基于输入多样性与模型集成的可迁移对抗样本生成方法

徐鹤[1,2]，张恒[1]，李鹏[1,2]，郑文龙[1]，朱枫[1,2]
（1.南京邮电大学计算机学院，南京 210023; 2.南京邮电大学网络安全与可信计算研究所，南京 210023）

摘要：对抗样本的研究能够挖掘深度学习模型的鲁棒性缺陷，推动防御机制的研发与优化，进而提升图像分类模型在实际场景中的安全应用能力。然而，现有的对抗样本攻击方法大多存在着黑盒迁移能力不强，生成的对抗样本泛化能力不足的问题。为解决上述问题，本文提出了一种基于输入多样性与模型集成的可迁移对抗样本生成方法 DE-TEG。首先，引入多尺度变换的输入多样性增强策略，通过随机缩放、自适应填充/裁剪操作扩展优化过程的输入分布，使对抗样本摆脱对特定图像尺度与位置的依赖，学习更具泛化性的梯度信息，提升其对输入变换的鲁棒性；其次，构建由多种异构网络架构组成的模型集成组，通过对多模型的对抗损失函数进行平均融合，引导对抗样本学习跨模型的通用对抗特征，从而提升其在黑盒场景下的迁移攻击性能；最后，在迭代优化过程中引入动量优化机制，通过累积历史梯度的方向信息形成优化惯性，有效稳定参数更新方向，加速复杂损失曲面下的收敛过程，避免优化过程陷入局部最优解。实验结果表明，DE-TEG 在两个数据集下相较于对比实验可迁移攻击成功率平均提升约 3.16%。

关键词：图像分类；对抗样本；黑盒攻击；模型集成
中图分类号：TP391; TP18 **文献标志码**：A

引言

深度学习技术的飞速发展推动了计算机视觉领域[1,2]的跨越式进步，图像分类[3]凭借优异的特征提取与分类能力，被广泛部署于智能驾驶、安防监控、医疗影像诊断等实际场景，为生活的智能化升级提供了技术支撑。但深度学习模型并非具备绝对的鲁棒性，对抗样本[4]的发现揭示了其在安全应用层面的显著缺陷，这类通过在原始样本中添加人类视觉难以察觉微小扰动的样本，能够恶意诱导模型做出错误分类决策，给深度学习系统的实际落地埋下严重安全隐患，也使得对抗样本生成与防御的相关研究成为计算机视觉安全领域的核心研究方向。

在各类对抗攻击方法中，对抗补丁攻击[5]因仅需在图像局部区域嵌入精心设计的扰动图案即可实现攻击目的，具备极强的实用性，成为近年来的研究热点。为提升对抗补丁的攻击效果与场景适配性，国内外学者已开展大量相关研究[6-11]，在白盒场景下[12]取得了良好的攻击效果，验证了对抗补丁在图像分类模型攻击中的有效性，为后续对抗样本生成算法的研究奠定了方法基础。但尽管潜力巨大，现有研究仍面临众多问题，传统算法多以单一白盒模型[13-16]为优化目标，生成的对抗补丁易过拟合于特定模型的决策边界，难以学习到跨模型的通用对抗扰动特征，在面对未知架构的黑盒模型[17]时迁移攻击能力大幅衰减；部分模型优化算法因未有效处理模型梯度噪声与不一致性问题，存在优化收敛缓慢、易陷入局部最优解的缺陷，导致对抗补丁的生成质量与攻击稳定性不佳。上述问题的存在，限制了对抗补丁攻击在真实黑盒场景中的实际应用，也使得提升对抗补丁的黑盒迁移性与优化稳定性成为该领域的关键研究难点。

为解决上述问题,本文提出了一种基于输入多样性与模型集成的可迁移对抗样本生成方法 DE-TEG。通过引入多尺度变换的输入多样性增强策略，扩展优化过程的输入分布，使对抗样本摆脱对特定图像尺度与位置的依赖，学习更具泛化性的梯度信息，提升对输入变化的鲁棒性；其次，构建由多种异构网络架构组成的模型集成组[18]，通过对多模型的对抗损失函数进行平均融合，引导对抗样本学习跨模型的通用对抗特征，从而提升其在黑盒场景下的迁移攻击性能；在多模型集成优化的场景下，由于需要同时优化多个模型的损失，损失曲面更

收稿日期:2026-03-18; 修回日期:2026-07-04
基金项目:国家自然科学基金(61902196)。

加复杂，梯度信息可能存在较大的噪声和不一致性，引入动量优化机制方法，通过累积历史梯度信息来平滑优化轨迹，能够有效抑制梯度噪声的影响，加速收敛过程，避免优化过程陷入局部最优解。

本文主要贡献如下:

(1)提出异构模型感知模块，通过理论梯度方差分析证明多异构模型联合优化可抑制单模型专属梯度噪声，解决对抗补丁容易过拟合单一模型的问题。结合动量累积策略平滑优化轨迹，解决了复杂损失曲面下的收敛不稳定问题，引导生成具备跨模型通用性的对抗扰动。

(2)引入随机缩放、自适应填充/裁剪的多尺度变换策略，消除对抗补丁对固定图像尺度、空间位置的依赖，学习通用梯度特征，进而生成迁移性更强的对抗样本。

(3)实验评估了本文所提方法基于 ImageNet 和 GTSRB 数据集，面向 VGG、DenseNet、MobileNet 和 ResNet 等系列图像分类模型，与其他先进的对照方法进行对比，实验证明，本文所提方法可有效提升攻击成功率和黑盒迁移能力，通过配对 t 检验验证本文方法具备统计学意义。

1 相关工作

1.1 图像分类模型

图像分类作为计算机视觉的基础任务，其模型架构的演进推动了各类视觉任务的发展，从浅层卷积网络到深层异构模型，研究人员不断突破模型的特征提取能力与泛化性能。Simonyan 等[19]提出的 VGG 系列模型，通过规整的 3×3 卷积核堆叠构建深度网络结构，简化了卷积层的设计逻辑，让模型具备更强的基础特征层级捕捉能力，成为图像分类领域的经典基准模型，但该系列模型存在参数量大、计算冗余的问题，推理效率受限。He 等[20]提出的 ResNet 系列模型创新性引入残差连接，有效缓解了深层网络训练中的梯度消失与梯度爆炸问题，实现了百层以上深层网络的稳定训练，其中 ResNet50、ResNet152 等模型因精度与效率的平衡，成为后续视觉任务与对抗研究的主流基础模型。Huang 等[21]提出的 DenseNet 系列模型则采用稠密连接机制，实现了网络各层间的特征高度复用，在减少参数量的同时提升了特征融合能力，DenseNet121 模型在小样本场景与复杂特征提取中表现优异。Howard 等[22]提出的 MobileNet 系列模型，基于深度可分离卷积设计轻量化架构，大幅降低了模型的计算量与参数量，MobileNetV2 模型成为端侧设备图像分类的核心选择。

1.2 对抗攻击

对抗攻击通过向输入图像添加人类视觉难以察觉的微小扰动，诱导模型输出错误预测，已成为评估模型鲁棒性的重要手段。近期研究发现，打印出来的对抗样本能有效欺骗神经网络模型。Thys 等[23]提出 AdvPatch，首次针对行人这类高类内多样性目标开展对抗补丁攻击，提出了可打印的小型对抗补丁生成方法，设计了融合可打印性损失、总变分损失与检测目标损失的联合优化函数，通过 Adam 算法优化补丁像素值，证实了安防监控系统中行人检测模型的对抗脆弱性，但该研究仅针对单一检测器开展攻击，生成的补丁跨模型迁移性较差，难以应对多模型融合的检测架构。Hu 等[24]开发基于环形裁剪的可扩展生成攻击方法制作 AdvTexture，包含训练可扩展生成器和优化潜在变量两个阶段，利用环形裁剪技术优化局部模式，通过平铺相同单元生成最终潜在变量，赋予生成器平移不变性，让不同位置生成的纹理有一致对抗效果。Karmon 等[25]提出了 LOVAN，针对传统对抗补丁攻击中补丁位置固定或随机导致的攻击效果有限问题，设计了联合优化补丁内容与位置的攻击策略。该方法通过梯度上升最大化交叉熵损失，在迭代过程中通过贪心搜索策略优化补丁位置，同时提出基于该攻击的对抗训练方法。Chao 等[26]提出 DorPatch，该方法将补丁分解为二进制掩码与图案的哈达玛积，通过两阶段优化解决混合整数规划问题，生成的补丁在数字域具备优异的攻击性能，但黑盒迁移性仍有提升空间。

2 方法

2.1 问题定义

对抗补丁攻击通过设计局部可见、可部署的对抗补丁，覆盖或附着在目标对象上，诱导深度学习模型产生错误预测。带有对抗补丁的恶意图片可表示为 x^{adv},攻击目标公式如式(1)所示，其中， $f(\cdot)$ 为模型对输入样本的预测概率， y_{true} 为数据真实标签。

$$\arg\max f(x^{adv}) \neq y_{\text{true}} \tag{2}$$

对抗补丁攻击优化目标是最小化真实标签的预测概率，如式(2)所示。

$$x^{\text{adv}} = \arg\min_{d(x,x^{\text{adv}})\leq\varepsilon} \mathrm{L}\left(f(x^{\text{adv}}), y_{\text{true}}\right) \tag{3}$$

其中， $\arg\min_{d(x,x^{adv})\leq\varepsilon}$ 表示约束原始样本x和对抗样本 x^{adv} 之间的差距$d(x,x')$不超过一个很小的阈值ϵ。$\mathcal{L}(,)$表示损失函数，衡量模型预测结果与真实标签之间的差异。

2.2 方法框架

为了提升对抗补丁的可迁移性,本文提出一种基于输入多样性与模型集成的可迁移对抗样本生成方法 DE-TEG。该模型由输入多样性变换和模型集成两部分组成,具体流程如图 1 所示。

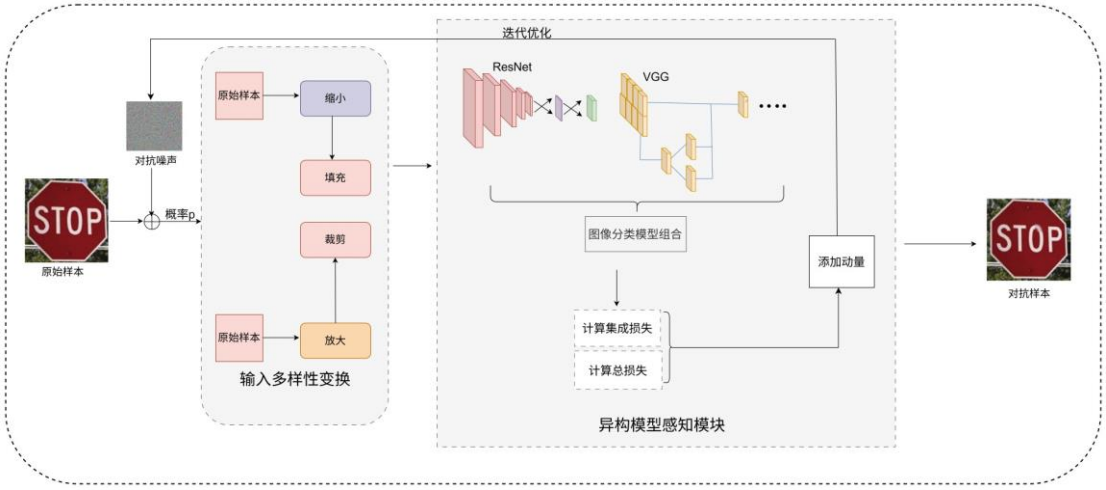

图 1　DE-TEG 生成框架图

Fig.1 Generation framework of DE-TEG

图 1 左半部分表示原始样本以随机概率 p 进行输入多样性变换，执行随机变换和自适应填充/裁剪操作，通过多尺度变换拓展输入分布，使模型能学习到更泛化的梯度信息，避免对抗样本过拟合于特定图像尺度或位置。图 1 的右半部分表示经过多样性变换后的样本被送入异构模型感知模块，该模块由 ResNet、VGG 等不同结构的图像分类模型组成，各异构模型对样本进行特征提取与分类预测后，计算多模型的集成损失与总损失，通过跨模型梯度融合抑制单一模型的特有噪声，引导对抗样本学习跨模型的通用对抗特征。最后，引入动量优化机制，累积历史梯度方向形成优化惯性，平滑梯度更新轨迹、稳定优化方向，避免陷入局部最优并加速收敛。经过多轮迭代优化后，最终生成具备强黑盒迁移能力的对抗样本。

2.3 输入多样性变换

在对抗攻击研究中，对抗样本的迁移性与其对输入变换的鲁棒性[27]密切相关。基于这一观察，在每次迭代优化过程中，对输入样本施加随机变换，使优化过程不依赖于特定的输入形式，从而学习到更具泛化能力的对抗扰动。本方法设计的输入多样性增强策略包含随机缩放与自适应填充/裁剪两个关键操作。

（1）随机缩放

随机缩放变换的目的是在每次优化迭代中对输入图像的空间尺度进行随机扰动，使生成的对抗补丁不依赖于特定的图像分辨率。设输入对抗样本为 x^{adv}，其维度为$C\times H\times W$，其中 C 表示图像的通道数，H 表示图像的高度，W 表示图像的宽度。本方法首先在预设的缩

放范围内随机采样缩放因子 s，设缩放因子的取值范围为$[\alpha_{\min}, \alpha_{\max}]$。缩放因子 s 从该区间内的连续均匀分布中随机采样，其数学表达式如公式(3)所示：

$$s \sim \mathrm{U}(\alpha_{\min}, \alpha_{\max}) \tag{4}$$

根据采样得到的缩放因子 s，计算变换后的目标图像尺寸。设变换后图像的高度为 H'，宽度为 W'， floor() 为向下取整，其计算如公式(4)所示：

$$H' = \mathrm{floor}(H \cdot s), W' = \mathrm{floor}(W \cdot s) \tag{5}$$

随后，采用双线性插值方法将原始图像重采样至目标尺寸。设重采样后的图像为 x_{resized}，重采样过程如公式(5)所示：

$$x_{\text{resized}} = \mathrm{Interpolate}(x^{\text{adv}}, (H', W'), \mathrm{mode} = \mathrm{bilinear}) \tag{6}$$

其中，*Interpolate*() 表示图像插值函数，mode = bilinear 指定采用双线性插值模式。双线性插值通过对目标像素位置周围的四个最近邻像素进行距离加权平均来计算新像素值，能够在保持图像平滑性的同时实现任意比例的尺度变换。

(2)自适应填充

当缩放因子 $s < 1$ 时，经过重采样后的图像尺寸小于原始尺寸。为保持神经网络输入维度的一致性，需要通过填充操作将缩小后的图像恢复至原始尺寸。首先，计算水平方向和垂直方向需要填充的总像素数。设水平方向的填充总量为Δ_W，垂直方向的填充总量为Δ_{H}。对于水平方向，需要将填充总量 Δ_W 分配至图像的左侧和右侧。设左侧填充量为 p_{left}，右侧填充量为 p_{right}。左侧填充量从离散均匀分布中随机采样，如公式(6)所示。在垂直方向也是同理，如公式(7)所示。确定四个方向的填充量后，使用常数值填充将缩小的图像扩展至原始尺寸。设填充后的图像为 $\tilde{x}$，填充操作如公式(8)所示：

$$p_{\text{left}} \sim \mathrm{U}_{\text{discrete}}(0, \Delta_W), p_{\text{right}} = \Delta_W - p_{\text{left}} \tag{7}$$

$$p_{\text{top}} \sim \mathrm{U}_{\text{discrete}}(0, \Delta_H), p_{\text{bottom}} = \Delta_H - p_{\text{top}} \tag{8}$$

$$\tilde{x} = \mathrm{Pad}\left(x_{\text{resized}}, \left(p_{\text{left}}, p_{\text{right}}, p_{\text{top}}, p_{\text{bottom}}\right), v\right) \tag{9}$$

其中， Pad() 表示填充函数，第一个参数 x_{resized} 为待填充的缩小图像，第二个参数(, , ,)为四个方向的填充量，第三个参数 v 为填充像素值。经过填充操作后，图像 $\tilde{x}$ 的尺寸从 $C \times H' \times W'$ 恢复为 $C \times H \times W$，可直接输入神经网络进行前向传播和梯度计算。

(3)自适应裁剪

当缩放因子 $s > 1$ 时，经过重采样后的图像尺寸大于原始尺寸。此时需要通过裁剪操作将放大的图像缩减至原始尺寸。本方法采用中心裁剪策略，从放大后图像的中心区域提取与原始尺寸相同的子图像。设裁剪的起始坐标为(c_x, c_y)，其中 c_x 表示水平方向的起始像素位置，c_y 表示垂直方向的起始像素位置。基于起始坐标，中心裁剪操作如公式（9）所示：

$$\tilde{x} = x_{\text{resized}}[:, :, c_y : c_y + H, c_x : c_x + W] \tag{10}$$

其中，切片操作表示保留第一维和第二维所有通道，在高度维度上提取从第 c_y 行到第 $(c_y + H - 1)$ 行共 H 行，在宽度维度上提取从第 c_x 列到第 $(c_x + W - 1)$ 列共 W 列。经过该操作，放大图像 x_{resized} 被裁剪为尺寸为 $C \times H \times W$ 的图像 $\tilde{x}$，等效于对原始图像进行了轻微的放大观察，突出图像的中心内容。

2.4 异构模型感知模块

单一模型生成的对抗补丁易过度拟合该模型特有的决策边界特征，导致其仅对该特定模型有效，而在其他未见过的异构模型上攻击效果显著下降。为提升对抗补丁的可迁移性，本文提出异构模型感知模块，通过同时在多个具有不同网络架构的深度学习模型上进行联合优化，使生成的对抗补丁能够学习到跨模型的通用对抗特征，从而提升其在未知目标模型上的迁移攻击能力。从稳定性角度展开理论分析，对比单模型策略[28]和集成模型策略的泛化误差上界，挖掘提升对抗补丁在不同模型间可迁移性的关键因素。

设*Var*为方差，*Cov*为协方差，单个模型梯度可分解为$g_k = g^* + \varepsilon_k$，其中$g^*$为所有模型共享的通用对抗方向，$\varepsilon_k$为第 k 个模型的特有噪声，满足$\mathrm{Var}(\varepsilon_k)=\sigma^2$，噪声相关系数为$\rho_\varepsilon = \dfrac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{\sigma^2}$，$K$为集成模型数量，则集成梯度$g_{ens}$的方差可表示为：

$$\mathrm{Var}(g_{\mathrm{ens}}) = \frac{\sigma^2}{K}\left[1+(K-1)\rho_\varepsilon\right] \tag{11}$$

证明：

由梯度分解式$g_k = g^* + \varepsilon_k$，其中$g^*$为确定值($\mathrm{Var}(g^*)=0$)，且$g^*$与$\varepsilon_k$相互独立，因此：

$$\mathrm{Var}(g_k) = \mathrm{Var}(g^* + \varepsilon_k) = \mathrm{Var}(g^*) + \mathrm{Var}(\varepsilon_k) = 0 + \sigma^2 = \sigma^2 \tag{12}$$

集成梯度定义为：

$$g_{\mathrm{ens}} = \frac{1}{K}\sum_{k=1}^{K} g_k = \frac{1}{K}\sum_{k=1}^{K}(g^* + \varepsilon_k) = g^* + \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k \tag{13}$$

对集成梯度取方差：

$$\begin{aligned}
\mathrm{Var}(g_{\mathrm{ens}}) &= \mathrm{Var}\left(g^* + \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k\right) \\
&= \mathrm{Var}\left(\frac{1}{K}\sum_{k=1}^{K}\varepsilon_k\right) \\
&= \frac{1}{K^2}\left[\sum_{k=1}^{K}\mathrm{Var}(\varepsilon_k) + 2\sum_{i<j}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)\right] \\
&= \frac{1}{K^2}\left[K\sigma^2 + 2\cdot\frac{K(K-1)}{2}\cdot\rho_\varepsilon\sigma^2\right] \\
&= \frac{\sigma^2}{K}\left[1+(K-1)\rho_\varepsilon\right]
\end{aligned} \tag{14}$$

$$\rho_\varepsilon = \frac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{\sqrt{\mathrm{Var}(\varepsilon_i)\mathrm{Var}(\varepsilon_j)}} = \frac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{\sigma^2} \tag{15}$$

在噪声相互独立($\rho_\varepsilon = 0$)的理想异构场景下，可简化为：

$$\mathrm{Var}(g_{\mathrm{ens}}) = \frac{\sigma^2}{K} \tag{16}$$

集成梯度g_{ens}的方差与噪声相关系数ρ_ε成正比，与集成模型数量 K 成反比。在异构模型场景下，模型间噪声相关性ρ_ε较低，随着 K 增大，方差显著减小。方差的降低意味着集成梯度更新方向更稳定，可有效抑制单模型特有噪声，避免对抗样本过拟合单一模型结构，从而显著提升对抗样本的黑盒可迁移性与泛化能力。

在多模型集成优化的场景下，由于需要同时优化多个模型的损失，损失曲面更加复杂，梯度信息可能存在较大的噪声和不一致性，引入动量优化机制方法，通过累积历史梯度信息来平滑优化轨迹，能够有效抑制梯度噪声的影响，加速收敛过程。

设当前迭代步为 t，其中 t 表示优化过程中的迭代计数。设第 t 步的原始梯度为 $\nabla L(\theta_t)$，该梯度通过对集成损失函数关于待优化参数求导得到。定义动量累积项 v_t，μ 表示动量衰减系数，控制历史梯度信息的保留程度。动量累积项更新规则如公式(16)所示：

$$v_t = \mu \cdot v_{t-1} + \nabla L(\theta_t) \tag{17}$$

基于动量累积项，参数更新规则定义如公式(17)所示：

$$\theta_{t+1} = \theta_t - \alpha_t \cdot \text{sign}(v_t) \tag{18}$$

其中，θ_t 表示第 t 步的待优化参数，在对抗补丁生成中对应于补丁的掩模参数或图案参数。α_t 表示第 t 步的学习率，控制每步参数更新的步长大小；本方法采用自适应学习率策略，当损失长期不下降时自动降低学习率。$\text{sign}(v_t)$ 表示动量累积项的符号向量，表示参数更新的方向，以实现损失函数的最小化。

3 实验与结果分析

3.1 实验设置

3.1.1 数据集及分类模型

为了验证本文所提方法的攻击性能,选择在 2 个数据集和 9 个图像分类模型上进行测试。数据集选择 ImageNet 数据集和 GTSRB 数据集，ImageNet 包含海量多样化的图像样本，覆盖 1000 个类别的图像，总图像数量超过 100 万张，能够充分模拟复杂真实场景下的图像分布特征。GTSRB 数据集专为交通标志识别任务设计，涵盖 43 类日常道路中常见的交通标志，包括禁令标志、警告标志、指示标志等，全面覆盖了日常驾驶场景中的核心交通规则与警示信息。选取了 9 个具有不同网络架构的图像分类模型作为攻击目标，具体包括 ResNet50、ResNet101、ResNet152、DenseNet121、DenseNet169、VGG13、VGG16、VGG19 和 MobileNetV2。

3.1.2 对比方法及实验参数

实验选取了 4 种当前主流的对抗补丁攻击算法作为对照实验，分别为 DorPatch、AdvPatch、AdvTexture 和 LOVAN。所有实验均在 NVIDIA A800 GPU 上运行，操作系统为 Ubuntu22.04，Python 版本为 3.9.15，基于 PyTorch 1.11.0 与 CUDA 11.3。其中数据实验参数设置攻击次数为 500 次，采用的集成模型组包含 4 个异构预训练模型，分别为 ResNet152、DenseNet121、VGG16 和 MobileNetV2，密度正则化系数为 0.002，结构正则化系数为 0.002，优化学习率为 0.008。

3.1.3 评价指标

本文采用攻击成功率（ASR）作为评价指标，定量评估所提方法的攻击有效性。ASR 的计算公式如式(18)所示。其中，N 表示总测试样本数，x_i 表示原始干净样本，δ_i 表示对抗扰动，$f(\cdot)$ 表示模型预测函数，y_i^{true} 表示真实标签。由计算公式可知，ASR 值越大，表明攻击算法的性能越优。

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N} \text{I}\left(f(x_i + \delta_i) \neq y_i^{\text{true}} \right) \tag{18}$$

3.2 白盒攻击对比实验及结果分析

为了验证本文所提方法的白盒攻击能力，通过在 5 种不同的白盒场景下，将 4 种基线攻击方法与所提方法进行白盒攻击实验，对比 ASR 值，可以看出所提方法在白盒模型上拥有较好的攻击成功率，实验结果如表 1 所示。

表 1 不同方法生成的对抗补丁在白盒场景下攻击成功率对比（%）

Table 1 Attack success rates (%) of adversarial patches from different methods in white-box scenarios

数据集	目标模型	DorPatch	AdvPatch	AdvTexture	LOVAN	DE-TEG
ImageNet	ResNet50	99.36	94.59	98.47	97.10	99.47
	ResNet152	98.26	94.12	94.58	94.01	98.91
	DenseNet121	98.15	94.03	94.27	93.95	98.83
	VGG16	98.52	94.37	95.11	94.89	99.18
	MobileNetV2	97.85	95.70	96.14	95.83	99.52
GTSRB	ResNet50	99.28	95.11	98.62	97.35	99.51
	ResNet152	98.31	94.45	94.79	94.26	98.97
	DenseNet121	98.20	94.18	94.43	94.10	98.89
	VGG16	98.60	94.52	95.27	95.05	99.23
	MobileNetV2	97.92	95.85	96.29	95.98	99.57

从 ImageNet 数据集和 GTSRB 数据集的实验结果来看，所提算法在所有白盒模型上均表现出最优的攻击性能，且维持了极高的攻击成功率。在 ImageNet 数据集上，针对 ResNet50 模型，所提算法的 ASR 达到 99.47%，略高于基线算法中表现最优的 DorPatch（99.36%），相较于 AdvPatch、AdvTexture、LOVAN 分别提升了 4.88%、1.00%、2.37%。对于 MobileNetV2，所提算法的 ASR 达到 99.52%，较 DorPatch、AdvPatch、AdvTexture、LOVAN 分别提升 1.67%、3.82%、3.38%、3.69%。在 GTSRB 数据集上，针对 VGG16 模型：所提方法 ASR 为 99.23%，较 DorPatch 提升 0.63%，较 AdvPatch、AdvTexture、LOVAN 分别提升 4.71%、3.96%、4.18%。由表 1 中的实验结果可知，本文所提方法在白盒攻击上优于其它四种基线方法。

3.3 黑盒攻击对比实验及结果分析

在白盒攻击中，所提算法展现了较好的攻击能力，为进一步验证本文所提方法的黑盒迁移能力，实验分别在 5 种不同的图像分类模型中测试黑盒攻击成功率，充分验证所提方法的黑盒迁移能力。其中对比方法采用在 DenseNet121 模型上生成对抗补丁，生成补丁效果图如图 2 所示，将其应用于 ResNet50、ResNet101、DenseNet169、VGG16、VGG19 五种模型上测试黑盒迁移能力，实验结果如表 2 所示。

表 2 不同方法生成的对抗补丁在黑盒场景下攻击成功率对比（%）

Table 2 Attack success rates (%) of adversarial patches from different methods in black-box scenarios

数据集	攻击方法	ResNet50	ResNet101	DenseNet169	VGG13	VGG19
ImageNet	DorPatch	45.16	35.08	46.40	47.93	41.52
	AdvPatch	42.01	33.16	36.51	41.06	43.20
	AdvTexture	34.52	27.55	29.37	37.71	34.61
	LOVAN	38.24	29.56	34.72	40.79	36.53
	DE-TEG	47.27	37.10	47.58	53.23	50.29
GTSRB	DorPatch	45.70	42.99	42.85	41.35	41.64
	AdvPatch	45.66	43.21	44.51	47.06	41.83
	AdvTexture	43.82	41.27	40.16	39.29	39.41
	LOVAN	40.24	33.79	41.72	44.03	44.52
	DE-TEG	46.10	45.56	48.24	50.91	48.62

从表 2 分析可知，攻击其余 5 个黑盒模型时，与基线方法相比，本文所提出的方法在黑盒设置下的攻击成功率得到明显提升。在 ImageNet 数据集上，所提方法在 ResNet50 与 VGG16 模型上分别取得 47.27%和 53.23%的黑盒攻击成功率，均显著优于 DorPatch、AdvPatch 等基线方法，尤其在 VGG16 上实现了超过 5%的性能提升。在 GTSRB 数据集上，所提方法在 DenseNet169 与 VGG16 模型上的攻击成功率分别达到 48.24%和 50.91%，相较于 DorPatch 分别提升了 5.39%和 3.85%，同时在轻量化模型 MobileNetV2 上保持了 48.62%的高攻击成功率。综合实验结果表明，即使针对不同骨干网络和不同结构的分类模型，本文所提方法仍能有效提升其黑盒迁移性能。

(a) 原始图片 (b) AdvPatch (c) AdvTexture (d) LOVAN (e) DorPatch (f) DE-TEG

图 2 DE-TEG 与 AdvPatch、AdvTexture、LOVAN、DorPatch 在 Densenet121 上生成补丁效果对比
Fig.2 Comparison of patch effects generated by DE-TEG, AdvPatch, AdvTexture, LOVAN and DorPatch on DenseNet121

3.4 统计显著性检验

为验证 DE-TEG 相比基线方法（尤其是 DorPatch）的性能提升是否具有统计学意义，对 ImageNet 数据集，ResNet50 检测器上的黑盒攻击成功率（ASR）进行了配对 t 检验，具体结果如表 3 所示。

表 3 DE-TEG 和 DorPatch 的配对 t 检验数据

Table 3 Paired t-test data of DE-TEG and DorPatch

配对编号	均值	标准差	95%CI	t	df	双侧 p
DE-TEG–DorPatch	1.894	1.178	[0.430,3.357]	3.594	4	0.023

采用配对样本 t 检验比较 DE-TEG 与基线方法 DorPatch 的攻击性能差异。结果显示，DE-TEG 的平均攻击成功率显著高于 DorPatch，均值差为 1.894%，95%置信区间为[0.430%, 3.357%]，$t(4)=3.594$，$p=0.023<0.05$。这说明本文方法的性能提升并非随机波动，而是具有统计学意义的稳定提升。

3.5 消融实验

为精准量化所提算法中各核心模块的独立贡献及模块间的协同作用，本文设计了一系列消融实验。实验以未引入输入多样性策略与异构模型感知模块的基础方案作为 Baseline，通过逐步添加关键模块（1 为多尺度变换的输入多样性策略，2 为异构模型感知模块），最终对比全模块融合方案的黑盒迁移攻击性能，验证各模块对提升方法迁移性能的实际贡献。实验基于 ImageNet 数据集，以 ResNet50、ResNet101、DenseNet169、VGG13、VGG19 为黑盒攻击目标，核心评价指标仍为攻击成功率 ASR，实验结果如表 4 所示。

表 4 不同模块下的消融结果（%）

Table 4 Ablation experiment results for different modules (%)

攻击方法	ResNet50	ResNet101	DenseNet169	VGG13	VGG19
Baseline	45.16	35.08	46.40	47.93	41.52
Baseline+1	45.93	36.25	46.81	49.11	44.63
Baseline+2	46.78	36.02	47.32	50.98	46.14
DE-TEG	47.27	37.10	47.58	53.23	50.29

从表 4 结果可知，在 Baseline 上引入输入多样性变换之后，所有黑盒模型的攻击成功率均实现稳定提升，如攻击 VGG19 模型 ASR 从 41.52%提升至 44.63%。对比 Baseline 与 Baseline+2 的实验结果可知，异构模型集成策略对迁移性能的提升更为显著，如 VGG13 从 47.93%提升至 50.98%。对比 Baseline 和 DE-TEG，其中在 VGG13 模型上 ASR 从 47.93%提升至 53.23%。综上可得，本文所提方法在实现较高白盒攻击成功率的同时，还能显著提升对抗样本的黑盒迁移性能。

4 结束语

本文提出了一种针对图像分类模型的对抗样本生成方法，以解决传统对抗样本生成算法存在的迁移性不足及优化效率与稳定性不佳等问题，通过融合多尺度变换的输入多样性增强策略，提升对抗样本对图像尺度变化与位置偏移的适配能力，使其学习更具泛化性的梯度信息，进而提高对抗样本的迁移性能；此外通过异构模型集成优化框架迫使对抗样本学习跨模型通用对抗特征，依托动量优化机制平滑优化轨迹并加速收敛。通过实验结果表明，此方法可有效提升白盒攻击成功率和黑盒迁移能力。未来的工作将进一步探讨如何在更多的真实业务场景中应用此方法，并基于该攻击思路研究对抗防御方案，以此提升图像分类模型的鲁棒性和安全性。

参考文献：

[1] SHEN M, LI C, YU H, et al. Decision-based query efficient adversarial attack *via* adaptive boundary learning[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(4): 1740-1753.

[2] MO K, TANG W, LI J, et al. Attacking deep reinforcement learning with decoupled adversarial policy[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(1): 758-768.

[3] XU N, FENG W, ZHANG T, et al. A unified optimization framework for feature-based transferable attacks[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 4794-4808.

[4] CHEN Z, LI B, WU S, et al. Query-efficient decision-based black-box patch attack[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 5522-5536.

[5] 何琨, 佘计思, 张子君, 等. 基于引导扩散模型的自然对抗补丁生成方法[J]. 电子学报, 2024, 52(2): 564-573.
HE Kun, SHE Jisi, ZHANG Zijun, et al. A guided diffusion-based approach to natural adversarial patch generation[J]. Acta Electronica Sinica, 2024, 52(2): 564-573.

[6] ZHOU M, ZHOU W, HUANG J, et al. Stealthy and effective physical adversarial attacks in autonomous driving[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 6795-6809.

[7] DOAN B G, XUE M, MA S, et al. TnT attacks! universal naturalistic adversarial patches against deep neural network systems[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3816-3830.

[8] YANG Y, LIN C, LI Q, et al. Quantization aware attack: Enhancing transferable adversarial attacks by model quantization[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 3265-3278.

[9] 王振, 韩纪庆, 何勇军, 等. 面向 Transformer 语音识别模型的高迁移通用对抗样本生成方法[J]. 数据采集与处理, 2026, 41(1): 109-116.
WANG Zhen, HAN Jiqing, HE Yongjun, et al. Universal adversarial example generation method with high transferability for transformer-based speech recognition models[J]. Journal of Data Acquisition & Processing, 2026, 41(1): 109-116.

[10] DING X, CHEN J, YU H, et al. Transferable adversarial attacks for object detection using object-aware significant feature distortion[C]//Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium

on Educational Advances in Artificial Intelligence. ACM, 2024: 1546-1554.

[11] LIU J, ZHOU J, ZENG J, et al. DifAttack: Query-efficient black-box adversarial attack via disentangled feature space[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(4): 3666-3674.

[12] 张洁, 张志昊. 基于 AE-WGAN 的定向对抗样本生成及利用[J]. 南京邮电大学学报(自然科学版), 2020, 40(1): 63-69.
ZHANG Jie, ZHANG Zhihao. Generating and utilizing targeted adversarial examples by AE-WGAN transformation[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2020, 40(1): 63-69.

[13] 陆宇轩, 刘泽禹, 罗咏刚, 等. 基于生成对抗网络的目标检测黑盒迁移攻击算法[J]. 软件学报, 2024, 35(7): 3531-3550.
LU Yuxuan, LIU Zeyu, LUO Yonggang, et al. Black-box transferable attack method for object detection based on GAN[J]. Journal of Software, 2024, 35(7): 3531-3550.

[14] ZHAO Y, ZHU H, LIANG R, et al. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London United Kingdom: ACM, 2019: 1989-2004.

[15] JING Y, YANG Y, FENG Z, et al. Neural style transfer: A review[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(11): 3365-3385.

[16] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC, Canada. IEEE, 2021: 2778-2788.

[17] 武阳, 刘靖. 面向图像分析领域的黑盒对抗攻击技术综述[J]. 计算机学报, 2024, 47(5): 1138-1178.
WU Yang, LIU Jing. A survey on black-box adversarial attack in image analysis[J]. Chinese Journal of Computers, 2024, 47(5): 1138-1178.

[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA: IEEE, 2017: 39-57.

[19] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. 2014: arXiv: 1409.1556. https://arxiv.org/abs/1409.1556

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

[21] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2261-2269.

[22] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. 2017: arXiv: 1704.04861. https://arxiv.org/abs/1704.04861

[23] THYS S, VAN RANST W, GOEDEMÉ T. Fooling automated surveillance cameras: Adversarial patches to attack person detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA: IEEE, 2020: 49-55.

[24] HU Z, HUANG S, ZHU X, et al. Adversarial texture for fooling person detectors in the physical world[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 13297-13306.

[25] KARMON D, ZORAN D, GOLDBERG Y. LaVAN: Localized and visible adversarial noise[EB/OL]. 2018: arXiv: 1801.02608. https://arxiv.org/abs/1801.02608

[26] HE C, MA X, ZHU B B, et al. DorPatch: Distributed and occlusion-robust adversarial patch to evade certifiable defenses[C]//Proceedings 2024 Network and Distributed System Security Symposium. San Diego, CA, USA:

Internet Society, 2024.
[27] KOMKOV S, PETIUSHKO A. AdvHat: Real-world adversarial attack on ArcFace face ID system[C]//2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021: 819-826.
[28] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 9185-9193.

作者简介:

徐鹤（1985-），男，博士，教授，研究方向：人工智能与大数据、物联网技术，Email: xuhe@njupt.edu.cn。

张恒（2001-），男，硕士研究生，研究方向：人工智能与大数据，Email:1224046001@njupt.edu.cn。

李鹏（1979-），通信作者，男，博士，教授，研究方向：高性能计算、大数据、人工智能和网络安全，Email: lipeng@njupt.edu.cn。

郑文龙（1999-），男，博士，主要研究方向为网络安全，Email: 2024040412@njupt.edu.cn。

朱枫（1987-），男，博士，讲师，研究方向：网络安全和系统安全，Email: zhufeng@njupt.edu.cn。

A Transferable Adversarial Example Generation Method Based on Input Diversity and Model Ensemble

XU He[1,2], ZHANG Heng[1], LI Peng[1,2*], ZHENG Wenlong[1], ZHU Feng[1,2]

(1.School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023,China; 2.Institute of Network Security and Trusted Computing, Nanjing University of Posts and Telecommunications, Nanjing 210023,China)

Abstract：Research on adversarial examples can unearth the robustness flaws of deep learning models, drive the research, development and optimization of defense mechanisms, and thereby enhance the secure application capability of image classification models in real-world scenarios. However, most existing adversarial example attack methods suffer from weak black-box transferability and insufficient generalization of the generated adversarial examples. To address the above issues, this paper proposes a transferable adversarial example generation method named DE-TEG based on input diversity and model ensemble. First, an input diversity enhancement strategy with multi-scale transformation is introduced, which expands the input distribution in the optimization process through random scaling and adaptive padding/cropping operations. This enables adversarial examples to break away from the dependence on specific image scales and positions, learn more generalized gradient information, and improve their robustness to input transformations. Second, a model ensemble group composed of multiple heterogeneous network architectures is constructed. By averaging and fusing the adversarial loss functions of multiple models, adversarial examples are guided to learn cross-model universal adversarial features, thus enhancing their transfer attack performance in black-box scenarios. Finally, a momentum optimization mechanism is incorporated into the iterative optimization process. By accumulating the directional information of historical gradients to form optimization inertia, it effectively stabilizes the direction of parameter update, accelerates the convergence process on complex loss surfaces, and prevents the optimization process from falling into local optima. The experimental results show that, compared with the baseline methods, DE-TEG improves the average success rate of transferable attacks by approximately 3.16% on both datasets.

Highlights：

1. We design a heterogeneous model-aware module to reduce gradient variance via multi-model ensemble. Experiments prove it enhances the transferability of adversarial patches. A momentum optimization mechanism is also adopted to stabilize parameter updates, speed up convergence and avoid local optima.
2. We apply a transformation strategy with random scaling, adaptive padding and cropping. It improves the adaptability and generalization of adversarial patches against scale and position changes, yielding more transferable adversarial examples.
3. We conduct experiments on ImageNet and GTSRB using VGG, DenseNet, MobileNet and ResNet. Compared with state-of-the-art methods, our approach achieves higher attack success rates and better black-box transferability.

Key words：image classification; adversarial examples; black-box attack; model ensemble

---

Received: 2026-03-18; Revised: 2026-07-04

Foundation Item: National Natural Science Foundation of China (No. 61902196)

Biographies: XU He, male, PhD, Professor; LI Peng (Corresponding Author), male, PhD, Professor, E-mail: lipeng@niupt.edu.cn