

输出：参数优化后的目标网络$f(\cdot)$

for $t=1,\cdots,T$ **do**
  $B\leftarrow 1$
  for $\{x_i,y_i|i=1,\cdots,n\}\sim D$ **do**
    $\delta=\prod_{[-\varepsilon,\varepsilon]}[\delta_{B-1}+\alpha\cdot sign(\nabla_{x_i}L(f_w(x+\delta_{B-1},\theta),y_i))]$
    $\rho_i,\bar{\rho}_i,\xi_i,\bar{\xi}\leftarrow\varphi(x_i,\delta,y_i)$ 注：函数$\varphi(x,\delta,y)$表示基于结构掩膜引导的对抗样本和其对应的平滑标签生成过程，其具体步骤详见正文公式(5)、(6)、(8)、(9)
    $\lambda\sim U[0,1]$
    $x_{adv}=\lambda\rho_i+(1-\lambda)\bar{\rho}_i$
    $y_{adv}=\lambda\xi_i+(1-\lambda)\bar{\xi}_i$
    $\theta\leftarrow\theta-\eta\nabla L_{adv}(f(x_{adv},\theta),y_{adv})$
    $\delta_B\leftarrow\delta$
    $B\leftarrow B+1$
  end for
end for
return θ

3 实验与分析

本文在公开数据集CIFAR-10、CIFAR-100(Krizhevsky 等，2009)和Tiny ImageNet(Deng 等，2009)上评估了所提方法在防御对抗样本方面的能力，分别选取了PGD-10、PGD-20、PGD-50(Madry 等，2018)、C&W(Carlini 和 Wagner，2017)和AA(Croce 等，2020)对训练所得模型施加攻击，将实验结果与现有的多种快速对抗训练方法作对比。

3.1 模型参数设置

在实验中，在CIFAR-10、CIFAR-100和Tiny ImageNet数据集上选择ResNet18(He 等，2016)、PreActResNet18(He 等，2016)、WideResNet34-10(Zagoruyko 等，2016)和DeiT-Ti(Touvron 等，2021)作为目标网络进行训练。训练周期设置为110次，采用SGD(Qian 等，1999)作为优化器并设置权重衰减系数为5e-4，动量参数为0.9，设置每个训练批次包含128个样本，对于学习率的设置，初始学习率设为0.1，在第100个周期和第105个周期将学习率以0.1的衰减因子进行衰减，所有实验扰动预算设为8/255。对于DeiT-Ti模型，采用预训练模型进行微调训练，与现有vision transformer对抗训练研究中常用的训练策略(Mo 等，2022)保持一致。对于本文方法的参数，正则化系数设置为12。

3.2 对比实验

本文将一些先进的快速对抗训练方法作为基线，即FGSM-RS(Wong 等，2020)、FGSM-CKPT(Kim 等，2021)、N-FGSM(Jorge 等，2022)、FGSM-UAP

(Pan 等,2024)。除此之外,本文还与标准训练框架,即常用的多步对抗训练方法PGD-AT(Madry 等,2018)作对比。实验将干净精度和不同攻击算法下的鲁棒精度作为主要的评价指标来综合评估模型性能,干净精度衡量的是模型对未经过扰动的原始样本的分类能力,鲁棒精度衡量的是模型在遭受对抗攻击时的稳健性。为确保公平比较,所有对比实验均遵循原始文献中的实验设置和训练流程。

3.2.1 CIFAR-10上性能对比与评估

在CIFAR-10数据集上将ResNet18作为目标模型进行训练,比较本文方法与其他快速对抗训练方法,针对每种方法,均选取了训练周期内鲁棒精度达到最佳的模型与训练完成时的最终模型进行性能评估,同时对干净样本的分类精度进行了评估,旨在综合验证本文方法的鲁棒性和泛化性,实验结果如表1所示。由表1可知,本文所提出的方法在不同攻击方式下相比于其他快速对抗训练方法均取得了最高的鲁棒精度。虽然所提出方法的干净精度未达到最佳,但也有良好的分类表现,在鲁棒精度和干净精度之间达到了较好的平衡。例如,在PGD-10的攻击下,本文方法的鲁棒精度比N-FGSM(Jorge 等,2022)方法高4.27%。在C&W的攻击下,相较于先进的FGSM-UAP(Pan 等,2024),本文方法提高了2.65%的性能且干净精度高了5.06%,在AutoAttack的攻击下,本文方法的鲁棒精度达到48.97%,比FGSM-CKPT(Kim 等,2021)方法高11.79%。从训练时长统计结果可以看出,本文方法的计算开销与现有快速对抗训练方法处于同一数量级,相比PGD-AT(Madry 等2018)具有明显的时间优势。实验表明,本文所提方法的性能显著优于多种现有方案,得益于先验结构信息指导生成的对抗样本具有针对性和多样性。

为进一步评估本文方法与其他先进方法的性能,绘制了各方法在训练周期内的鲁棒精度变化曲线,如图3所示。从图3中可以看出,随着训练周期的增加,FGSM-RS出现了灾难性过拟合现象,本文方法鲁棒精度稳步提升,且鲁棒精度整体高于其他算法,说明本文方法能够有效解决快速对抗训练中的灾难性过拟合问题,有效提升模型鲁棒性,证明了该方法的有效性。

为了验证本文方法对不同网络的适应性,在同样的参数设置下,将目标网络更换为PreActResNet18,对模型的鲁棒精度变化曲线作图,结果如图4所示,在目标网络不同的情况下,鲁棒精度的

表1 CIFAR-10上训练ResNet18的各方法测试结果

Tab. 1 Comparative test results of different methods on ResNet-18 using CIFAR-10

目标网络	方法	模型	干净精度%	鲁棒精度%					训练时长 h
				PGD-10	PGD-20	PGD-50	C&W	AA	
ResNet-18	PGD-AT	最好	82.42	53.75	52.96	52.62	51.22	**48.86**	4.40
		最后	82.74	52.13	52.51	52.26	51.07	48.71	
	FGSM-RS	最好	74.25	42.98	41.77	41.49	39.95	37.41	**0.80**
		最后	83.87	00.07	00.03	00.02	0.00	0.00	
	FGSM-CKPT	最好	**89.63**	44.33	42.96	40.35	41.95	37.18	1.25
		最后	**88.91**	42.82	42.07	40.35	41.03	37.18	
	N-FGSM	最好	80.63	50.52	49.82	49.03	48.84	47.12	**0.80**
		最后	79.82	50.13	49.36	48.78	47.25	46.17	
	FGSM-UAP	最好	79.43	52.71	52.39	52.11	49.36	48.55	1.39
		最后	78.91	52.24	51.98	51.76	49.36	47.89	
	本文方法	最好	84.49	**54.79**	**53.45**	**53.09**	**52.01**	**48.97**	1.12
		最后	83.83	**54.23**	**53.06**	**52.82**	**51.35**	48.21	

注:加粗字体为每列最优两个值。

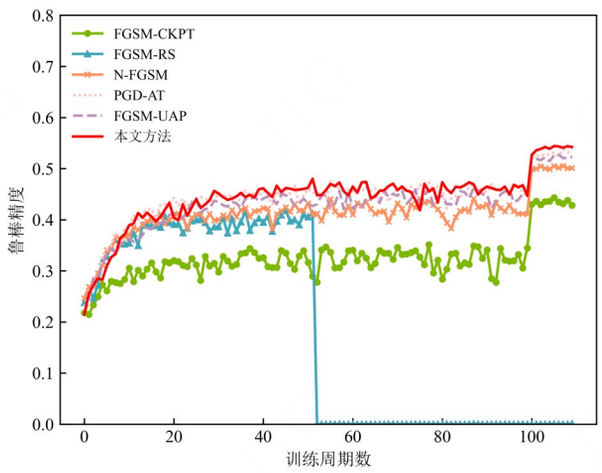


图3　ResNet18框架下训练CIFAR-10数据集时不同对抗训练方法结果对比

Fig. 3　Comparison of different adversarial training methods in training ResNet18 on CIFAR-10

变化趋势与目标网络为ResNet18时保持一致，该结果表明本文方法在不同卷积神经网络结构下呈现出一致的鲁棒性变化趋势。

为进一步验证本文方法在不同网络范式下的适用性，本文在Transformer架构的DeiT-Ti网络上进行对抗训练实验，结果如表2所示，相比FGSM-RS，本文方法在PGD-10攻击下的鲁棒精度提高了3.02%，与PGD-AT相比，本文方法在显著降低计算开销的同时，仍能保持与其接近的鲁棒精度，并具有更高的干净精度。上述结果表明，本文方法在vision transformer架构下同样具有良好的适用性，并在鲁棒性与效率之间取得了较好的平衡，证明了该方法在不同模型结构下具有良好的泛化能力。

3.2.2　CIFAR-100上性能对比与评估

将ResNet18在复杂度更高的CIFAR-100数据集上进行训练，各方法对比结果如表3所示，本文方法依然保持了性能优势，所提方法的鲁棒性在PGD-10的攻击下比FGSM-CKPT（Kim 等，2021）提高了7.77%，在C&W攻击下，其鲁棒精度比N-FGSM（Jorge 等，2022）提高了1.63%，在AutoAttack攻击下比FGSM-UAP（Pan 等，2024）提高了1.1%且保持了良好的干净精度。在该框架下，分析各个算法鲁棒精度的变化趋势，结果如图5所示，由此可知，本文所提方法具有以下优越性：1）采用跨批次对抗样本初始化和先验结构引导的对抗样本生成，可以提高对抗样本的针对性和多样性，有效避免灾难性过拟合。2）在标准交叉熵损失的基础上，融入正则项，从而构建的正则化损失，能够有效地提高系统的鲁棒性。

本文方法与较先进的FGSM-MEP（Jia 等，2022）方法对比，实验结果如表4所示。在WideResNet34-10的复杂结构框架下，本文方法在干净精度和鲁棒精度上均取得了最佳性能，鲁棒性相

比于FGSM-MEP提高了4.66%。对比结果如图6所

表2　CIFAR-10上训练DeiT-Ti的各方法测试结果

Tab. 2　Comparative test results of different methods on DeiT-Ti using CIFAR-10

目标网络	方法	干净精度%	鲁棒精度%					训练时长h
			PGD-10	PGD-20	PGD-50	C&W	AA	
DeiT-Ti	PGD-AT	75.46	48.62	48.10	47.92	45.40	43.62	3.19
	FGSM-RS	80.48	42.37	41.92	41.83	39.05	37.29	0.64
	本文方法	81.50	45.39	44.92	44.55	42.03	40.18	0.93

示。从图6中可以看出，FGSM-MEP出现了灾难性过拟合现象。并且，FGSM-MEP需要维护两个全数据集的张量，而本文方法只需要维护一个批次的张量，大大节省了计算机设备的内存。由此可见，本文方法在处理更复杂的数据和网络结构方面具有显著的优势。

3.2.3　Tiny ImageNet上性能对比与评估

在分辨率更高且结构特征更复杂的Tiny ImageNet数据集上对目标网络ResNet-18进行训练，实验对比结果如表5所示，由表5可知，实验结果与CIFAR系列数据集基本一致，本文方法在保持较高干净精度的同时，鲁棒精度提升较为明显，在高分辨率场景下依然保持稳定优势。表明先验结构引导机

表3　CIFAR-100上训练ResNet18的各方法测试结果

Tab. 3　Comparative test results of different methods on ResNet-18 using CIFAR-100

目标网络	方法	模型	干净精度%	鲁棒精度%					训练时长h
				PGD-10	PGD-20	PGD-50	C&W	AA	
ResNet-18	PGD-AT	最好	58.15	**29.47**	28.62	28.46	**28.41**	**25.46**	4.73
		最后	57.98	29.36	28.53	28.40	**28.32**	**25.46**	
	FGSM-RS	最好	50.66	23.57	23.02	22.83	22.65	19.29	**1.17**
		最后	63.87	00.32	00.13	00.07	0.00	0.00	
	FGSM-CKPT	最好	**64.02**	21.97	21.23	20.82	20.77	18.37	1.60
		最后	**64.14**	21.21	20.76	20.19	20.64	18.22	
	N-FGSM	最好	57.15	27.71	27.32	26.89	24.52	22.91	**1.17**
		最后	57.45	27.35	27.09	26.73	24.47	22.69	
	FGSM-UAP	最好	55.83	28.02	27.92	27.15	25.34	23.31	1.82
		最后	55.59	27.41	27.11	27.02	25.19	23.07	
	本文方法	最好	62.52	**29.74**	**29.62**	**29.06**	26.15	24.41	1.68
		最后	62.53	29.32	**29.11**	**28.93**	26.12	24.29	

注：加粗字体为每列最优两个值。

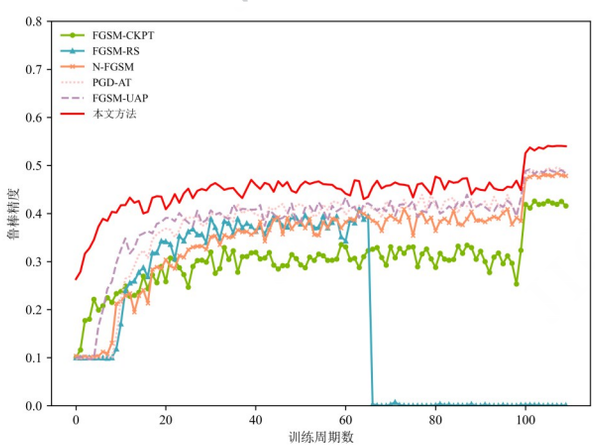


图4　PreActResNet18框架下训练CIFAR-10数据集时不同对抗训练方法结果对比

Fig. 4　Comparison of different adversarial training methods in training PreActResNet18 on CIFAR-10

制在复杂视觉结构下仍然有效，具有良好的跨数据集泛化能力。

3.3　性能分析

为了证明本文所提方法可以提升快速对抗训练的性能，将本文方法和其他快速对抗训练方法进行定性分析比较，并将损失景观进行了可视化，可视化结果如图7所示。其中，random为随机方向，adversarial为对抗扰动方向。为了保证公平性，训练好的模型均采用PGD-10攻击，为了探索模型决策边界的局部线性特征，绘制了交叉熵损失沿对抗方向和随机方向的损失曲面。通过观察图7可以发现，与FGSM-RS（Wong 等，2020）相比，FGSM-RS损失表面发生了严重的扭曲，表明其发生了灾难性过拟合，所提方法损失面平坦，说明其能较好地解决灾难性过拟合问题。与FGSM-CKPT（Kim 等，2021）、N-FGSM（Jorge 等，2022）和FGSM-UAP（Pan 等，2024）相比，本文方法能够更好地保持目标网络模型的局部线性，损失变化值更小损失面更加平坦，证明了所提方法的优越性。

除了上述可视化分析外，模型中正则化权重γ的取值也会影响结构约束对训练过程的调节效果。为验证本文方法对该参数的稳定性，进一步进行了超参数敏感性分析。在CIFAR-10数据集上将γ取值设为$\{4,8,12,16\}$，并在PGD-10攻击下分别评估模型的干净精度和鲁棒精度。如表6所示，γ的变化对模型干净精度影响较小，不同取值下的最大波动不超过约1.5%。相比之下，鲁棒精度随γ增大呈现先上升后趋于平稳的趋势，当γ从4增加至12时，模型鲁棒性能明显提升，而继续增大γ时性能提升幅度有限。总体来看，本文方法在较宽范围内对γ的取值并不敏感，模型性能保持稳定。综合干净精度

表4　CIFAR-100上训练WideResNet34-10的各方法测试结果

Tab. 4　Comparative test results of different methods on WideResNet34-10 using CIFAR-100

目标网络	方法	模型	干净精度%	鲁棒精度%				
				PGD-10	PGD-20	PGD-50	C&W	AA
WideResNet34-10	FGSM-MEP	最好	48.35	25.11	24.75	23.89	19.52	19.12
		最后	**72.63**	19.15	14.09	10.73	12.47	2.69
	本文方法	最好	**67.15**	**29.77**	**27.62**	**27.36**	**24.15**	**24.41**
		最后	67.07	**29.16**	**27.11**	**27.04**	**24.61**	**24.29**

注：加粗字体为每列最优两个值。

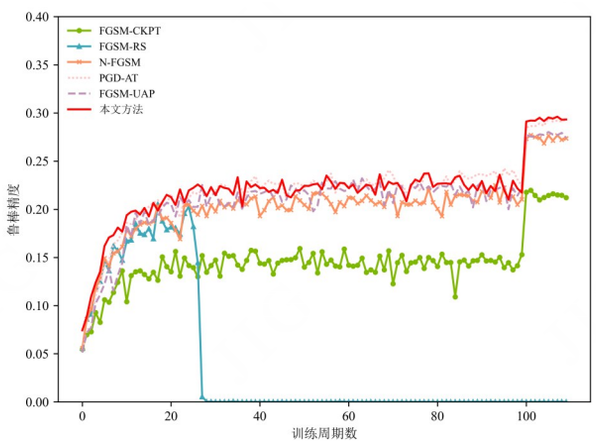


图5　ResNet18框架下训练CIFAR-100数据集时不同对抗训练方法结果对比

Fig. 5　Comparison of different adversarial training methods in training ResNet18 on CIFAR-100

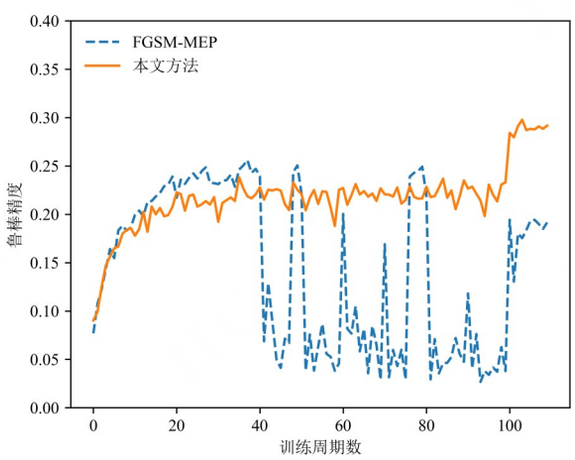


图6　WideResNet34-10框架下训练CIFAR-100数据集时FGSM-MEP与本文方法结果对比

Fig. 6　Comparison of FGSM-MEP and the proposed method in training WideResNet34-10 on CIFAR-100

表5　Tiny ImageNet上训练ResNet-18的各方法测试结果

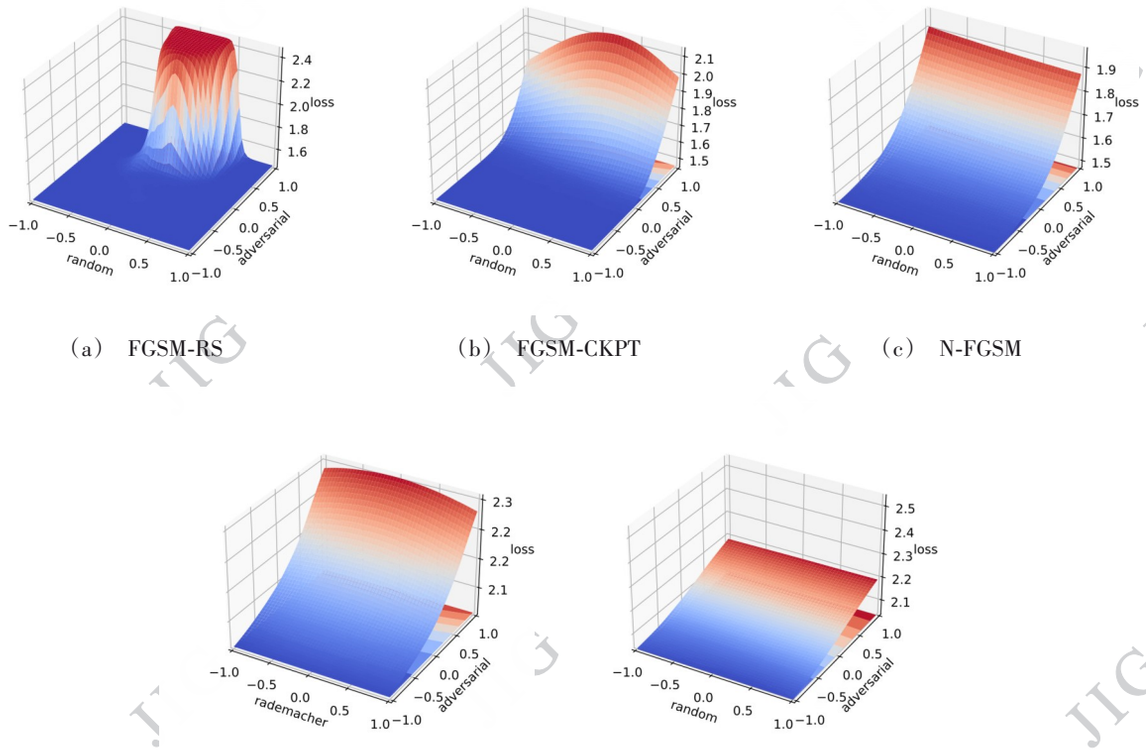
Tab. 5　Comparative test results of different methods on ResNet-18 using Tiny ImageNet

目标网络	方法	干净精度%	鲁棒精度%		训练时长h
			PGD-50	AA	
ResNet-18	PGD-AT	58.15	20.73	17.10	10.85
	FGSM-RS	45.24	0.00	0.00	2.12
	FGSM-CKPT	50.95	16.18	12.57	2.66
	N-FGSM	44.98	19.21	16.01	2.12
	FGSM-UAP	45.83	18.75	16.29	3.86
	本文方法	48.63	19.59	16.53	2.83

与鲁棒精度的平衡，后续实验统一采用γ=12作为默认设置。

4　结　论

为了解决快速对抗训练中常见的灾难性过拟合问题，同时提升模型鲁棒性与训练效率。本文提出了一种先验结构引导的快速对抗训练方法，利用结构二值化掩膜分解对抗扰动并进行随机混合，增强对抗样本的多样性和针对性，提高生成对抗样本的质量，利用批次间的对抗扰动传递，将上一批次的对抗扰动作为当前批次的扰动初始化，平滑了扰动生成过程，增强了训练的稳定性。在CIFAR-10和CIFAR-100以及Tiny ImageNet数据集上的综合评估表明，所提出的方法解决了对抗训练过程中的灾难



（a） FGSM-RS （b） FGSM-CKPT （c） N-FGSM

（d） FGSM-UAP （e） 本文方法

图7 不同快速对抗训练方法的损失景观可视化

Fig. 7 Loss landscape visualization for different fast adversarial training methods

表6 正则化权重γ对模型性能的影响

Tab. 6 Effect of the regularization weight γ on model performance

γ	干净精度%	鲁棒精度%
4	85.45	52.98
8	84.93	53.83
12	84.49	54.79
16	84.07	54.12

性过拟合问题，在面对PGD、C&W和AA等强对抗攻击时，使得模型能够保持高鲁棒性，在干净精度和鲁棒精度权衡问题上取得了更佳的平衡，与多步对抗训练方法相比，在鲁棒精度相当的情况下，对抗训练的效率也获得有效的提升。

尽管本文方法在提高模型鲁棒性和泛化能力方面取得了显著成果，但也存在一定的局限性，与FGSM-RS训练方法相比，本文方法在时间上并未占据优势，为了进一步提升本文方法的效率，未来的研究工作可以重点关注如何优化时间消耗，探索更多的效率提高策略。

参考文献(References)

Abdukhamidov E, Abuhamad M, Juraev F, Chan-Tin E and AbuHmed T. 2021. AdvEdge: optimizing adversarial perturbations against interpretable deep learning//Computational Data and Social Networks. Cham: Springer: 93 - 105 [DOI: 10.1007/978-3-030-91434-9_9]

Adachi H, Hirakawa T, Yamashita T, Fujiyoshi H, Ishii Y and Kozuka K. 2023. Masking and mixing adversarial training//Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Lisbon: SCITEPRESS: 74 - 82 [DOI: 10.5220/0011653300003417]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//European Conference on Computer Vision. Cham: Springer: 213 - 229 [DOI: 10.1007/978-3-030-58452-8_13]

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//IEEE Symposium on Security and Privacy. San

Jose: IEEE: 39 - 57 [DOI: 10.1109/SP.2017.49]

Chen Y Y, Tian D X, Lin C M and Yin H B. 2024. Survey of end-to-end autonomous driving systems (陈妍妍, 田大新, 林椿眄, 殷鸿博. 2024. 端到端自动驾驶系统研究综述). Journal of Image and Graphics, 29(11): 3216 - 3237. [DOI: 10.11834/jig.230787]

Croce F and Hein M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks//International Conference on Machine Learning. Virtual: PMLR: 2206 - 2216

De Jorge Aranda P, Bibi A, Volpi R, Sanyal A, Torr P H S, Rogez G, et al. 2022. Make some noise: reliable and efficient single-step adversarial training//Neural Information Processing Systems (NeurIPS). Virtual: NeurIPS: 12881 - 12893

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database//IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE: 248 - 255 [DOI: 10.1109/CVPR.2009.5206848]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples//International Conference on Learning Representations. San Diego: ICLR

He K, Zhang X, Ren S and Sun J. 2016. Deep residual learning for image recognition//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770 - 778 [DOI: 10.1109/CVPR.2016.90]

He K, Zhang X, Ren S and Sun J. 2016. Identity mappings in deep residual networks//European Conference on Computer Vision. Cham: Springer: 630 - 645 [DOI: 10.1007/978-3-319-46493-0_38]

Huang Z, Fan Y, Liu C, Zhang W, Zhang Y, Salzmann M, et al. 2023. Fast adversarial training with adaptive step size. IEEE Transactions on Image Processing, 32: 6102 - 6114. [DOI: 10.1109/TIP.2023.3326398]

Jia X, Zhang Y, Wei X, Wu B, Ma K, Wang J, et al. 2022. Prior-guided adversarial initialization for fast adversarial training//European Conference on Computer Vision. Cham: Springer: 567 - 584 [DOI: 10.1007/978-3-031-19772-7_33]

Jiang C, Wang J, Dong M, Gui J, Shi X, Cao Y, et al. 2025. Improving fast adversarial training via self-knowledge guidance. IEEE Transactions on Information Forensics and Security, 20: 3772 - 3787. [DOI: 10.1109/TIFS.2025.3554041]

Kim H, Lee W and Lee J. 2021. Understanding catastrophic overfitting in single-step adversarial training//AAAI Conference on Artificial Intelligence. Virtual: AAAI: 8119 - 8127 [DOI: 10.1609/aaai.v35i9.16989]

Krizhevsky A and Hinton G. 2009. Learning multiple layers of features from tiny images. Technical Report, University of Toronto

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards deep learning models resistant to adversarial attacks//International Conference on Learning Representations. Vancouver: ICLR

Mo Y, Wu D, Wang Y, Guo Y and Wang Y. 2022. When adversarial training meets vision transformers: recipes from training to architecture//Neural Information Processing Systems (NeurIPS). Virtual: NeurIPS: 18599 - 18611

Pan C, Li Q and Yao X. 2024. Adversarial initialization with universal adversarial perturbation: a new approach to fast adversarial training//AAAI Conference on Artificial Intelligence. Vancouver: AAAI: 21501 - 21509 [DOI: 10.1609/aaai.v38i19.30147]

Pan X Y, Jia N X, Mu Y Z and Gao X R. 2023. Survey of small object detection (潘晓英, 贾凝心, 穆元震, 高炫蓉. 2023. 小目标检测研究综述). Journal of Image and Graphics, 28(9): 2587 - 2615. [DOI: 10.11834/jig.220455]

Qian N. 1999. On the momentum term in gradient descent learning algorithms. Neural Networks, 12(1): 145 - 151 [DOI: 10.1016/S0893-6080(98)00116-6]

Sui C H, Wang A, Zhou S W, Zang A K, Pan Y H, Liu H, et al. 2023. A survey on adversarial training for robust learning (隋晨红, 王奥, 周圣文, 臧安康, 潘云豪, 刘颢, 等. 2023. 面向鲁棒学习的对抗训练技术综述). Journal of Image and Graphics, 28(12): 3629 - 3650. [DOI: 10.11834/jig.220953]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J, et al. 2014. Intriguing properties of neural networks//International Conference on Learning Representations. Banff: ICLR

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H. 2021. Training data-efficient image transformers and distillation through attention//International Conference on Machine Learning. Virtual: PMLR: 10347 - 10357

Wang Y, Ma X, Bailey J, Yi J, Zhou B and Gu Q. 2019. On the convergence and robustness of adversarial training//International Conference on Machine Learning. Long Beach: PMLR: 6586 - 6595

Wong E, Rice L and Kolter J Z. 2020. Fast is better than free: revisiting adversarial training//International Conference on Learning Representations. Virtual: ICLR

Zagoruyko S and Komodakis N. 2016. Wide residual networks//British Machine Vision Conference. York: BMVA Press

Zhao Z Q, Zheng P, Xu S T and Wu X. 2019. Object detection with deep learning: a review. IEEE Transactions on Neural Networks and Learning Systems, 30(11): 3212 - 3232. [DOI: 10.1109/TNNLS.2018.2876865]

作者简介

朱进东,男,硕士研究生,研究方向为对抗攻防。 E-mail: m325123102@sues.edu.cn

张玉金,通信作者,男,副教授,主要研究方向为多媒体内容安全。 E-mail:yjzhang@sues.edu.cn

张涛,男,教授,主要研究方向为多媒体内容安全。E-mail: tzhang@cslg.edu.cn

王永琦,女,副教授,主要研究方向为多媒体内容安全。 E-

mail:wangyongqi17008@163.com
吴飞,男,教授,主要研究方向为人工智能安全,计算机分布式处理技术。 E-mail:fei_wu1@163.com