计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(6): 1627−1638, 2026

基于扩散模型的分布双向引导对抗扰动净化防御

赵新宇　李　司
（公安部第一研究所　北京　100048）
（zhaoxy_ys@sina.com）

Probability Distribution Bidirectional Guidance for Adversarial Purification with Diffusion Model

Zhao Xinyu and Li Si
（*First Research Institute of the Ministry of Public Security of P.R.C*, *Beijing* 100048）

Abstract　The defense methods against adversarial perturbations holds significant importance for the secure application of recognition models. Existing adversarial defense methods typically either leverage fixed adversarial samples to augment classifier training or perform sample reconstruction prior to model input. However, in practical defense scenarios, such approaches suffer from inherent limitations, including inadequate generalizability, loss of critical sample components, and unstable data distribution. To tackle this problem, we propose probability distribution bidirectional guidance for adversarial purification with diffusion model. Specifically, we propose a novel bidirectional guidance strategy encompassing inter-class loops and intra-class aggregation. For the inter-class loop component, cross-domain mapping between adversarial and clean domains is achieved by leveraging the inverse time sequence of stochastic differential equations (SDEs), with targeted interference elimination enabled via the circular consistency principle. Subsequently, latent variable refinement is employed for distribution regularization, enhancing the integrity of noise-free components and the stability of distribution characteristics. For the intra-class aggregation component, it is exploited to strengthen the confidence of category attributes at each time step along the path from the clean domain to the adversarial domain, thereby facilitating the shift toward high-density intervals of the probability distribution. Extensive experiments conducted on benchmark datasets including CIFAR-10, CIFAR-100, and ImageNet demonstrate that our proposed method achieves state-of-the-art (SOTA) performance. Specifically, it yields a 10.9 percentage point (pp) improvement in Standard Acc and a 4.36 pp gain in Robust Acc compared with existing approaches. Furthermore, our method exhibits remarkable advantages in preserving the integrity of noise-free components and demonstrates robust stability throughout the purification process.

Key words　adversarial perturbations; defense methods; recognition models; probability distribution bidirectional guidance; diffusion model

摘　要　对抗扰动的净化防御对识别模型安全应用具有重要的研究意义。现有的防御方法通常采用固定的对抗攻击方法进行分类模型增强训练，或者对输入样本进行数据重构，但在实际防御中存在方法泛化性低、信息丢失或分布不稳定等问题。为解决上述问题，提出了基于扩散模型的分布双向引导对抗扰动净化防御方法。具体而言，设计了类间循环和类内聚合的双向引导方法，通过随机微分方程逆时序实现对抗域与干净域之间相互跨域映射，依据循环一致性思想完成干扰的有向消除，利用潜在变量增强进行分布调控，提高干净分量完整性和分布稳定性，并在干净域到对抗域的通路中引用类内聚合，提升各时刻类别属性的置信度，促使其向概率分布高密度区间偏移。实验结果表明，在数据集 CIFAR-10, CIFAR-100,

收稿日期：2025-07-01；**修回日期**：2025-12-08

ImageNet 上,所提方法与基线防御算法相比,Standard Acc 平均提升 10.9 个百分点,Robust Acc 平均提升 4.36 个百分点,在干净分量完整性保持方面表现出显著的优势,并且该方法在净化过程中也展现出较强的稳定性。

关键词 对抗扰动;净化防御;识别模型;分布双向引导;扩散模型

中图法分类号 TP311.13;TP309

DOI: 10.7544/issn1000-1239.202550515 **CSTR:** 32373.14.issn1000-1239.202550515

在当今数字化时代,图像作为重要的信息载体,广泛应用于安防监控、医疗影像诊断、自动驾驶视觉识别等诸多关键领域。然而,随着深度学习技术在图像识别任务中的深度嵌入,基于图像媒介的识别算法面临着严峻的攻击威胁。通过向图像中添加具备欺骗深度神经网络的细微扰动,诱使神经网络分类器给出错误判断,从而引起神经网络模型使用安全问题。自 Szegedy 等人[1] 发现神经网络易受对抗扰动攻击以来,对抗扰动防御技术逐渐成为当下计算机视觉领域中一项紧迫且关键的研究课题。对抗扰动防御方法旨在通过降低恶意扰动对识别模型判别干扰,确保基于图像的识别或检测模型能够稳定、准确地运行,从而为各领域的安全应用提供坚实保障。

近年来,对抗扰动防御方法研究主要集中在模型鲁棒性增强[2-4] 和数据重构去扰动[5-7] 两方面。其中模型鲁棒性增强通过把对抗样本放入训练样本中,利用监督学习机制使模型中神经感知层实现对抗扰动的抑制。虽然该种方法降低了目标分类器对扰动的敏感性,一定程度提高模型自身鲁棒性,但是对于未知的对抗扰动仍然不具备免疫能力。而基于数据重构去扰动的方式是在目标分类器预测之前,对样本进行像素重构,实现对抗干扰消除。该种防御方法直接作用于样本数据,不受对抗攻击类型的限制,具备强大的泛化性,对各种对抗攻击都具有较好的防御性,但是样本在重构过程中容易引入新的不确定信息与丢失干净样本细节信息,导致去扰动结果在分类预测存在较大的置信度损失,甚至重构结果分布不稳定。

综上,针对图像对抗防御中方法泛化性低、干净分量缺失及结果分布不稳定等问题,本文提出了一种双向引导对抗扰动净化防御方法,在扩散重构过程中加入类间循环引导与类内聚合引导,通过信息定向生成实现对抗扰动消除。类间循环引导在扩散模型逆向过程中实现对抗域与干净域之间相互跨域映射,依据循环一致性思想构建自监督式干扰消除框架,利用潜在变量增强对逆时序中各时刻生成状态进行分布调控,旨在提高干净分量完整度与净化结果分布的稳定性。类内聚合引导在干净域到对抗域生成环节使用分类器置信度最大化方式促使各时刻状态向概率分布的高密度区间偏移。本文主要贡献有 3 点:

1)提出了一种基于扩散模型的分布双向引导对抗扰动净化防御方法,结合循环一致性思想实现了对抗样本的无监督净化,有效解决了防御中净化方向不明确的问题,提高了净化鲁棒性。

2)提出了分布调控方法,依靠潜在变量增强对逆向还原过程进行修正,提高干净域内干净分量保持能力与净化结果分布的稳定性,避免逆向还原中内容生成失控现象。

3)通过实验表明,与基线防御方法相比,本文方法在 Standard Acc 与 Robust Acc 上具有明显的提升,并且净化结果中干净分量的完整性与分布的稳定性也表现出显著优势。

1 相关工作

1.1 模型鲁棒性增强

模型鲁棒性增强旨在利用对抗样本对被攻击模型进行对抗训练。Garmon 等人[2] 分析对抗样本扰动的有界性,提出有限数据集鲁棒性增强的训练方法。Dong 等人[3] 针对传统模型鲁棒性增强训练中干净样本置信度变低的潜在缺陷,提出逆对抗训练(inverse adversarial training, IAT)方法。Naman 等人[4] 提出使用 ConvSetm 代替 PatchStem,以提高模型在基于范数对抗攻击下的稳定性。模型鲁棒性增强方法虽然通过降低被攻击模型中神经感知单元对扰动的敏感性来提高抵御对抗扰动的能力,但是对于未知的对抗扰动仍然不具备免疫能力。因此受训练中对抗样本类型制约,该类型防御方法泛化性有限。

1.2 数据重构去扰动

数据重构去扰动的方式旨在使用生成模型对对抗样本进行编辑重构,剔除对抗样本中的扰动纹理,

破坏对抗扰动的连续性与迁移性，生成具备原始特性的干净样本。He 等人[5]提出了基于矢量量化 U-Net（vector quantization U-Net, VQUNet）的对抗攻击防御方法，通过将矢量量化技术融入 U-Net 架构，在低维矢量空间实现对样本中对抗干扰的约束与调整。Ren 等人[6]提出可学习主成分分析（learnable principal component analysis, learnable PCA）估计免疫空间的方法，利用去噪自编码器与特征向量选择网络实现低维空间干扰消除。Samangouei 等人[7]通过训练生成对抗网络（generative adversarial network, GAN）学习正常数据分布，利用其数据修复与净化机制处理对抗样本。上述基于编解码生成模型的数据重构去扰动方法在潜在空间特征还原过程中实现了干扰分量抑制，该种方法直接作用于样本数据，不受对抗攻击类型的限制，具备强大的泛化性，对各种对抗攻击都具有较好的防御性，但受编解码模型固有特性限制，基于编解码的生成方式在编码过程损失太多信息，解码中难以重新创建高质量样本，并在生成过程中容易引入新的不确定信息，导致结果在分类预测存在较大的置信度损失，甚至出现类别分布的迁移。

随着生成式模型发展，基于概率空间变换思想的模型得到的广泛的研究和发展，其中扩散模型（diffusion model, DM）[8-10]在多种视觉处理领域表现出强大的潜力，并且在对抗扰动消除方面已经取得较多成果[11-14]。利用扩散模型强大的分布拟合能力，在逆向还原中逐步剔除样本中扰动分量，破坏对抗扰动语义关联性，钝化扰动对识别模型的干扰能力，以此消除对抗扰动在类别空间的干扰。Nie 等人[11]应用扩散模型进行对抗防御，并证明扩散模型在对抗扰动消除方面的优越性。Zhang 等人[13]提出了分类器置信度引导方法，通过加入类别置信度最大化引导，阻止净化过程中样本分布向类别边界扩散，实现了类别置信度的提高。Lin 等人[14]提出了通过损失对抗实现净化中准确性与鲁棒性的平衡。尽管基于扩散模型的净化方法在对抗扰动防御任务中取得较好的成果，但是由于扩散模型生成过程不可控，净化后结果出现细节缺失，且随着的时间步不断推移，净化结果无法稳定在干净样本分布内。

综上所述，针对对抗扰动防御任务，现有的鲁棒性增强防御方法存在泛化性差的问题，防御性能往往只能针对单一对抗攻击方法。数据重构去扰动方法中，依靠编解码架构的生成模型，对抗防御处理后的结果容易引入不确定信息且分布置信度损失较大；而基于扩散模型的净化防御方法，虽然具备较强的扰动消除能力与分布拟合能力，但是现有方法均表现出生成结果细节缺失与分布不稳定的现象。因此，本文为提高防御的泛化性与分布迁移稳定性，提出一种基于扩散模型的分布双向引导对抗扰动净化防御方法，在扩散模型强大分布拟合能力基础上，设计了类间循环引导与类内聚合引导，增加了扩散模型逆时序过程中分布域间迁移规则，不仅明确了净化中类别迁移方向，同时也提高了迁移后的域内稳定性，增强了生成结果细节保持能力。

2 本文方法

本文提出一种基于扩散模型的分布双向引导对抗扰动净化防御，具体过程如图 1 所示，对抗样本通

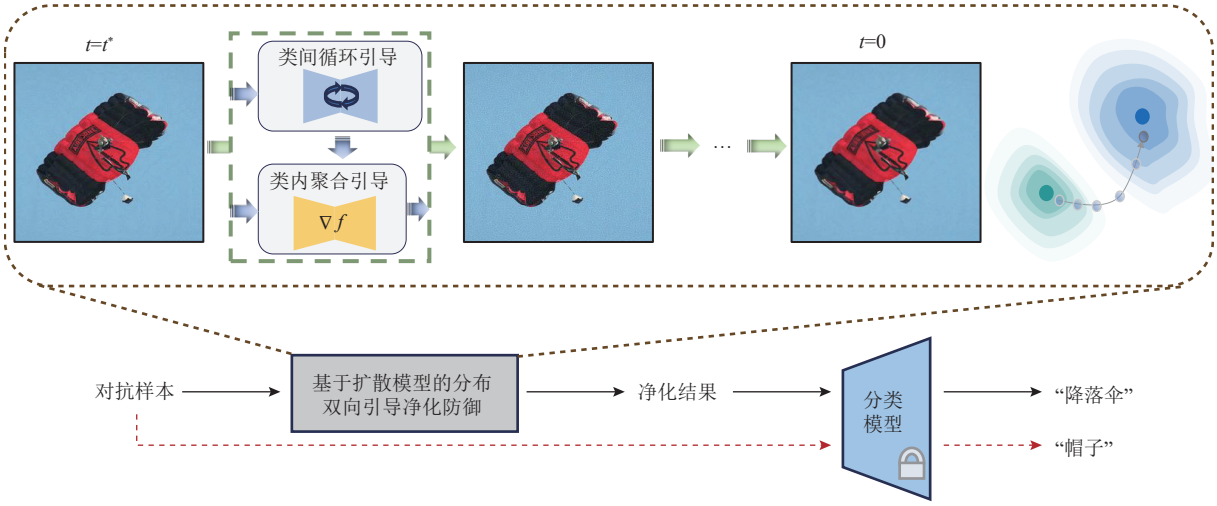


Fig. 1 Probability distributed bidirectional guidance for adversarial purification framework

图 1 分布双向引导对抗扰动净化防御框架

过基于扩散模型的分布双向引导净化方法实现样本中对抗干扰分量的消除。基于扩散模型的分布双向引导净化方法是在扩散还原过程中，对每一时刻重构分量进行类间循环引导，完成当前时刻扰动的消除，并对当前时刻扰动消除分量进行类内聚合引导，实现概率分布向高密度区间偏移。类间循环引导中，通过分数随机方程逆时序进行对抗域与干净域之间的循环回归，实现扰动净化方向的自监督引导，并利用分布调控调整逆时序各时刻状态的分布，以便增强干净分量保持能力与净化结果分布的稳定性；类内聚合引导中，在干净域到对抗域生成过程中，通过最大化类别置信度使净化结果保持在高概率密度空间，避免低概率密度区间类别不稳定现象。

2.1 分数随机微分方程扩散模型

分数随机微分方程扩散模型[8]是在传统扩散模型的基础上，利用随机微分方程对扩散过程进行建模的一种方法。本文采用 Score SDE 逆时序作为对抗样本重构的基础操作，通过重构方式提高防御方法在对抗攻击中的泛化性。利用 Score SDE 对样本进行跨域变换，其中分数函数用于描述数据分布的局部梯度信息，通过将分数函数与随机微分方程相结合，构建出样本的分布空间控制模型。

在 Score SDE 正时序过程，其前向扩散过程 $\{\boldsymbol{x}_t\}_{t\in[0,t^*]}$ 相邻时刻样本变化增量定义为

$$\mathrm{d}\boldsymbol{x}=f(\boldsymbol{x}_t,t)\mathrm{d}t+g(t)\mathrm{d}\boldsymbol{\omega}\,, \tag{1}$$

其中时间 $t^*\in(0,1]$ 且 $\boldsymbol{x}_{t^*}=\boldsymbol{x}_{\mathrm{adv}}$，$\boldsymbol{x}_{\mathrm{adv}}$ 为对抗样本。$f(\cdot,\cdot)$ 为漂移系数，$g(\cdot)$ 为扩散系数，$\boldsymbol{\omega}$ 为正时序标准布朗运动，$\mathrm{d}\boldsymbol{\omega}$ 可视为服从无穷小正态分布。在 SDE 框架中，$f(\cdot,\cdot)$ 与 $g(\cdot)$ 分别表示为 $-\frac{1}{2}\beta(t)\boldsymbol{x}_t$ 与 $\sqrt{\beta(t)}$，$\beta(\cdot)$ 为时间噪声系数函数且 $\beta(t)=(\beta_{\max}-\beta_{\min})t+\beta_{\min}$。在 Score SDE 逆时序过程，其逆向还原过程 $\{\boldsymbol{x}_t\}_{t\in[t^*,0]}$ 相邻时刻样本变化增量定义为

$$\mathrm{d}\boldsymbol{x}=[f(\boldsymbol{x}_t,t)-g(t)^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})]\mathrm{d}t+g(t)\mathrm{d}\bar{\boldsymbol{\omega}}\,, \tag{2}$$

其中时间 $t\in[t^*,0]$，$p_t(\boldsymbol{x})$ 为时刻 t 的 $\boldsymbol{x}_t$ 边缘分布。$\bar{\boldsymbol{\omega}}(t)$ 为逆时序布朗运动。SDE 逆时序 $\{\boldsymbol{x}_t\}_{t\in[t^*,0]}$ 与 SDE 正时序 $\{\boldsymbol{x}_t\}_{t\in[0,t^*]}$ 对应时间下的 $\boldsymbol{x}_t$ 一致。$\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})$ 为分数函数，反映了 $p_t(\boldsymbol{x})$ 在样本分布空间中各个方向上的变化率，因此，通过调整 $\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})$ 可实现 SDE 逆时序干扰消除的方向控制。

2.2 类间循环引导

2.2.1 循环回归

在基于 Score SDE 的净化方法中，一般采用深度神经网络 $s_\theta(\boldsymbol{x},t)$ 预估 $\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x})$，在重构生成过程中破坏对抗纹理内在语义关联实现样本的净化。这种单纯依靠干扰破坏实现对抗扰动消除的方式，由于重构方向不明确，造成重构结果逐步脱离期望分布。为明确净化方向，在 SDE 逆向还原中，本文设计循环回归方式，依靠循环一致性思想实现重构方向引导，通过构建对抗域与干净域之间的循环回归完成对抗样本的自引导净化，如图 2 所示。

对于 Score SDE 生成空间，假设对抗域到干净域 SDE 逆时序过程为 $\{\boldsymbol{x}_t^{\mathrm{pos}}\}_{t\in[t^*,0]}$，干净域到对抗域 SDE 逆时序过程[15]为 $\{\boldsymbol{x}_t^{\mathrm{neg}}\}_{t\in[t^*,0]}$，且 $\boldsymbol{x}_t^{\mathrm{pos}}=\boldsymbol{x}_{t^*-t}^{\mathrm{neg}}$，初始状态满足 $\boldsymbol{x}_{t^*}^{\mathrm{pos}}=\boldsymbol{x}_0^{\mathrm{neg}}$。首先在对抗域到干净域变换中，对带噪 $\boldsymbol{x}_t$ 进行 SDE 逆时序净化变换得到 $\boldsymbol{x}_{t'}^{\mathrm{pos}}$，SDE 逆时序净化变换可表示为

$$\boldsymbol{x}_{t'}^{\mathrm{pos}}=\boldsymbol{x}_t+[f(\boldsymbol{x}_t,t)-g(t)^2s_\theta(\boldsymbol{x}_t,t)]\mathrm{d}t+g(t)\mathrm{d}\bar{\boldsymbol{\omega}}\,, \tag{3}$$

其中 t' 为时刻 t 在时间序列 $[t^*,0]$ 下一时刻，且 $\boldsymbol{x}_{t^*}=\boldsymbol{x}_{\mathrm{adv}}$。然后在干净域到对抗域变换中，对干净 $\boldsymbol{x}_{t'}^{\mathrm{pos}}$ 进行 SDE 逆时序加噪变换得到 $\boldsymbol{x}_{t^*-t}^{\mathrm{neg}}$，SDE 逆时序加噪变换表示为

$$\boldsymbol{x}_{t^*-t}^{\mathrm{neg}}=\boldsymbol{x}_{t'}^{\mathrm{pos}}+[f(\boldsymbol{x}_{t'}^{\mathrm{pos}},t^*-t')-g(t^*-t')^2s_\theta(\boldsymbol{x}_{t'}^{\mathrm{pos}},t^*-t')]\mathrm{d}t+g(t^*-t')\mathrm{d}\bar{\boldsymbol{\omega}}\,。 \tag{4}$$

固定 $\boldsymbol{x}_{t'}^{\mathrm{pos}}$，对式(4)利用分类模型修正分数函数 $s_\theta(\boldsymbol{x}_{t'}^{\mathrm{pos}},t^*-t')$ 加噪方向，有

$$\nabla_{\boldsymbol{x}}\log p_{t^*-t'}(\boldsymbol{x})=s_\theta(\boldsymbol{x}_{t'}^{\mathrm{pos}},t^*-t')+\nabla_s MSE_{\mathrm{cls}}(f_{\mathrm{cls}}(\boldsymbol{x}_t),f_{\mathrm{cls}}(\boldsymbol{x}_{t^*-t}^{\mathrm{neg}}))\,, \tag{5}$$

其中 $f_{\mathrm{cls}}(\cdot)$ 为分类模型，$\nabla_s MSE_{\mathrm{cls}}(f_{\mathrm{cls}}(\boldsymbol{x}_t),f_{\mathrm{cls}}(\boldsymbol{x}_{t^*-t}^{\mathrm{neg}}))$ 表示 $\boldsymbol{x}_t$ 与 $\boldsymbol{x}_{t^*-t}^{\mathrm{neg}}$ 表征特征均方误差关于分布空间变化率 $s_\theta(\boldsymbol{x}_{t'}^{\mathrm{pos}},t^*-t')$ 的梯度。通过分数函数 $\nabla_{\boldsymbol{x}}\log p_{t^*-t'}(\boldsymbol{x})$ 更

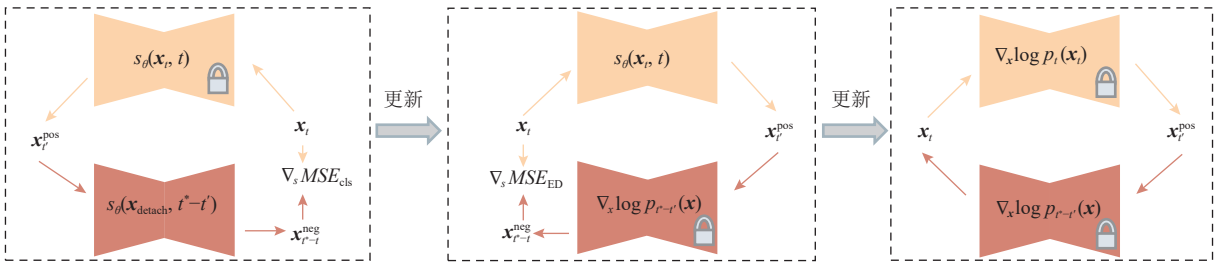


Fig. 2 The cyclic regression

图 2 循环回归

新$\boldsymbol{x}_{t^*-t}^{\text{neg}}$，有

$$\boldsymbol{x}_{t^*-t}^{\text{neg}}=\boldsymbol{x}_{t'}^{\text{pos}}+[f(\boldsymbol{x}_{t'}^{\text{pos}},t^*-t')-g(t^*-t')^2\nabla_{\boldsymbol{x}}\log p_{t^*-t'}(\boldsymbol{x})]\mathrm{d}t+g(t^*-t')\mathrm{d}\bar{\boldsymbol{\omega}}\text{。}\tag{6}$$

最后，固定$\boldsymbol{x}_{t^*-t}^{\text{neg}}$，根据$\boldsymbol{x}_t$与$\boldsymbol{x}_{t'}^{\text{neg}}$均方误差更新分数函数$s_\theta(\boldsymbol{x}_t,t)$净化方向，则有

$$\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)=s_\theta(\boldsymbol{x}_t,t)+\nabla_s MSE_{\text{ED}}(\boldsymbol{x}_t,\boldsymbol{x}_{t'}^{\text{neg}})\text{。}\tag{7}$$

通过分数函数$\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)$更新$\boldsymbol{x}_{t'}^{\text{pos}}$，有

$$\boldsymbol{x}_{t'}^{\text{pos}}=\boldsymbol{x}_t+[f(\boldsymbol{x}_t,t)-g(t)^2\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)]\mathrm{d}t+g(t)\mathrm{d}\bar{\boldsymbol{\omega}}\text{。}\tag{8}$$

循环回归中，通过固定$\boldsymbol{x}_{t'}^{\text{pos}}$与$\boldsymbol{x}_{t^*-t}^{\text{neg}}$实现对加噪变换与净化变化中分数函数的梯度控制。

2.2.2 分布调控

循环回归虽然通过干净域与对抗域间循环变换实现了净化的自监督，但重构方式采用无监督的扩散模型，净化结果可能出现干净分量的丢失或畸变，甚至结果分布的不稳定，因此本文引入潜在变量增强方法[10]进行调控，增加无条件下净化过程的动态约束，提高净化后干净分量的完整性与准确性。

净化任务中对抗分量属于高频信息，理想情况中SDE逆时序$\{\boldsymbol{x}_t\}_{t\in[t^*,0]}$各时刻$\boldsymbol{x}_t$包含的干净分量是相同的，且干净分量相对于对抗分量属于低频信息，因此满足$\phi_N(\boldsymbol{x}_t)=\phi_N(\boldsymbol{x}_{t'})$，$\phi_N(\cdot)$表示按$N$倍进行下采样和上采样的线性低通滤波器。如图3所示，分布调控利用干净分量偏差去修正当前状态，以此减小扩散模型逆向还原中分布的逐步偏离，保障了干净分量完整性与净化结果分布的稳定性。分布调控可表示为

$$\boldsymbol{x}_{t'}^{N}=\boldsymbol{x}_{t'}^{\text{pos}}+\phi_N(\boldsymbol{x}_t)-\phi_N(\boldsymbol{x}_{t'}^{\text{pos}})\text{。}\tag{9}$$

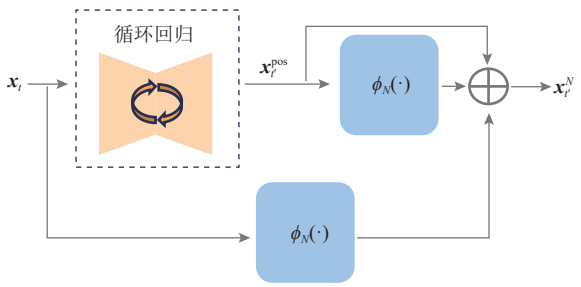


Fig. 3 The distribution regulation

图3 分布调控

2.3 类内聚合引导

依靠类间循环引导进行扰动消除可使样本在类别空间中向干净域靠近，但受生成噪声及净化能力限制，净化后的结果可能无法移动到干净域高概率密度区域。为使净化后的样本远离分类器决策边界和提高样本类别置信度，本文引入最大化分类器置信度引导[13]思想，在干净域到对抗域变换中，通过最大化置信度提高类内聚合。

如图4所示，假设$\boldsymbol{x}_t$通过类间循环引导得到$\boldsymbol{x}_{t'}^{N}$，在干净域到对抗域变换中，$\boldsymbol{x}_{t'}^{N}$通过最大化分类器置信度进行类内迁移，有

$$\boldsymbol{x}_{t'}=\boldsymbol{x}_{t'}^{N}+\lambda\nabla_{\boldsymbol{x}_{t'}^{N}}MSE(\boldsymbol{x}_t,\boldsymbol{x}_t^{\text{neg}N})\text{，}\tag{10}$$

其中λ为置信度调整参数，用于控制$\boldsymbol{x}_{t'}^{N}$的类内迁移幅度，本文选取$\lambda=0.001$。$\boldsymbol{x}_t^{\text{neg}N}$是$\boldsymbol{x}_t$的预估量，代表干净域到对抗域变化中$\boldsymbol{x}_{t'}^{N}$的下一时刻逆时序生成量，可表示为

$$\boldsymbol{x}_t^{\text{neg}N}=\boldsymbol{x}_{t'}^{N}+[f(\boldsymbol{x}_{t'}^{N},t^*-t')-g(t^*-t')^2\nabla_{\boldsymbol{x}^N}\log p_{t^*-t'}(\boldsymbol{x})]\mathrm{d}t+g(t^*-t')\mathrm{d}\bar{\boldsymbol{\omega}}\text{，}\tag{11}$$

$$\nabla_{\boldsymbol{x}^N}\log p_{t^*-t'}(\boldsymbol{x})=s_\theta(\boldsymbol{x}_{t'}^{N},t^*-t')+\nabla_{\boldsymbol{x}}\log\max f_{\text{conf}}(\boldsymbol{x}_{t'}^{N})\text{，}\tag{12}$$

其中$\max f_{\text{conf}}(\cdot)$为分类模型类别最大置信度。通过利用$\boldsymbol{x}_t$和$\boldsymbol{x}_t^{\text{neg}N}$均方差在$\boldsymbol{x}_{t'}^{N}$上的梯度$\nabla_{\boldsymbol{x}_{t'}^{N}}MSE(\boldsymbol{x}_t,\boldsymbol{x}_t^{\text{neg}N})$进行$\boldsymbol{x}_{t'}^{N}$的偏差调节，弥补净化中的置信度损失。

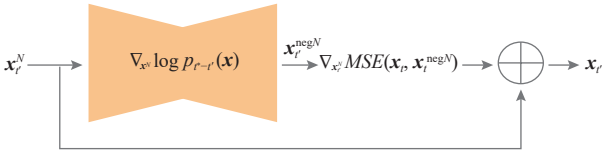


Fig. 4 The intra-class aggregation

图4 类内聚合

2.4 分布双向引导净化防御

本文在Score SDE生成模型基础上，构建了一个基于对抗域与干净域双向映射的自回归干扰消除方法。对抗样本$\boldsymbol{x}_{\text{adv}}$在净化通路中，利用类间循环引导对各时刻状态$\boldsymbol{x}_t$进行对抗干扰消除，在分布回迁的同时提高了干净分量完整性与分布稳定性，并在各时刻相应类别空间内进行类内聚合来引导操作，推动样本分布向高概率密度区域转移。本文所提基于扩散模型的分布双向引导对抗扰动净化防御算法流程如算法1所示。

算法1. 基于扩散模型的分布双向引导对抗扰动净化防御。

输入：初始时间步长t^*，时间步长度Δt，预训练Score SDE模型$s_\theta(\boldsymbol{x},t)$，对抗样本$\boldsymbol{x}_{\text{adv}}$；

输出：净化结果$\boldsymbol{x}$。

① 初始化$\boldsymbol{x}_t\leftarrow\boldsymbol{x}_{\text{adv}}$，$t\leftarrow t^*$；

② while $t\geqslant 0$ do

③ $\quad\boldsymbol{x}_{t'}^{\text{pos}}\leftarrow$式(3)；/*逆时序净化变换*/

④ $\quad$固定$\boldsymbol{x}_{t'}^{\text{pos}}$；/*截断反向传播梯度*/

⑤ $\quad\boldsymbol{x}_{t^*-t}^{\text{neg}}\leftarrow$式(4)；/*逆时序加噪变换*/

⑥ $\nabla_x \log p_{t^*-t'}(\boldsymbol{x})$ ←式(5); /*更新加噪变换分数函数*/
⑦ 根据式(6)更新 $\boldsymbol{x}_{t^*-t}^{\text{neg}}$;
⑧ 固定 $\boldsymbol{x}_{t^*-t}^{\text{neg}}$; /*截断反向传播梯度*/
⑨ $\nabla_x \log p_t(\boldsymbol{x})$ ←式(7); /*更新净化变换分数函数*/
⑩ 根据式(8)更新 $\boldsymbol{x}_{t'}^{\text{pos}}$;
⑪ $\boldsymbol{x}_{t'}^{N}$ ←式(9); /*分布调控*/
⑫ $\nabla_{x^N} \log p_{t^*-t'}(\boldsymbol{x})$ ←式(12);
⑬ $\boldsymbol{x}_t^{\text{neg}N}$ ←式(11); /*类内聚合*/
⑭ $\boldsymbol{x}_t$ ←式(10);
⑮ if $t-\Delta t \leqslant 0$ then
⑯ $t \leftarrow \Delta t$;
⑰ else
⑱ $t \leftarrow t-\Delta t$;
⑲ end if
⑳ end while
㉑ return $\boldsymbol{x}_t$。

3 实验及结果分析

3.1 实验环境

为验证对抗防御算法净化效果,本文实验采用 CIFAR-10, CIFAR-100, ImageNet 这 3 种数据集。对于实验中的分类模型,数据集 CIFAR-10 和 CIFAR-100 选用 WideResNet-28-10, WideResNet-70-16 这 2 种框架的分类预训练模型, ImageNet 选用 ResNet50 框架预训练模型。在对比实验中,本文对抗攻击方法采用标准 AutoAttack 作为攻击基准进行算法评估,实验中利用标准 AutoAttack 的 l_2 与 l_∞ 攻击分别对干净样本添加扰动,其中标准 AutoAttack 为 APGD-ce, APGD-t, FAB-t 这 3 种白盒攻击和 Square 黑盒攻击形成的组合攻击。为了与其他对抗净化方法进行公平比较,实验还使用了自适应攻击方法 BPDA+EOT,这种攻击方法专门针对净化防御方法而设计。考虑了模型在面对不同攻击策略时的适应性, BPDA+EOT 通过引入可微近似函数及随机变换平均梯度策略,提高了离散梯度破碎和随机梯度扰动处理能力,形成多种防御机制的全面自适应攻击能力。因此 BPDA+EOT 攻击方法可以更真实地模拟实际应用中模型可能遇到的攻击情况,使评估结果更具说服力。

由于自适应攻击及净化算法的计算成本高,因此实验仅从数据集中随机抽取的 512 个样本作为固定子集进行净化方法性能评估,并且有研究证明大多数基线的精度在不同测试子集上没有太大变化[11]。

算法防御性能评估指标选用标准准确率(Standard Acc)与鲁棒准确率(Robust Acc)进行性能衡量。Standard Acc 是衡量防御方法作用于干净样本后保持样本原特征的性能, Robust Acc 是衡量防御方法对对抗样本净化后保持干净样本特征的性能。样本干净分量完整性评价指标选用峰值信噪比(peak signal to noise ratio, PSNR)与结构相似度指数 SSIM(structural similarity index measure)。

本文实验采用的计算机配置为 Intel Xeon E5-2697 v3 处理器, Nvidia L40s 显卡。运行环境为 Pytorch 1.11.0。实验中当 AutoAttack 采用 l_∞ 对抗攻击时设置 $t^*=0.1$, 对抗攻击扰动阈值 ε 设置为 8/255; 当 AutoAttack 采用 l_2 对抗攻击时设置 $t^*=0.075$, 对抗攻击扰动阈值 ε 设置为 0.5; BPDA+EOT 攻击时设置 $t^*=0.1$。CIFAR-10 与 CIFAR-100 数据集下分数模型采用 VP-SDE 模型作为预训练模型, ImageNet 数据集采用 Guilded Diffusion 作为预训练模型,其中时间噪声系数 $\beta_{\min}=0.1$, $\beta_{\max}=20$。

3.2 防御性能对比实验

为了验证本文所提净化方法的有效性,实验分别在 CIFAR-10, CIFAR-100, ImageNet 固定子集中与基线防御方法进行比较。实验中选用 AutoAttack 与 BPDA+EOT 作为攻击方法,分析净化后结果在分类模型中的 Standard Acc 和 Robust Acc 表现。

3.2.1 CIFAR-10 净化性能分析

为验证本文算法净化性能,针对 CIFAR-10 数据集,分别对 WideResNet-28-10 与 WideResNet-70-16 这 2 种分类模型设计标准 AutoAttack 与 BPDA+EOT 攻击净化防御对比实验,分析不同分类模型、不同扰动模式、不同攻击方法下的净化结果。表 1 与表 2 分别是 WideResNet-28-10 与 WideResNet-70-16 分类模型针对 AutoAttack l_∞ 攻击的防御性能表现,表 3 与表 4 分别是 WideResNet-28-10 与 WideResNet-70-16 分类模型针对 AutoAttack l_2 攻击防御的性能表现,表 5 是 WideResNet-28-10 模型针对 BPDA+EOT l_∞ 攻击的防御性能表现。

1) AutoAttack l_∞ 防御性能

实验结果如表 1 与表 2 所示,在 CIFAR-10 数据集中,本文方法在 AutoAttack l_∞ 攻击下展现出显著优势。对于 WideResNet-28-10 分类模型,本文方法 Standard Acc 为 94.92%,表明对干净样本的原始特征保留完整,较 Peng 等人[16]的方法(93.27%)提升了 1.65 个百分点,证明本文防御方法引入额外特征损

Table 1 Defense Performance of WideResNet-28-10 Against Attack of AutoAttack l_∞ on CIFAR-10 Dataset

表 1 WideResNet-28-10 对 CIFAR-10 数据集中 AutoAttack l_∞ 攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Gowal 等人[18]	2021	88.74	66.11
Nie 等人[11]	2022	89.02	70.64
Wang 等人[12]	2022	84.85	71.18
Lin 等人[19]	2022	90.62	72.85
Wang 等人[20]	2023	93.25	70.69
Peng 等人[16]	2023	93.27	71.07
Bi 等人[21]	2024	91.81	68.60
Lin 等人[17]	2024	90.82	78.12
Zhang 等人[13]	2024	90.04	73.05
本文		94.92	78.91

Table 2 Defense Performance of WideResNet-70-16 Against Attack of AutoAttack l_∞ on CIFAR-10 Dataset

表 2 WideResNet-70-16 对 CIFAR-10 数据集中 AutoAttack l_∞ 攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Gowal 等人[22]	2020	85.29	59.57
Rebuffi 等人[23]	2021	88.54	64.46
Gowal 等人[18]	2021	88.74	66.60
Nie 等人[11]	2022	91.04	71.84
本文		95.12	79.10

Table 3 Defense Performance of WideResNet-28-10 Against Attack of AutoAttack l_2 on CIFAR-10 Dataset

表 3 WideResNet-28-10 对 CIFAR-10 数据集中 AutoAttack l_2 攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Rebuffi 等人[23]	2021	91.79	78.32
Nie 等人[11]	2022	91.38	78.98
Wang 等人[12]	2022	95.16	83.68
Zhang 等人[13]	2024	92.58	83.13
Bai 等人[24]	2024	93.75	84.38
Lin 等人[17]	2024	90.82	86.84
本文		97.07	87.50

失较小。本文方法 Robust Acc 为 78.91%，较 Lin 等人[17]的方法(78.12%)提升 0.79 个百分点，说明通过双向引导机制，对抗样本的扰动清除效率显著提高，并且没有过度净化干净样本特征。在更深的 WideResNet-70-16 模型上，本文 Standard Acc 进一步提升至 95.12%，Robust Acc 提升至 79.10%，验证了本文净化方法对复杂架构的适应性，表明模型深度与本文引导净化策略结合后，能更高效地分离干净特征与扰动，实现干净特征保留与强扰动净化的双重优化。

Table 4 Defense Performance of WideResNet-70-16 Against Attack of AutoAttack l_2 on CIFAR-10 Dataset

表 4 WideResNet-70-16 对 CIFAR-10 数据集中 AutoAttack l_2 攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Gowal 等人[22]	2020	90.90	74.03
Rebuffi 等人[23]	2021	92.41	80.86
Nie 等人[11]	2022	93.24	81.17
本文		97.46	83.00

Table 5 Defense Performance of WideResNet-28-10 Against Attack of BPDA+EOT l_∞ on CIFAR-10 Dataset

表 5 WideResNet-28-10 对 CIFAR-10 数据集中 BPDA+EOT l_∞ 攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Song 等人[25]	2018	95.00	9.00
Yang 等人[26]	2019	94.00	15.00
Hill 等人[27]	2021	84.12	54.90
Yoon 等人[28]	2021	86.14	70.01
Wang 等人[12]	2022	84.85	76.22
Nie 等人[11]	2022	89.02	81.40
Zhang 等人[13]	2024	90.04	83.20
本文		94.92	85.74

2)不同扰动模式的鲁棒性分析

实验结果如表 3 与表 4 所示，针对 AutoAttack l_2 攻击，本文方法在 WideResNet-28-10 分类模型上的 Robust Acc 为 87.50%，超越 Lin 等人[17]的方法(86.84%)，且 Standard Acc 保持在 97.07%，体现了准确率与鲁棒性的平衡优势。结合表 1 结果分析可知，本文方法对不同扰动模式均有良好净化能力，在 l_2 全局平滑扰动与 l_∞ 局部剧烈扰动间具有较好的泛化性。在 WideResNet-70-16 模型中 Robust Acc 达 83%，较 Nie 等人[11]的方法(81.17%)提升 1.83 个百分点，结合表 2 证明本文净化方法在复杂分类网络中对不同扰动模式具有较好的鲁棒性。

3)BPDA+EOT 强自适应攻击防御

BPDA+EOT 强自适应攻击具有自适应和随机性,并且是专门针对基于扩散模型净化方法设计的攻击方式,因此该攻击方法是评估防御方法泛化能力的关键指标[11]。实验结果如表 5 所示,在 BPDA+EOT 攻击防御中,本文方法的 Robust Acc 为 85.74%,远超 Zhang 等人[13]的方法(83.20%),这表明通过双向引导的净化方式能有效抵御依赖梯度近似的攻击,凸显了本文方法在强对抗场景下的稳定性与泛化性。

3.2.2 CIFAR-100 多类别净化鲁棒性验证

在 CIFAR-100 数据集上,语义类别复杂度显著增加,设计 WideResNet-28-10 与 WideResNet-70-16 这 2 种多类别分类模型在标准 AutoAttack l_∞对抗攻击下的净化实验。如表 6 与表 7 所示,本文方法在 WideResNet-28-10 模型上 Robust Acc 为 53.13%,较 Lin 等人[17]的方法(46.09%)提升 7.04 个百分点,且 Standard Acc 为 94.33%,显著高于 Lin 等人[17]的方法(69.73%)。对于 WideResNet-70-16 模型,本文方法的 Robust Acc 达 55.08%,高于 Wang 等人[12]的方法(42.67%),方法的 Standard Acc 为 94.72%,同样优于其他基线方法。此实验表明本文方法在多种类别净化中具有较强的鲁棒性,在语义多样场景下仍能保持较好的净化能力与特征保持能力。

Table 6 Defense Performance of WideResNet-28-10 Against Attack of AutoAttack l_∞ on CIFAR-100 Dataset

表 6 WideResNet-28-10 对 CIFAR-100 数据集中 AutoAttack l_∞攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Rebuffi 等人[23]	2021	62.41	32.06
Pang 等人[29]	2022	63.66	31.08
Wang 等人[12]	2022	72.58	38.83
Cui 等人[30]	2024	65.93	32.52
Lin 等人[17]	2024	69.73	46.09
Bi 等人[21]	2024	64.88	34.58
本文		94.33	53.13

Table 7 Defense Performance of WideResNet-70-16 Against Attack of AutoAttack l_∞ on CIFAR-100 Dataset

表 7 WideResNet-70-16 对 CIFAR-100 数据集中 AutoAttack l_∞攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Rebuffi 等人[23]	2021	63.56	34.64
Pang 等人[29]	2022	65.56	33.05
Wang 等人[12]	2022	75.22	42.67
本文		94.72	55.08

3.2.3 ImageNet 高分辨率图像净化

为了验证净化方法对高分辨率图像的净化效果,在 ImageNet 数据集上,设计 ResNet50 分类模型在标准 AutoAttack l_∞对抗攻击下的净化实验。如表 8 所示,本文方法的 Robust Acc 为 65.43%,较 Bi 等人[21]方法的 63.09% 提升 2.34 个百分点。结合本文方法的 Standard Acc(94.72%),说明本文方法在高分辨率样本中仍有效保留了图像信息,对像素复杂变化具有鲁棒性,验证了本文方法净化过程对图像内容的语义保留优势明显,未因分辨率增加导致特征畸变。

Table 8 Defense Performance of ResNet50 Against Attack of AutoAttack l_∞ on ImageNet Dataset

表 8 ResNet50 对 ImageNet 数据集中 AutoAttack l_∞攻击的防御性能 %

方法	年份	Standard Acc	Robust Acc
Wong 等人[31]	2020	55.62	26.95
Salman 等人[32]	2020	64.02	37.89
Bai 等人[33]	2021	67.38	35.51
Wang 等人[12]	2022	62.56	31.06
Nie 等人[11]	2022	68.22	43.89
Bi 等人[21]	2024	61.05	63.09
本文		94.72	65.43

3.3 干净分量完整性对比实验

为评估基于扩散模型的净化方法在对抗攻击下的干净分量保留能力,设计实验对比 Nie 等人[11]和 Zhang 等人[13]的方法与本文方法在 PSNR 和 SSIM 指标上的表现,验证本文方法在保全样本干净分量方面的优越性。实验采用 AutoAttack l_∞对抗攻击对 CIFAR-10 与 ImageNet 固定子集中的样本进行攻击。

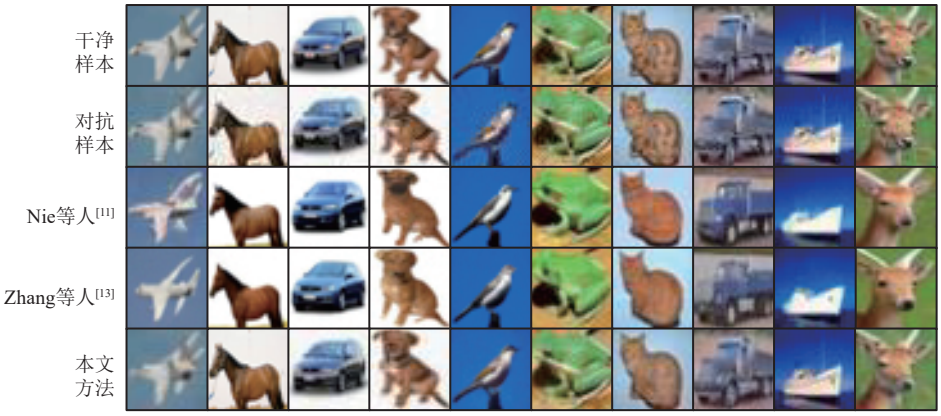
从表 9 中可以看出,在 CIFAR-10 数据集中本文方法的 PSNR 与 SSIM 分别为 28.92 dB 与 0.93,在 ImageNet 数据集中本文方法的 PSNR 与 SSIM 分别为 34.31 dB 与 0.87,通过与基线方法比较可知,本文方法不受

Table 9 Integrity Performance of Purification Results

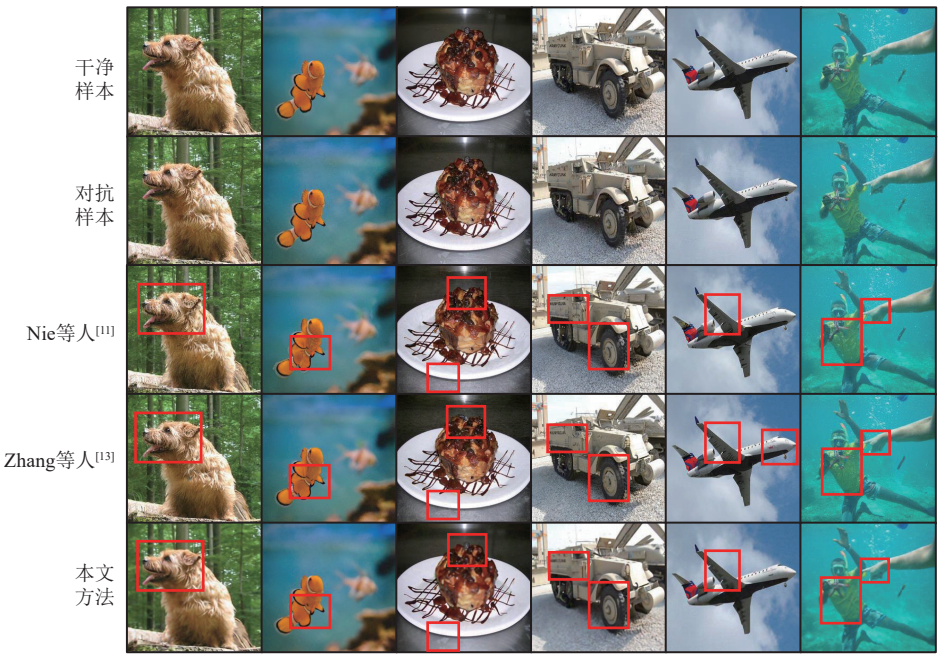
表 9 净化结果完整性表现

方法	CIFAR-10		ImageNet	
	PSNR/dB	SSIM	PSNR/dB	SSIM
Nie 等人[11]	19.33	0.73	31.38	0.72
Zhang 等人[13]	19.34	0.73	31.42	0.73
本文	28.92	0.93	34.31	0.87

分辨率限制，在像素级噪声抑制与内容重构方面具有显著优势，净化后的样本在纹理结构、边缘细节等信息的保留上更接近原始图像。图 5 为干净分量完整性对比实验部分结果图，从 CIFAR-10 与 ImageNet 实验图例中可以看出，Nie 等人[11] 与 Zhang 等人[13] 等基线方法净化后的图像存在明显模糊、细节丢失现象，例如物体边缘出现伪影及纹理结构难以辨认。而本文方法能够有效保存干净样本的结构特征，净化结果在结构和细节方面更接近干净状态下的样本，如物体轮廓清晰、色彩过渡自然，在细节保持（如毛发纹理、背景阴影）方面接近原始图像，进一步验证了量化指标的结论，说明本文方法在细节保持及干净样本还原度方面明显高于目前基线净化防御算法。



(a) CIFAR-10实验结果



(b) ImageNet实验结果

Fig. 5 Partial results of comparative experiment on the integrity of noise-free component

图 5 干净分量完整性对比实验部分结果

3.4 消融实验

在本节中，对本文方法及各个设计部分效能进行消融验证。实验在 CIFAR-10 固定子集上，主要对 AutoAttack l_∞攻击 WideResNet-28-10 分类模型进行防御，分别分析循环回归、分布调控、类内聚合对净化鲁棒性与干净分量完整性的影响。其中表 10 实验 1 中算法为本文所提完整方法，实验 2 使用 Score SDE 逆时序代替循环回归，实验 3 去掉分布调控，实验 4 去掉类内聚合。

由实验 2 与实验 1 对比可以看出，替换循环回归后 Robust Acc 下降 6.65 个百分点，验证了循环回归干扰定向消除的有效性，说明本文所提自监督的净

Table 10 Ablations Study on the Key Operations

表 10 关键操作的消融实验

实验	循环回归	分布调控	类内聚合	Standard Acc/%	Robust Acc/%	PSNR/dB	SSIM
1	√	√	√	94.92	78.91	28.92	0.93
2	×	√	√	93.36	72.26	28.89	0.91
3	√	×	√	90.23	75.39	19.33	0.73
4	√	√	×	92.38	73.63	28.91	0.93

化方式可使净化结果更加贴近干净样本分布状态。循环回归通过干净域与对抗域自引导方式抑制对抗样本中的噪声干扰来确保样本中核心特征完整有效。当该模块缺失后，对抗样本的干扰噪声无法定向消除，导致净化后结果对分类模型误导依然存在。同时本文方法的 Standard Acc 下降 1.56 个百分点，表明循环回归提高样本重构的鲁棒性。

实验 3 与实验 1 对比可知，当去掉分布调控后，Standard Acc, PSNR, SSIM 分别降低 4.69 个百分点、9.59dB、0.2, PSNR 与 SSIM 明显下降表明分布调控是结构完整及细节保持的核心模块，是减少扩散模型在逆向还原过程中干净分量丢失与畸变的关键步骤。

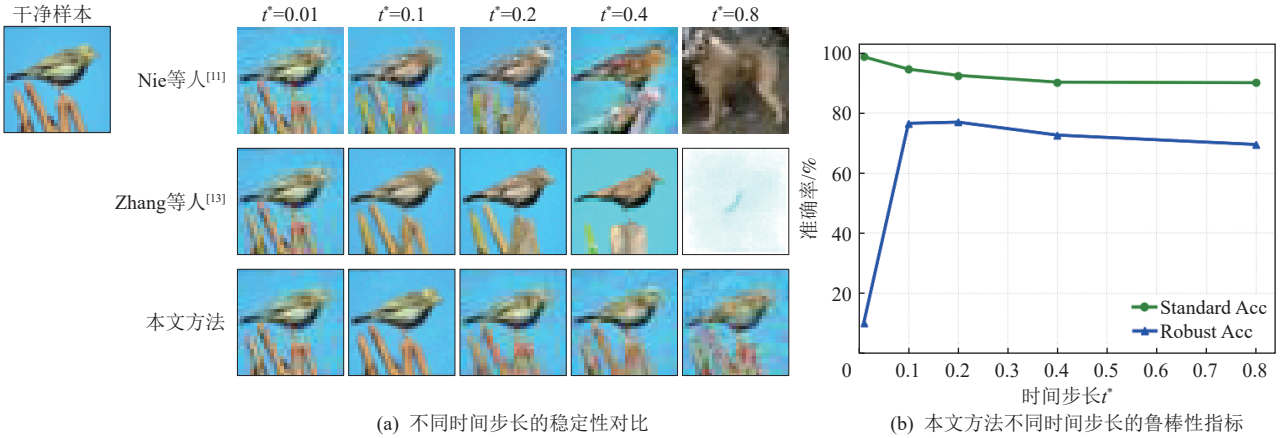
实验 4 与实验 1 对比可知，去掉类内聚合后 Standard Acc, Robust Acc 分别降低 2.54 个百分点、5.28 个百分点。这 2 项指标验证了类内聚合具有引导净化过程中样本向着概率分布高密度区间偏移，从而避免了样本分布概率处于决策边界而出现的类别不稳定现象。该模块通过先验引导弥补净化中类内方向缺失，优化各时刻生成量特征决策边界，提高样本类别置信度。

3.5 稳定性分析对比实验

基于扩散模型的净化方法中时间步长 t^* 是决定去干扰的关键参数，为了分析 t^* 对净化稳定性的影响，设计不同 t^* 的比较实验，其中数据集采用 CIFAR-10 固定子集，攻击方法和分类模型分别为 AutoAttack l_∞ 攻击与 WideResNet-28-10 模型。如表 11 所示，分别对 t^*=0.01, 0.1, 0.2, 0.4, 0.8下净化结果进行性能评估，随着时间步长 t^* 的增加，Nie 等人[11] 与 Zhang 等人[13] 等基线方法的 Standard Acc 与 Robust Acc 大幅度降低，表明净化后的结果逐渐发生分布偏离，生成的内容脱离干净样本类别，而本文方法对不同 t^* 具有较高的鲁棒性，Standard Acc 与 Robust Acc 指标相对稳定，在样本分布和干净分量保持上具有显著的优势。图 6(a)为抽取的实验样例，可以看出本文方法净化结果的稳定性。图 6(b)为本文不同时间步长 t^* 下 Standard Acc 与 Robust Acc 的表现。由图 6(b)可以看出 t^* = 0.1时净化效果最佳。当 t^*($t^* < 0.1$)较小时，Robust Acc 处于较低位置表明净化力度不足，此时对抗干扰仍然具备较强的对抗性；随着 t^*($t^* > 0.1$)增大，Standard Acc 与 Robust Acc 均保持稳定，表明本文方法净化结果的分布具有较强的稳定性。

3.6 固定子集实验评估的有效性验证

为提高实验的高效性，实验中所用评估数据均采用样本量为 512 张的固定子集，即从公共数据集中随机抽取 512 个样本构建的测试集。该种评估方式在对抗净化方法中普遍应用，Nie 等人[11] 通过实验证明固定子集与完整数据集实验评估的相对一致性。为进一步说明本文实验在固定子集评估的有效性，印证固定子集与完整数据集的实验结果一致性，针对本文所提净化方法，在 CIFAR-10 中分别抽取 512, 2048, 6 144, 12 288 个样本进行 Standard Acc 和 Robust Acc 分析，实验采用 AutoAttack l_∞ 攻击，分类模型选用 WideResNet-28-10 模型与 WideResNet-70-16 模型，实验结果如表 12 所示。从表 12 中可以看出，以 512



(a) 不同时间步长的稳定性对比 (b) 本文方法不同时间步长的鲁棒性指标

Fig. 6 Comparative experiment of stability analysis

图 6 稳定性分析对比实验

Table 11 Stability Analysis of t^*

表 11 t^*的稳定性分析

方法	t^*=0.01		t^*=0.1		t^*=0.2		t^*=0.4		t^*=0.8	
	Standard Acc/%	Robust Acc/%	Standard Acc/%	Robust Acc/%	Standard Acc/%	Robust Acc/%	Standard Acc/%	Robust Acc/%	Standard Acc/%	Robust Acc/%
Nie 等人[11]	95.90	60.16	89.02	70.64	82.42	53.13	62.50	43.95	0.04	0.04
Zhang 等人[13]	96.88	62.30	90.04	73.05	83.98	55.47	58.79	42.97	0.00	0.00
本文	100.00	0.02	94.92	78.91	92.77	79.69	88.09	71.29	88.09	68.95

Table 12 Effectiveness Validation of the Fixed Subset

表 12 固定子集有效性验证 %

固定子集样本数量	WideResNet-28-10		WideResNet-70-16	
	Standard Acc	Robust Acc	Standard Acc	Robust Acc
512	94.92	78.91	95.12	79.10
2 048	94.92	78.91	95.12	79.00
6 144	94.87	78.86	95.07	78.86
12 288	94.90	78.83	95.14	79.03

个样本的固定子集作为基准，Standard Acc 和 Robust Acc 指标最大浮动不超过 0.24 个百分点，且不会随着样本量的扩增出现指标值的提升或降低，验证了 Nie 等人[11]实验评估方法的一致性，说明了本文用固定子集进行鲁棒性指标评估是有效的。

4 结 论

为提高净化鲁棒性、干净分量完整性与净化结果分布的稳定性，本文提出基于扩散模型的分布双向引导对抗扰动净化防御方法，在扩散还原过程引入分布双向引导方法，实现了干扰的有向消除，并通过类间与类内的分布调整，提高了干净分量保持能力与结果的稳定性。其中类间循环引导依靠对抗域与干净域之间循环回归完成了有向干扰消除，通过分布调控进一步提高净化结果的稳定性与信息保持能力。类内聚合在干净域到对抗域生成中最大化各时刻生成量类别置信度，实现概率分布向高密度区域集中。本文在数据集 CIFAR-10, CIFAR-100, ImageNet 上的各项对比实验证明了本文方法的有效性，且在净化鲁棒性、干净分量完整性、净化稳定性等方面优于目前基线净化防御方法。消融实验也进一步验证了本文方法各设计部分在净化中的作用及有效性。由于本文所提的基于扩散模型的分布双向引导方法需在干净域与对抗域中进行多次还原生成操作，因此算法的时效性不高，在今后工作中，我们将进一步研究如何在保持净化指标基础上提高净化速率。

作者贡献声明：赵新宇提出了算法思路，实现算法和执行实验，并撰写了论文主要内容；李司负责设计实验方案并参与论文修改。

参 考 文 献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312.6199, 2013

[2] Carmon Y, Raghunathan A, Schmidt L, et al. Unlabeled data improves adversarial robustness[J]. Advances in Neural Information Processing Systems, 2019, 15: 8−14

[3] Dong Junhao, Seyed-Mohsen M, Lai Jianhuang, et al. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adver sarial training[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2023: 24678–24687

[4] Naman D S, Francesco C, Matthias H. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models[J]. Advances in Neural Information Processing Systems, 2023, 36: 13931−13955

[5] He Zhixun, Mukesh S. VQUNet: Vector quantization U-Net for defending adversarial attacks by regularizing unwanted noise[C]//Proc of the 7th Int Conf on Machine Vision and Application. New York: ACM, 2024: 69−76

[6] Ren Min, Zhu Yuhao, Wang Yunlong, et al. Perturbation inactivation based adversarial defense for face recognition[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2947−2962

[7] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models[J]. arXiv preprint, arXiv: 1805.06605, 2018

[8] Yang Ling, Zhang Zhilong, Song Yang, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2023, 56(4): 1−39

[9] Cao Hanqun, Tan Cheng, Gao Zhangyang, et al. A survey on generative diffusion models[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 13(2): 162−196

[10] Choi J, Kim S, Jeong Y, et al. ILVR: Conditioning method for denoising diffusion probabilistic models[C]//Proc of the 18th Int Conf on Computer Vision. Los Alamitos, CA: IEEE Computer Society, 2021: 1415−1428

[11] Nie Weili, Guo B, Huang Yujia, et al. Diffusion models for adversarial purification[C]//Proc of the 39th Int Conf on Machine Learning. New

York: PMLR, 2022: 16805−16827

[12] Wang Jinyi, Lyu Zhaoyang, Lin Dahua, et al. Guided diffusion model for adversarial purification[J]. arXiv preprint, arXiv: 2205.14969, 2022

[13] Zhang Mingkun, Li Jianing, Chen Wei, et al. Classifier guidance enhances diffusion-based adversarial rurification by preserving predictive information[C]//Proc of the 27th European Conf on Artificial Intelligence. Berlin: Springer, 2024: 2234−2241

[14] Lin Guang, Tao Zerui, Zhang Jianhai, et al. Robust diffusion models for adversarial purification[J]. arXiv preprint, arXiv: 2403.16067, 2024

[15] Li Minghui, Jiang Peipei, Wang Qian, et al. Adversarial attacks and defenses for deep learning models[J]. Journal of Computer Research and Development, 2021, 58(5): 909−926 (in Chinese)
(李明, 江沛佩, 王骞, 等. 针对深度学习模型的对抗性攻击与防御[J]. 计算机研究与发展, 2021, 58(5): 909−926)

[16] Peng Shengyun, Xu Weilin, Cornelius C, et al. Robust principles: Architectural design principles for adversarially robust CNNs[J]. arXiv preprint, arXiv: 2308.16258, 2023

[17] Lin Guang, Tao Zerui, Zhang Jianhai, et al. Adversarial guided diffusion models for adversarial purification[J]. arXiv preprint, arXiv: 2403.16067, 2024

[18] Gowal S, Rebuffi S A, Wiles O, et al. Improving robustness using generated data[J]. Advances in Neural Information Processing Systems, 2021, 34: 4218−4233

[19] Lin Guang, Li Chao, Zhang Jianhai, et al. Adversarial training on purification: Advancing both robustness and generalization[J]. arXiv preprint, arXiv: 2401.16352, 2024

[20] Wang Zekai, Pang Tianyu, Du Chao, et al. Better diffusion models further improve adversarial training[C]//Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 36246−36263

[21] Bi Xiuli, Yang Zonglin, Liu Bo, et al. ZeroPur: Succinct training-free adversarial purification[J]. arXiv preprint, arXiv: 2406.03143, 2024

[22] Gowal S, Qin Chongli, Uesato J, et al. Uncovering the limits of adversarial training against norm-bounded adversarial examples[J]. arXiv preprint, arXiv: 2010.03593, 2020

[23] Rebuffi S A, Gowal S, Calian D A, et al. Fixing data augmentation to improve adversarial robustness[J]. arXiv preprint, arXiv: 2103.01946, 2021

[24] Bai Mingyuan, Huang Wei, Li Tenghui, et al. Diffusion models demand contrastive guidance for adversarial purification to advance[C]//Proc of the 41st Int Conf on Machine Learning. New York: PMLR, 2024: 2640−3498

[25] Song Yang, Kim T, Nowozin S, et al. PixelDefend: Leveraging generative models to understand and defend against adversarial examples[C]//Proc of the 6th Int Conf on Learning Representations. Washington: ICLR, 2018: 1890−1910

[26] Yang Yuzhe, Zhang Guo, Katabi D, et al. ME-Net: Towards effective adversarial robustness with matrix estimation[C]//Proc of the 36th Int Conf on Machine Learning. New York: PMLR, 2019: 984−1006

[27] Hill M, Mitchell J, Zhu Songchun. Stochastic security: Adversarial defense using long-tun dynamics of energy-based models[C]//Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021: 4735−4753

[28] Yoon J, Hwang S J, Lee J. Adversarial purification with score-based generative models[C]//Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 12062–12072

[29] Pang Tianyu, Lin Min, Yang Xiao, et al. Robustness and accuracy could be reconcilable by (proper) definition[C]//Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 17258–17277

[30] Cui Jiequan, Tian Zhuotao, Zhong Zhisheng, et al. Decoupled Kullback-Leibler divergence loss[J]. Advances in Neural Information Processing Systems, 2024, 37: 74461−74486

[31] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training[J]. arXiv preprint, arXiv: 2001.03994, 2020

[32] Salman H, Ilyas A, Engstrom L, et al. Do adversarially robust imagenet models transfer better?[J]. Advances in Neural Information Processing Systems, 2020, 33: 3533−3545

[33] Bai Tao, Luo Jinqi, Zhao Jun, et al. Recent advances in adversarial training for adversarial robustness[C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2021: 4312–4321



Zhao Xinyu, born in 1990. Master, associate professor. His main research interests include computer vision and artificial intelligence security.
赵新宇, 1990 年生。硕士, 副研究员。主要研究方向为计算机视觉、人工智能安全。



Li Si, born in 1982. PhD candidate, professor. His main research interests include signal processing and artificial intelligence.
李　司, 1982 年生。博士研究生, 研究员。主要研究方向为信号处理、人工智能。