通过融合 Top-1 分解特征的逻辑值提升对抗迁移性

Juanjuan Weng
厦门大学 厦门，中国

Zhiming Luo 厦门大学 厦门，中国

Dazhen Lin
厦门大学
厦门 ，中国

Shaozi Li 厦门大学 厦门，中国

Zhun Zhong
特伦托大学 特伦托，意大利

arXiv:2305.01361v3 [cs.CV] 5 Jul 2023

摘要

近期研究表明，深度神经网络（DNNs）极易受到对抗样本的攻击，这些对抗样本具有很强的可迁移性，可用于攻击其他未知的黑盒模型。为提高对抗样本的可迁移性，研究人员提出了多种基于特征的对抗攻击方法，以破坏中间层中的神经元激活。然而，当前最先进的基于特征的攻击方法通常需要额外的计算开销来估计神经元的重要性。为应对这一挑战，我们提出了一种基于奇异值分解（SVD）的特征级攻击方法。我们的方法受到以下发现的启发：从中间层特征分解得到的、与较大奇异值相关联的特征向量表现出更优越的泛化能力和注意力属性。具体而言，我们通过保留分解得到的与 Top-1 奇异值相关联的特征来进行攻击，并利用该特征计算输出逻辑值；随后，将输出逻辑值与原始逻辑值相结合，以优化对抗样本。大量实验结果验证了我们所提出方法的有效性。该方法可以轻松集成到各种基线方法中，从而显著增强对抗样本的可迁移性，以干扰正常训练的卷积神经网络和先进的防御策略。本研究的源代码可在 链接。

CCS 概念

• **安全与隐私**； • **计算方法** → **图像表示**；

关键词

黑盒攻击；奇异值分解；对抗样本；深度神经网络

1 引言

过去十年来，深度神经网络（DNNs）在各类计算机视觉任务中取得了显著进展，例如图像分类 [16], 目标检测 [29], 和图像分割 [30]。然而，Goodfellow 等人 [13] 揭示了在原始输入图像中添加几乎无法察觉的对抗扰动，会诱使深度神经网络产生错误的预测结果，从而给许多应用带来严重的安全问题，例如人脸识别 [8] 和自动驾驶 [20]。因此，这一点至关重要

通过预先开发更有效的攻击方法来识别深度神经网络的弱点，进而可利用这些方法提升深度神经网络的鲁棒性。

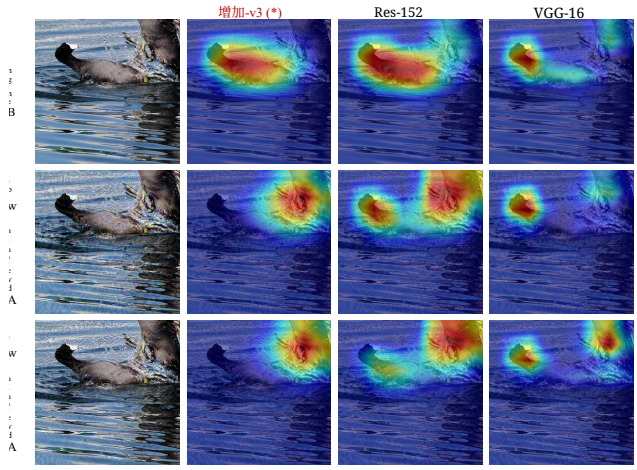


图 1：注意力转移的可视化。我们使用 Eigen-CAM [26]可视化由我们的方法在不使用/使用奇异值分解的情况下生成的良性图像（第 1 行）和对抗图像（第 2、3 行）的注意力图。第 1 行展示了三个具有代表性的模型（ Inception V3 、ResNet-152 和 VGG-16）在良性图像上的注意力图。第 2、3 行展示了在源模型（Inception-V3）上生成、并迁移用于攻击目标模型（ResNet-152 和 VGG-16）的对抗图像。结果表明，使用奇异值分解生成的对抗样本更能将视觉注意力从真实的前景物体转移到不同模型共有的无关区域。

与白盒攻击相比，白盒攻击中攻击者已知模型结构和参数，而黑盒攻击无法获取这些信息，因此更适用于现实应用。此外，对抗样本更棘手的问题是其良好的可迁移性 [33]。因此，许多研究 [6, 37, 41] 利用可迁移性攻击其他未知的黑盒模型。此外，还有另一种黑盒攻击方法，即基于查询的攻击，通过多次查询来近似梯度信息，从而生成对抗样本。本研究专注于提升基于迁移的黑盒攻击的可迁移性。

近年来，人们从多个不同角度开发了多种方法来提高对抗样本的可迁移性，包括基于梯度的方法、基于损失函数的方法、基于输入

基于变换的方法 [1, 41], 和基于特征的方法 [17, 38, 43]。前三类方法主要通过操控模型的输入或输出发挥作用。相比之下，最后一类基于特征的方法会扰乱替代模型中间层的特征图，从而提高可迁移性，因为不同模型通常共享重要特征 [39]。

现有的基于特征的攻击主要旨在扰乱对抗样本相较于原始干净样本的神经元激活。特征破坏攻击（FDA） [10] 通过优化各中间层中的神经元激活，使其趋近于平均激活值来实施攻击。FDA 并未将每个神经元一视同仁，也有一些方法尝试考虑每个神经元对攻击的重要性。特征重要性感知攻击（FIA） [38] 通过计算多个随机掩蔽输入的梯度来估计重要性。基于神经元归因的攻击（NAA） [43] 则基于神经元归因方法 [5]计算神经元重要性。此外，Wu 等人 [39] 利用注意力图衡量特征重要性，并通过扰乱注意力图来正则化优化过程。然而，这些方法通常需要大量额外计算来获取神经元或特征重要性。

另一方面，近期一项关于域适应的研究 [3]从频谱角度分析了深层特征，发现较大的奇异值所对应的特征向量对不同域之间的泛化贡献最大。此外，Muhammad 等人 [26] 提出了 Eigen-CAM，用于可视化最后一个卷积层的注意力图。通过将特征图投影到通道维度上最大的特征向量，激活图能够稳健且可靠地定位前景目标，如图 1 （第 1 行）所示。因此，我们假设，最大奇异值所对应的特征向量能够凭借其优越的泛化能力和注意力属性提升对抗迁移性。

受上述发现启发，我们提出了一种利用奇异值分解（SVD）的新型基于特征的攻击方法。具体而言，我们的方法从替代卷积神经网络模型中选择一个具有充分语义信息的中间层，并通过 SVD 对其特征图进行分解 $X = USV^T$。随后，我们仅保留奇异值最大的分解后 Top-1 特征 $Z = s_1 u_1 v_1^T$，并将其与原始特征并行 X 传入所有后续层，分别获得相应的输出逻辑值。最后，将这两个逻辑值混合，以计算损失函数，并通过反向传播更新对抗样本。比较图 1中的第 2 行和第 3 行可以发现，采用 SVD 学得的对抗样本比未采用 SVD 的原始对抗样本具有更好的可迁移性。为评估所提方法的有效性，我们使用不同的训练基线，在白盒和盒盒设置下进行了大量实验。

总而言之，本研究的主要贡献如下：

- 我们提出了一种基于奇异值分解的简单有效的特征分解对抗攻击方法，通过可靠地定位卷积神经网络模型中的关键特征来增强可迁移性。此外，所提出的特征分解攻击方法可以轻松集成到其他攻击方法中，以提升其性能。
- 我们进行了大量实验来验证该攻击方法的有效性。实验结果表明，该方法可以显著提高对抗样本对正常训练的卷积神经网络模型和防御策略的可迁移性。

2 相关工作

2.1 对抗攻击

过去几年，研究者提出了许多对抗攻击方法 [1, 2, 6, 9, 13, 21, 25, 37, 43] ，用于生成具有破坏性的对抗样本。此外，对抗样本具有黑盒可迁移性，即针对某个源模型生成的对抗样本可以重复用于欺骗其他未知模型。因此，这推动了黑盒攻击的研究。为了生成更具破坏性且可迁移性更强的对抗样本，研究者开发了许多侧重于不同方面的方法。我们将它们归为以下几类。

基于梯度的方法：这类方法通过先进的基于梯度的优化来提高可迁移性。FGSM [13]是最早的基于梯度的攻击方法，它通过最大化交叉熵损失函数，在单步优化中直接生成对抗样本。I-FGSM [21] 将 FGSM 扩展为迭代更新方法，在梯度方向上以较小的步长反复添加扰动。MI-FGSM [6]将动量项融入 I-FGSM，以促进优化方向的稳定。类似地，NI-FGSM [22]结合 Nesterov加速梯度（NAG），在训练迭代过程中累积梯度。投影梯度下降（PGD） [24]是 FGSM 的一种变体，它先在允许的范数球内进行随机初始化，然后采用 I-FGSM 的优化策略。LinBP [15] 通过跳过部分非线性激活函数，从更平滑的表面近似梯度，从而减少替代模型中的非线性路径数量。Huang 等人 [18]引入积分梯度以提高可迁移性，该方法利用从黑色参考图像到输入图像的梯度线积分来优化扰动。

基于损失的方法：一些研究着重于开发新的损失函数，通过优化生成对抗样本。C&W损失 [2] 直接扩大良性样本与对抗样本的逻辑值之间的间隔距离。在 [28], 中，提出了相对交叉熵（RCE）损失，以确保对抗样本与干净样本之间的置信度具有相对较大的间隔。 [44] 观察到使用交叉熵损失构造目标对抗样本时存在饱和问题，并提出了一种Logit损失，在训练时将目标逻辑值的负值作为损失。为了学习具有破坏性的通用对抗扰动（UAP）， [42] 通过最小化真实样本与其对应对抗样本之间的余弦相似度，利用余弦相似度损失进行优化。

基于输入变换的方法： 为进一步提升可迁移性，已有多种输入变换方法被提出，用于在输入层面扩增样本。例如，多样化输入（DI） [41] 在每次训练迭代中对输入图像进行随机缩放与填充。平移不变（TI） [7] 发现，对输入图像进行平移以获得一组对抗图像，可以提升攻击效果。为进一步提高效率，TI 方法通过对输入图像的梯度进行卷积来实现，

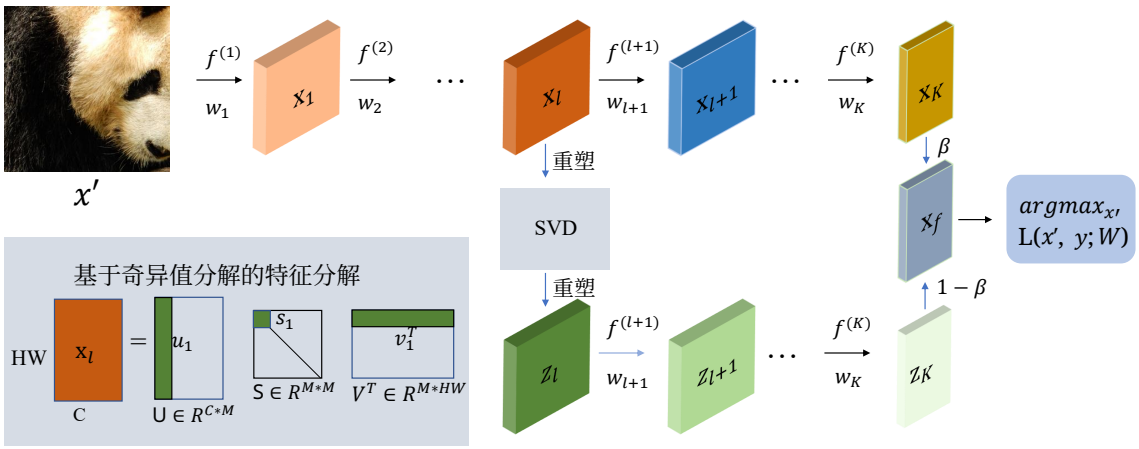


图 2：我们提出的基于SVD的特征分解对抗攻击方法的整体框架。X_f是最终逻辑值，保留了由分解后的 Top-1 奇异值相关特征 Z_l 计算得到的逻辑值 Z_k 以及原始逻辑值 X_k。

并使用预定义核。尺度不变（SI）[22] 在输入图像的多个尺度副本（$S_i(x) = x/2^i$）上学习对抗扰动。基于对象的多样化输入（ODI）[1] 将输入图像投影到多个三维对象上，以进行数据增强。此外，Long等人 [23]提出了一种用于训练的频域数据增强方法，这可以显著提升可迁移性。这些基于输入变换的方法是附加步骤，可直接集成到其他对抗攻击方法中。

基于特征： 不同于主要基于卷积神经网络模型的最终决策生成对抗样本，一些研究尝试利用中间层特征来提升可迁移性。中间层级攻击 [17] 采用两步优化来增强可迁移性。第一步执行标准攻击，第二步沿着前一步的优化方向，增大卷积神经网络模型预先指定层上的扰动。特征破坏攻击（FDA）[10] 通过破坏卷积神经网络各层的特征来生成对抗样本。基于不同卷积神经网络模型的分类注意力图非常相似这一观察，注意力引导迁移攻击（ATA）[39] 通过进一步破坏注意力图来提升对抗样本的可迁移性。特征重要性感知攻击（FIA）[38]使用多个随机掩码输入图像，根据中间层的累积梯度计算特征重要性，然后通过破坏加权特征图来计算扰动。基于神经元归因的攻击（NAA）[43] 根据神经元属性计算中间层的神经元重要性，并通过对神经元加权来执行特征级攻击。

与其他基于特征的方法相比，我们的方法实际上不需要显式计算不同神经元的特征重要性。相反，SVD分解能够隐式识别有助于提升可迁移性的重要特征。此外，我们提出的方法是一种附加组件，可以融入各种不同的攻击方法中，以提升其性能。

2.2 对抗防御

近年来，人们提出了许多方法来缓解对抗攻击带来的安全问题。这些方法主要分为基于对抗训练的方法和输入预处理方法。基于对抗训练的方法 [24]是提升模型鲁棒性的一种简单有效的方式：将对抗样本与干净训练样本一起扩充训练集，再对模型进行重新训练。另一方面，输入预处理方法主要旨在将图像输入卷积神经网络之前，消除对抗扰动的影响，例如，JPEG压缩 [14], 随机缩放与填充（R&P）[40], 以及高层表示引导去噪器（HGD）[22]。此外，为了进一步提升防御效果，许多研究尝试结合对抗训练和输入预处理方法的优势，例如，NeurIPS-r3方案 [34] 和神经表示净化器（NRP）[27]。在本研究中，我们利用这些防御方法来验证所提出的对抗攻击的破坏性。

3 预备知识

我们将用于图像分类的卷积神经网络（CNNs）视为函数 f，它可表示为一系列连续函数 $f = f(K) \circ f(K-1) \circ \cdots \circ f(1)$，其参数为 $W = (\mathbf{w}_K, \mathbf{w}_{k-1}, \cdots, \mathbf{w}_1)$。其中，$f(l)$ 是第 l 层，$\mathbf{w}_l$ 是其对应参数。K 表示层的总数。此外，对于输入图像 x，我们将第 l 层的输出记为 $\mathbf{X}_l$。

给定卷积神经网络模型 f、带有真实标签 y的干净图像 x，我们的主要目标是构造一个对抗样本 x'，使分类器 f产生意料之外的预测。为确保对原始图像 x 施加的扰动尽可能小，通常会将对抗样本限制在以 x 为中心、半径为 ϵ的 l_p-范数球内。遵循先前的研究 [11, 13, 41], ，本文采用广受推崇的 l_∞-范数。更正式地说，可以将其概括如下：

$$\arg\max f(x') \neq y, \quad \text{s.t.} \quad ||x' - x||_\infty \leq \epsilon. \tag{1}$$

设 $\mathcal{L}(x_{\prime}y;W)$ 表示式 1中用于构造对抗样本的损失函数，其形式如下：

$$\arg\max_{x'} \mathcal{L}(x', y; W), \tag{2}$$

其中，L($x_{\prime}y;W$) 通常表示交叉熵损失。

在现实场景中，攻击者无法访问受害模型 f 的参数 W。按照常见做法，我们利用替代模型生成的对抗样本的可迁移性来欺骗受害模型。此外，我们采用先进的基于梯度的优化方法来增强可迁移性，例如迭代快速梯度符号法（I-FGSM）。I-FGSM 以较小的步长 α 沿梯度方向迭代添加扰动：

$$x'_0 = x, \quad x'_{i+1} = x'_i + \alpha \cdot \text{sign}(\nabla_{x'} \mathcal{L}(x'_i, y; W)), \tag{3}$$

其中，$x_{\prime i}$ 表示第 i_{th} 次迭代中的对抗图像，而 $\alpha = \epsilon/T$ 确保在最多进行 T 次迭代优化时，$x_{\prime}$ 的范数不超过 l_p。

4 方法

本节介绍我们提出的基于 SVD 的攻击方法。我们从中间层 l 提取 Top-1 奇异值特征，以获得更具泛化性的 信息，从而提升对抗样本的可迁移性。最后，我们介绍用于优化的损失函数。

4.1 整体架构

我们提出的特征分解攻击框架的整体架构如图 2 所示。对于输入图像 x，我们将其输入替代卷积神经网络模型，计算每一层的特征 $f_{(i)}$。在中间层 l，我们通过奇异值分解对相应的特征 X_l 进行分解，并获得具有最大奇异值的 Top-1 特征 Z_l。随后，原始特征 X_l 和 Top-1 特征Z_l 同时传入卷积神经网络的其余层，分别在两个并行分支中计算相应的输出 X_k 和 Z_k。最后，我们根据 X_k 和 Z_k计算攻击损失函数，并通过梯度反向传播更新对抗样本。

4.2 基于奇异值分解的特征分解

近期一项研究 [3] 从谱分析的角度研究了域适应任务中的深层特征。通过对特征表示进行奇异值分解，较大的奇异值所对应的特征向量通常有助于提升跨域分类的泛化能力。另一方面，[26] 提出了一种 Eigen-CAM，用于可视化卷积神经网络中学到的特征。通过将最后一个卷积层的特征图投影到其通道维度上最大的特征向量，激活图无需进行反向传播计算，即可稳健、可靠地定位前景目标。

为了增强对未知黑盒卷积神经网络的攻击可迁移性，建议解决对抗样本偏向训练代理卷积神经网络所导致的过拟合问题。受 [3, 26]，中研究结果的启发，我们还采用基于奇异值分解的特征分解来增强对抗迁移性。具体计算过程如下。

给定中间层 l 的特征图 $X_l \in R^{C\times H\times W}$，我们首先将其重塑为大小为 $X_l \in R^{C\times HW}$ 的矩阵。令 $M = \min(C, HW)$，然后我们进行奇异值分解，将特征分解为：

$$X_l = USV^T, \tag{4}$$

其中，$S \in R^{M\times M}$ 是矩形对角奇异值矩阵，$U \in R^{C\times M}$ 是左正交奇异向量矩阵，$V \in R^{HW\times M}$ 是右正交奇异向量矩阵。

分解后，我们根据最大的奇异值 s_1 计算 Top-1 分解特征 Z_l：

$$Z_l = s_1 u_1 v_1^T, \tag{5}$$

其中，u_1 和 v_1 分别是对应的左奇异向量和右奇异向量。将 Z_l 重塑回 $Z_l \in R^{C\times H\times W}$ 后，进一步用于计算分解后的逻辑值。此外，我们评估了使用不同 Top-k 奇异值获取分解特征的攻击性能，即，$Z_l = \sum_{i=1}^{k} s_i u_i v_i^T$。

4.3 攻击算法

由于我们提出的基于奇异值分解的特征分解可集成到不同的对抗基线方法中，以生成对抗样本，因此本节仅介绍本研究中用于学习对抗扰动的损失函数。

获得 Top-1 分解特征 Z_l后，我们将其与 X_l 并行输入卷积神经网络剩余的各层，从而分别得到输出逻辑值 Z_k 和 X_K。随后，我们将二者合并以得到最终逻辑值：

$$X_f = \beta X_K + (1-\beta) Z_k, \tag{6}$$

其中，β 是一个超参数，用于控制分解特征的影响。

基于融合后的逻辑值 X_f，我们可以通过交叉熵计算最终的损失函数：

$$\mathcal{L}(x', y; W) = -\mathbb{1}_y^T \cdot \log\left(\text{softmax}(X_f)\right), \tag{7}$$

其中，y 是输入 x的真实标签。

5 为什么 SVD方法有效?

本节将分析为何将奇异值分解用于生成对抗样本能够提升攻击性能。表征相似度（RS）指标已被广泛用于定量比较给定数据集上不同神经元集合的激活值。遵循 [4], ，我们采用线性中心核对齐（CKA）来衡量干净样本与对抗样本之间的 RS。线性 CKA 的计算方法如下：

$$\text{CKA}_L(X, Y) = \frac{||Y^T X||_F^2}{||X^T X||_F ||Y^T Y||_F}, \tag{8}$$

其中，$X \in \mathbb{R}^{n\times d_1}$ 和 $Y \in \mathbb{R}^{n\times d_2}$ 是一个包含 n 个样本的集合中两组特征激活值。d_1 和 d_2 是相应的特征维度。为计算线性 CKA，我们使用 Inc-v3 作为源模型，为 ImageNet-Compatible 数据集中的 1000 张图像生成对抗样本。优化过程中采用 S^2I-TI-DIM [23] 。更多训练细节请参见实验部分。

首先，我们计算源 Inc-v3 模型中干净样本与对抗样本特征之间的线性 CKA。在

表 1：源 Inc-v3 模型不同层计算得到的干净样本与对抗样本特征之间的线性 CKA。

	二维卷积_4a	混合层_5d	混合层_6e	混合层_7c	Pool	FC
w/o	0.9279	0.7183	0.6138	0.2765	0.1278	0.0885
w/	0.9251	0.6902	0.4697	0.2120	0.1140	0.0731

表2：将由 Inc-v3 生成的对抗样本迁移至 ResNet-152 和 VGG-16 时，特征之间的线性 CKA。

	Inc-v3 →ResNet-152			Inc-v3 → VGG-16		
	干净样本	对抗样本（无）	对抗样本（有）	干净样本	对抗样本（无）	对抗样本（有）
Pool	0.3769	0.2066	0.2273	0.2160	0.1107	0.1165
FC	0.4407	0.2182	0.2447	0.3196	0.1793	0.1896

层 l处，我们将所有干净样本、对抗样本（无奇异值分解）和对抗样本（有奇异值分解）的特征分别记为 $\boldsymbol{X} \in \mathbb{R}^{1000\times d_l}$、$\tilde{\boldsymbol{X}} \in \mathbb{R}^{1000\times d_l}$和 $\tilde{\boldsymbol{X}}_{svd} \in \mathbb{R}^{1000\times d_l}$，其中 d_l表示层 l的特征维度。然后，我们计算 $\mathrm{CKA}_L(\boldsymbol{X}, \tilde{\boldsymbol{X}})$和 $\mathrm{CKA}_L(\boldsymbol{X}, \tilde{\boldsymbol{X}}_{svd})$。如表 1所示，干净样本与对抗样本（有奇异值分解）的表征相似度远低于无奇异值分解时的相似度，这种差异在较高层尤为明显。这些结果表明，使用奇异值分解生成的对抗样本在攻击白盒模型时具有更强的破坏性。

此外，根据 [4], “高可迁移性表明目标网络学到了相似的决策边界”。因此，与分类相关的表示越相似，在源模型上训练的对抗样本就越容易迁移到目标模型。具体而言，我们计算了最后一个全局池化层和全连接层的线性 CKA。由于不同模型的架构各不相同，这些层与分类高度相关。在这部分中，我们直接使用由 Inc-v3 生成的对抗样本，计算 Inc-v3、ResNet-152 和 VGG-16 的特征。如表2所示，对于 1000 个干净样本，Inc-v3 与 ResNet-152 在池化层和全连接层的线性 CKA 分别为 0.3769 和 0.4407。对抗样本（无奇异值分解）的线性 CKA 会降至 0.2066 和 0.2182。另一方面，我们发现，对抗样本（有奇异值分解）的线性 CKA 会略微升至 0.2273 和 0.2447，这意味着从 Inc-v3 迁移到 ResNet-152 的对抗样本（有奇异值分解）的表示，比对抗样本（无奇异值分解）的表示更加相似。此外，在 Inc-v3 与 VGG-16 之间也能发现类似结果。这些结果表明，我们提出的基于奇异值分解的优化方法能够增强源模型与目标模型之间对抗样本中与分类相关的表示相似性，从而提高可迁移性。

通过使用线性 CKA 分析表示相似性，我们可以观察到，使用奇异值分解生成对抗样本能够有效扰乱白盒模型中的特征激活。此外，这也能提升对抗样本攻击黑盒模型时的可迁移性。

6 实验

6.1 实验设置

数据集： 遵循先前的研究 [6, 23], , 我们在具有挑战性的 ImageNet-Compatible 数据集[1]上开展实验。该数据集包含 1,000 张图像，每张图像对应一个唯一的类别标签，共有 1,000 个类别，这些类别与原始 ImageNet 数据集中的类别相对应。

模型： 在本研究中，我们使用四种经过常规训练的 ImageNet 卷积神经网络模型作为源模型来制作对抗样本，即 Inception-v3 (Inc-v3) [32], Inception-v4 (Inc-v4) [31], Inception-Resnet-v2 (IncRes-v2) [31] 和 Resnet-v2-152 (Res-152) [16]。为了评估对抗样本的性能，我们使用这四种训练源模型以及 Resnet-v2-50 (Res-50) 和 Resnet-v2-101 (Res-101) 作为受害模型。此外，为了进一步评估对防御模型的攻击效果，我们选择了四种基于 PGD-AT [24]和集成对抗训练 [**的对抗训练模型?**],, 后者通过使用其他静态预训练模型生成的对抗样本扩充模型的训练数据，从而使模型对黑盒攻击具有更强的鲁棒性。具体而言，对抗训练模型包括一个经过对抗训练的 Inception-v3 模型 (Inc-v3$_{adv}$)、三个经过对抗训练的 Inception-v3 模型组成的集成 (Inc-v3$_{ens3}$)、四个经过对抗训练的 Inception-v3 模型组成的集成 (Inc-v3$_{ens4}$)，以及三个经过对抗训练的 Inception-Resnet-v2 模型组成的集成 (IncRes-v2$_{ens}$)。此外，我们还选择了涵盖随机缩放与填充 (R&P) [40], NIPS-r3 [34], HGD [22], JPEG [14],随机平滑 (RS) [19] 和 NRP [27]) 的先进防御方法；这些方法能够抵御黑盒攻击。

基线方法： 为了全面评估性能，我们将本方法融入多种知名的对抗攻击基线方法，包括 DI-FGSM [41], TI-FGSM [7], MI-FGSM [6],VT-FGSM [35], SI-NI-FGSM [22], PI-TI-DI-FGSM [12], 和 Admix [36]。此外，我们还将本方法融入一种当前的 SOAT 攻击方法S^2-I-FGSM [23],, 该方法采用基于频率的频谱变换进行数据增强。此外，我们将这些简单的基线方法组合起来，构建更强的扰动攻击基线方法。

参数： 在所有实验中，最大扰动为 $\epsilon = 16$，迭代次数为 $T = 10$，步长为 $\alpha = \epsilon/T$。对于 MI- FGSM [6], , 我们将衰减因子设为 $\mu = 1.0$。对于 DI-FGSM [41], , 我们将变换概率设为 $p = 0.5$。对于 TI-FGSM [7], , 我们将核长度设为 $k = 7$。对于 VT-FGSM [35], , 我们将超参数设为 $\beta = 1.5$，采样样本数为 20。对于 SI-NI-FGSM [22], , 我们将副本数设为 $m_1 = 5$。对于 Admix [36], , 我们将副本数设为 $m_1 = 5$，样本数设为 $m_2 = 3$，Admix 比率设为 $\eta = 0.2$。对于 PI-TI-DI-FGSM [12],, 我们将放大因子设为 $\beta = 10$，投影因子设为 $\gamma = 16$，核长度设为 $k_w = 7$。对于 S^2-I-FGSM [23], , 我们将 M 的调节因子设为 $\rho = 0.5$，频谱变换次数设为 $N = 20$。对于我们提出的基于 SVD 的攻击，Top-k 分解特征为 $k = 1$，超参数为 $\beta = 0.5$。用于基于奇异值分解的特征分解的中间层为 Inception-v3 (Incv3) 的 Mixed-6e，即最后的 Inception-B 模块

[1]https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/ examples/nips17_对抗_competition/dataset

表3：六种常规训练模型上的攻击成功率（%）。对抗样本分别通过 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 构造。

模型	攻击	SVD	Inc-v3	Inc-v4	IncRes-v2	Res-152	Res-50	Res-101	Avg
Inc-v3	MI-FGSM	w/o	100.0	50.6	45.8	41.2	47.1	42.1	54.5
		w/	100.0	55.2	52.8	44.9	49.9	46.5	**58.2**(↑ 3.7)
	DI-FGSM	w/o	99.9	49.6	37.7	32.7	37.4	33.5	48.5
		w/	99.5	53.1	44.7	34.2	43.0	37.6	**52.0**(↑ 3.5)
	S2I-FGSM	w/o	99.7	63.6	57.5	49.9	56.5	52.0	63.2
		w/	99.8	69.8	65.9	57.2	64.2	60.1	**69.5**(↑ 6.3)
	SI-NI-FGSM	w/o	100.0	76.9	74.7	68.0	72.2	69.1	76.8
		w/	100.0	80.2	79.1	70.7	76.2	72.3	**79.8**(↑ 3.0)
	VT-MI-FGSM	w/o	100.0	75.0	69.9	62.6	66.8	62.3	72.8
		w/	99.8	80.6	78.7	69.0	73.7	70.6	**78.7**(↑ 5.9)
Inc-v4	MI-FGSM	w/o	60.7	99.9	45.5	42.3	47.7	43.5	56.6
		w/	63.5	100.0	48.7	44.9	50.1	46.3	**58.9**(↑ 2.3)
	DI-FGSM	w/o	52.5	99.2	36.0	29.9	32.7	30.5	46.8
		w/	56.0	99.5	40.2	32.6	39.2	33.7	**50.2**(↑ 3.4)
	S2I-FGSM	w/o	71.5	99.6	55.4	49.0	57.1	48.9	63.6
		w/	76.1	99.6	62.6	54.3	60.4	55.3	**68.1**(↑ 4.5)
	SI-NI-FGSM	w/o	86.0	100.0	79.1	73.6	77.1	73.7	81.6
		w/	87.1	100.0	81.4	75.3	79.6	75.9	**83.2**(↑ 1.6)
	VT-MI-FGSM	w/o	80.1	99.8	70.6	64.2	65.8	63.8	74.1
		w/	85.0	99.9	76.5	69.9	71.4	68.6	**78.6**(↑ 4.5)
IncRes-v2	MI-FGSM	w/o	60.2	53.4	99.2	44.1	49.6	46.7	58.9
		w/	71.5	61.2	99.7	50.8	57.4	54.8	**65.9**(↑ 7.0)
	DI-FGSM	w/o	56.3	51.6	97.7	37.1	39.8	36.3	53.1
		w/	64.2	58.8	98.9	39.6	46.0	41.9	**58.2**(↑ 5.1)
	S2I-FGSM	w/o	77.3	68.0	98.4	56.1	62.4	58.1	70.1
		w/	80.0	72.3	99.1	60.9	66.6	62.6	**73.6**(↑ 3.5)
	SI-NI-FGSM	w/o	86.6	84.0	99.8	76.5	80.0	77.5	84.1
		w/	90.0	87.8	100.0	81.8	84.6	83.6	**88.0**(↑ 3.9)
	VT-MI-FGSM	w/o	81.1	76.6	99.0	66.7	70.1	68.7	77.0
		w/	88.0	84.9	99.7	77.1	79.2	78.4	**84.6**(↑ 7.6)
Res-152	MI-FGSM	w/o	56.3	50.8	46.0	99.4	82.9	86.2	70.3
		w/	63.9	53.9	50.1	99.4	88.0	89.1	**74.1**(↑ 3.8)
	DI-FGSM	w/o	55.9	53.2	46.3	99.6	83.2	84.3	70.4
		w/	61.4	56.9	50.4	99.5	86.3	88.5	**73.8**(↑ 3.4)
	S2I-FGSM	w/o	66.6	63.3	57.6	99.8	93.5	93.9	79.1
		w/	71.6	64.9	62.5	99.8	95.2	95.2	**81.5**(↑ 2.4)
	SI-NI-FGSM	w/o	76.3	72.5	72.6	99.8	94.2	95.3	85.1
		w/	78.5	76.5	75.9	99.9	95.5	95.9	**87.0**(↑ 1.9)
	VT-MI-FGSM	w/o	73.7	68.7	65.8	99.5	93.2	93.8	82.5
		w/	80.2	75.9	73.4	99.5	93.9	95.7	**86.4**(↑ 3.9)

分别为 Inception-v4（Inc-v4）的混合层-6b、Inception-Resnet-v2（IncRes-v2）的混合层-6a，以及 Resnet-v2-152（Res-152）第4层中 block3 的最后一层。

6.2 攻击通常训练的模型

本节中，我们通过使用五种不同的基线方法训练对抗样本，验证我们的方法对常规训练模型的有效性。根据表 3中的结果，我们可以得出以下发现：**1**）对于每个源模型，比较时性能都有显著提升

采用奇异值分解与未采用奇异值分解的不同基线方法结果。例如，当 Inc-v3 用作源模型时，平均攻击成功率分别提高了 3.7%（MI-FGSM）、3.5%（DI-FGSM）、6.3%（S^2I-FGSM）、3.0%（SI-NI-FGSM）和 5.9%（VT-MI-FGSM）。**2**）攻击黑盒模型时，应用奇异值分解后，对抗迁移性也显著提升。例如，当 VT-MI-FGSM 和 IncRes-v2 分别用作基线方法和替代模型时，针对 Inc-v3、Inc-v4、Res-152 和 Res-50 的迁移攻击成功率分别从 81.1% 提高到 88.0%、从 76.6% 提高到 84.9%、从 66.7% 提高到 77.1%，以及从 70.1% 提高到 79.2%。

表 4：在十种防御方法上的攻击成功率（%）。对抗样本通过 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 生成。

模型	攻击	SVD	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	NIPS-r3	JPEG	RS	NRP	AVG.
Inc-v3	TI-DIM	w/o	45.5	43.1	43.2	28.6	36.2	30.0	35.4	54.6	36.2	24.9	37.8
		w/	49.2	48.4	45.5	32.0	40.0	34.3	41.2	59.8	39.0	24.3	41.4
	PI-TI-DI-FGSM	w/o	44.6	44.0	46.0	33.4	32.4	33.7	37.0	46.7	64.7	36.9	41.9
		w/	48.0	46.6	47.9	35.8	34.2	36.5	40.3	50.2	66.7	38.9	44.5
	VT-TI-DIM	w/o	64.3	65.0	64.3	47.6	58.8	51.7	57.6	74.9	42.6	35.2	56.2
		w/	67.9	69.4	67.3	51.9	61.7	53.5	62.3	80.2	45.1	37.8	59.7
	Admix-TI-DIM	w/o	77.0	75.6	73.9	60.0	69.6	62.2	70.3	84.7	55.9	46.3	67.6
		w/	81.6	82.7	79.3	64.8	73.4	57.5	75.5	88.2	58.9	47.3	70.9
	S^{2}I-TI-DIM	w/o	82.8	81.9	79.2	68.8	77.1	70.0	75.9	87.2	61.2	57.2	74.1
		w/	85.8	84.8	83.5	72.9	80.6	75.1	80.1	88.9	64.4	59.1	77.5
Inc-v4	TI-DIM	w/o	38.8	38.9	37.1	29.1	36.3	29.9	33.4	48.7	35.9	19.2	34.7
		w/	41.3	42.3	41.6	29.7	37.0	30.8	36.2	53.3	37.5	21.3	37.1
	PI-TI-DI-FGSM	w/o	40.6	41.6	44.2	34.4	31.7	33.1	32.0	45.4	63.5	34.2	40.1
		w/	43.6	44.8	45.8	35.2	33.4	35.6	37.8	47.7	65.5	36.9	42.6
	VT-TI-DIM	w/o	57.9	61.4	59.7	48.4	58.8	51.9	55.1	70.8	41.3	32.0	53.7
		w/	61.3	64.7	62.7	52.8	61.8	53.6	59.1	76.1	42.0	36.1	57.0
	Admix-TI-DIM	w/o	77.3	78.7	74.5	65.7	75.8	69.2	74.0	83.7	54.5	49.1	70.3
		w/	78.7	81.2	77.0	69.0	76.7	72.3	75.8	86.2	56.0	47.6	72.1
	S^{2}I-TI-DIM	w/o	80.2	80.8	77.7	71.0	77.2	72.4	76.4	84.7	60.4	56.5	73.7
		w/	82.1	82.2	78.8	72.9	78.3	74.3	78.2	86.3	63.6	59.0	75.6
IncRes-v2	TI-DIM	w/o	43.1	46.4	40.9	39.4	43.4	39.0	41.5	56.6	36.7	22.2	40.9
		w/	52.3	56.8	50.2	41.6	50.2	45.2	50.9	68.6	40.4	26.9	48.3
	PI-TI-DI-FGSM	w/o	47.8	48.7	49.5	45.6	39.4	45.8	46.5	51.1	66.3	40.9	48.2
		w/	54.9	54.7	56.1	50.3	45.7	50.2	51.6	56.1	70.3	45.8	53.6
	VT-TI-DIM	w/o	61.8	66.4	61.7	59.5	64.0	61.4	62.1	72.6	43.4	35.4	58.8
		w/	72.8	78.0	71.7	67.8	72.4	69.7	73.1	84.7	47.4	42.8	68.0
	Admix-TI-DIM	w/o	83.5	84.8	83.4	79.0	82.9	81.6	82.3	88.5	64.0	61.2	79.1
		w/	88.0	89.3	87.5	85.1	86.4	84.6	86.8	92.4	65.5	62.2	82.8
	S^{2}I-TI-DIM	w/o	81.8	81.8	78.5	78.6	79.9	79.1	80.4	85.0	63.0	60.7	76.9
		w/	88.7	88.7	85.7	85.0	86.2	85.2	87.1	91.0	70.5	66.3	83.4
Res-152	TI-DIM	w/o	55.7	55.9	52.5	43.4	55.3	46.5	51.9	66.0	46.5	31.4	50.5
		w/	58.7	58.7	55.3	43.6	56.4	47.2	55.6	70.7	48.4	33.0	52.8
	PI-TI-DI-FGSM	w/o	53.1	52.7	55.4	46.8	42.2	46.8	51.0	54.3	71.2	47.2	52.1
		w/	56.0	57.2	59.1	49.1	47.4	49.4	51.6	58.5	73.0	46.9	54.8
	VT-TI-DIM	w/o	67.9	73.4	69.3	60.5	70.4	64.9	68.3	80.7	52.7	43.0	65.1
		w/	74.2	77.1	72.6	64.8	74.8	67.7	71.8	83.2	55.6	46.8	68.9
	Admix-TI-DIM	w/o	79.0	83.0	78.7	70.4	80.2	74.8	78.3	86.4	65.2	57.7	75.4
		w/	81.1	86.1	82.1	73.2	81.4	77.5	81.0	88.4	67.4	59.3	77.8
	S^{2}I-TI-DIM	w/o	86.8	86.6	83.9	79.0	85.1	80.9	85.3	89.3	72.6	68.6	81.8
		w/	88.8	87.8	85.2	79.9	85.9	83.1	85.8	91.3	75.0	69.5	83.2

这些发现验证了我们提出的方法在提升针对常规训练模型的攻击性能方面的有效性。此外，与现有攻击方法结合时，该方法还能进一步增强对抗样本的黑盒迁移性。

6.3 攻击防御模型

为了进一步评估我们方法的有效性，我们评估了针对十种不同防御策略的攻击成功率，其中包括四种对抗训练模型（Inc-v3$_{adv}$、Inc-v3$_{ens3}$、Inc-v3$_{ens4}$、IncRes-v2$_{ens}$）以及六种更先进的方法（HGD、R&P、NIPS-r3、随机缩放、JPEG、NRP）。如表4所示，我们发现，所提出的方法在采用不同训练基线的源模型上均显著提升了性能。例如，采用 TI-DIM、PI-TI-DI-FGSM、VT-TI-DIM、Admix-TI-DIM 和 S^{2}I-TI-DIM 攻击 Inc-v3 的平均成功率分别提升了 3.6%、2.6%、3.5%、3.4% 和 3.4%。结果表明，Top-1 分解特征对最终预测有显著贡献。然而，当前的防御策略并未利用这些特征来增强鲁棒性。

6.4 计算成本分析

我们以 Inc-v3 为源模型，对 ImageNet-Compatible 数据集中的 1000 张图像发起攻击，并计算不同基线方法所需的额外计算成本。结果如表 8 所示。从表中可以看出，我们提出的奇异值分解

方法会使不同基线方法的额外计算成本增加约 25%–30%。

6.5 对抗样本的图像质量

我们进一步计算了 [?] 使用不同基线方法（无论是否采用我们提出的方法）时，基于 Inc-v3 模型生成的原始样本与对抗样本之间的结构相似性指数（SSIM）分数。SSIM 是一种常用的图像质量评估指标，SSIM 分数越高，表示图像质量越好、失真越小。如表6所示，我们发现，未采用和采用我们提出的方法时得到的 SSIM 分数非常接近。这表明，我们提出的方法不会显著改变原始样本与对抗样本之间的感知相似性。

6.6 消融研究

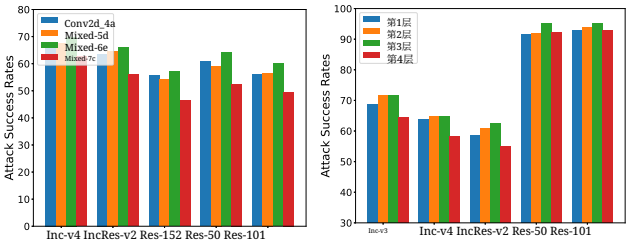
分解目标层：首先，我们评估使用不同层进行 SVD 特征分解时的性能。在图3 (a) 和 (b) 中，我们分别绘制了以 Inc-v3 和 Res-152 为源模型时的迁移攻击成功率。对于每个源模型，我们比较了四个卷积块中具有不同特征分辨率的四个层。如图3所示，Inception-v3 在 Mixed-6e 层、ResNet-152 在第3层时性能最佳。这一现象的主要原因是，前两层的分辨率较高，且语义信息较少。Mixed-7c 和第4层分别是 Inc-v3 和 Res-152 中最后的卷积层，泛化能力较弱。此外，最后一层中具有最大奇异值的分解特征与原始特征较为接近，从而削弱了 logits 组合的作用。

参数：超参数 β 对 Inc-v3 作为源模型时的权衡影响如图4a所示。我们可以发现，攻击成功率随着 β的增加而上升。对于全部五个黑盒模型，最佳整体性能均在 $\beta = 0.5$ 左右达到。当 β 进一步增大时，攻击成功率会下降，这表明仅使用 Top-1 分解特征不足以获得更好的性能。此外，当$\beta = 0$时，未使用原始特征，因此性能显著低于其他取值。当 $\beta = 1$时，未使用基于SVD的特征。因此，这些结果表明，原始特征对于取得更好的攻击性能至关重要。

Top-k： 图中4b，我们展示了使用 Top-k 分解特征 $Z_l = \sum_{i=1}^{k} s_i \boldsymbol{u}_i \boldsymbol{v}_i^T$时的可迁移性表现。与原始特征（即，Top-0）相比，我们发现，在所有模型中，使用 Top-1 分解特征都能显著提高攻击成功率。然而，使用 Top-k （$k > 2$） 分解特征时，性能会开始下降。此外，当 $k >= 5$时，攻击成功率会在某个值附近波动，因为$\sum_{i=1}^{k} s_i \boldsymbol{u}_i \boldsymbol{v}_i^T$会逐渐等同于原始特征。

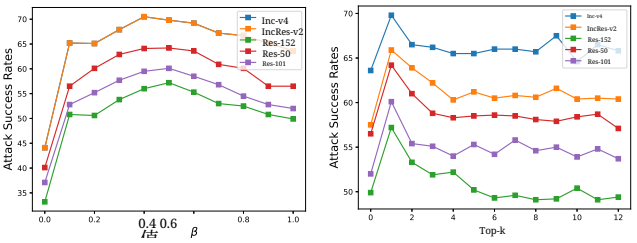
6.7 与基于特征的方法的比较

在本部分中，我们从攻击成功率和训练成本两方面，将我们的方法与其他基于特征的方法进行比较。我们采用与 NAA 相同的设置和数据增强方式制作对抗样本 [43]。根据表7，与



(a) Inception-v3 **(b) ResNet-152**

图 3：使用不同层进行特征分解的黑盒性能。



(a) 参数 β 的敏感性 **(b) Top-k 的影响**

图 4：在黑盒设置下针对 Inception-v3 的不同超参数消融研究。

FDA [10],FIA [38],和 NAA [43]方法相比，我们的方法在攻击成功率方面表现更优。此外，我们比较了在 2080Ti GPU 上以 Inc-v3 作为源模型、针对 1000 张图像进行训练时的总训练成本，并与其他基于特征的方法进行了比较。如表 8所示，我们的方法比 FIA [38], NAA [43], 和 FDA [10], 更高效，训练仅需 1029 秒。

7 结论

本文提出了一种基于 SVD 特征分解的攻击方法。将与最大奇异值相关的分解特征融入攻击中，可以充分发挥使用泛化能力更强的特征的优势，并有针对性地定位前景物体。大量实验表明，所提出的攻击方法能够显著提升攻击黑盒模型和防御策略的迁移能力。另一方面，我们可以考虑保留与最大奇异值对应的干净奇异向量，以构建更稳健的防御模型。

参考文献

[1] Junyoung Byun、Seungju Cho、Myung-Joon Kwon、Hee-Seon Kim 和 Changick Kim。2022。通过基于对象的多样化输入提升定向对抗样本的可迁移性。载于*CVPR*。15244–15253。[2] Nicholas Carlini 和 David Wagner。2017。迈向神经网络鲁棒性的评估。载于*IEEE Symposium on Security and Privacy*。39–57。[3] Xinyang Chen、Sinan Wang、Mingsheng Long 和 Jianmin Wang。2019。可迁移性与可判别性：用于对抗域适应的批量谱惩罚。载于*ICML*。1081–1090。[4] Christian Cianfarani、Arjun Nitin Bhagoji、Vikash Sehwag、Ben Zhao、Heather Zheng 和 Prateek Mittal。2022。通过表征相似性的视角理解鲁棒学习。*NeurIPS* 35（2022），34912–34925。[5] Kedar Dhamdhere、Mukund Sundararajan 和 Qiqi Yan。2018。神经元有多重要？*arXiv preprint arXiv:1805.12233*（2018）。

表5：使用和不使用所提出的基于 SVD 的攻击时，对 ImageNet-Compatible 数据集中的1000张图像进行攻击所需的不同基线方法的训练成本

SVD	MI-FGSM	DI-FGSM	S2I-FGSM	SI-NI-FGSM	VT-MI-FGSM	TI-DIM	PI-TI-DI-FGSM	SI-NI-TI-DIM	VT-TI-DIM	Admix-TI-DIM	S^2I-TI-DIM
w/o	23.39秒	27.60秒	723.84秒	110.09秒	452.02秒	27.83秒	40.58秒	125.96秒	525.35秒	348.98秒	815.24秒
w/	30.02秒	34.98秒	890.65秒	143.37秒	594.03秒	35.16秒	50.23秒	164.40秒	684.05秒	438.35秒	998.17秒

表 6：未经和采用我们提出的方法时，不同基线方法训练得到的对抗样本与原始样本之间的结构相似性指数（SSIM）。

SVD	MI-FGSM	DI-FGSM	S2I-FGSM	SI-NI-FGSM	VT-MI-FGSM	TI-DIM	PI-TI-DI-FGSM	SI-NI-TI-DIM	VT-TI-DIM	Admix-TI-DIM	S^2I-TI-DIM
w/o	0.8336	0.9505	0.9407	0.8313	0.8389	0.8112	0.7410	0.8110	0.8260	0.8079	0.8043
w/	0.8332	0.9500	0.9425	0.8294	0.8617	0.8109	0.7414	0.8093	0.8444	0.8068	0.8034

表 7：将我们的方法与基于特征的最先进方法进行比较。在四种未防御模型和四种对抗训练模型上的攻击成功率（%）。

模型	攻击	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	FDA	81.3	42.9	36.0	35.4	19.3	8.9	6.4	2.3
	FIA	98.3	83.2	79.1	71.6	53.3	36.1	37.0	20.0
	NAA	98.1	85.0	82.4	77.1	61.5	50.5	50.8	31.5
	Ours	**99.8**	**87.8**	**85.2**	**80.7**	**64.2**	**61.2**	**58.8**	**38.6**
Inc-v4	FDA	84.8	99.6	71.9	68.7	27.9	18.4	17.2	7.3
	FIA	84.1	95.7	78.6	72.0	45.3	38.0	37.2	19.4
	NAA	86.0	96.5	81.0	75.5	52.4	50.5	49.4	30.8
	Ours	**90.5**	**99.5**	**86.3**	**80.9**	**60.8**	**65.7**	**63.4**	**48.1**
IncRes-v2	FDA	69.3	67.7	78.3	56.3	36.4	16.2	22.3	17.9
	FIA	81.6	77.1	88.7	71.0	63.8	49.8	46.6	34.1
	NAA	82.4	78.0	93.0	74.4	64.9	60.0	56.7	47.5
	Ours	**92.9**	**92.2**	**99.3**	**88.0**	**76.7**	**78.0**	**74.7**	**70.4**
Res-152	FDA	83.9	84.1	73.9	89.1	51.2	27.9	23.6	11.5
	FIA	83	81.6	78.4	98.9	58.2	49.1	44.9	29.3
	NAA	85.9	**85.9**	**83.6**	98.2	66.1	61.6	59.2	46.7
	Ours	**86.4**	84.8	83.1	**99.9**	**70.7**	**70.4**	**67.6**	**57.2**

表 8：我们的方法与其他基于特征的方法的训练成本比较。

方法	FDA [10]	FIA [38]	NAA [43]	Ours
耗时（秒）	1421	7081	7163	1029

[6] Yinpeng Dong、Fangzhou Liao、Tianyu Pang、Hang Su、Jun Zhu、Xiaolin Hu 和 Jianguo Li。2018。利用动量提升对抗攻击。载于*CVPR*。9185–9193。[7] Yinpeng Dong、Tianyu Pang、Hang Su 和 Jun Zhu。2019。通过平移不变攻击规避针对可迁移对抗样本的防御方法。载于*CVPR*。4312–4321。[8] Yinpeng Dong、Hang Su、Baoyuan Wu、Zhifeng Li、Wei Liu、Tong Zhang 和 Jun Zhu。2019。针对人脸识别的高效基于决策的黑盒对抗攻击。载于*CVPR*。7714–7722。[9] Xia Du 和 Chi-Man Pun。2020。使用多样本和最可能集成方法进行对抗图像攻击。载于*ACM MM*。1634–1642。[10] Aditya Ganeshan、Vivek BS 和 R Venkatesh Babu。2019。FDA：特征破坏攻击。载于*ICCV*。8069–8079。[11] Lianli Gao、Yaya Cheng、Qilong Zhang、Xing Xu 和 Jingkuan Song。2021。通过统计对齐进行特征空间定向攻击。*arXiv*预印本 *arXiv:2105.11645*（2021）。[12] Lianli Gao、Qilong Zhang、Jingkuan Song、Xianglong Liu 和 Heng Tao Shen。2020。用于欺骗深度神经网络的逐块攻击。载于*ECCV*。307–322。[13] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。2015。解释并利用对抗样本。载于*ICLR*。[14] Chuan Guo、Mayank Rana、Moustapha Cisse 和 Laurens Van Der Maaten。2017。使用输入变换抵御对抗图像。*arXiv* 预印本 *arXiv:1711.00117*（2017）。[15] Yiwen Guo、Qizhang Li 和 Hao Chen。2020。通过线性反向传播提升对抗样本的可迁移性。*NeurIPS* 33（2020），85–95。[16] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于*CVPR*。770–778。[17] Qian Huang、Isay Katsman、Horace He、Zeqi Gu、Serge Belongie 和 Ser-Nam Lim。2019。通过中间层级攻击增强对抗样本的可迁移性。载于*ICCV*。4733–4742。[18] Yi Huang 和 Adams Wai-Kin Kong。2022。基于积分梯度的可迁移对抗攻击。*ICLR*（2022）。[19] Jinyuan Jia、Xiaoyu Cao、Binghui Wang 和 Neil Zhenqiang Gong。2019。通过随机平滑认证抵御对抗扰动的 top-k 预测鲁棒性。*arXiv* 预印本 *arXiv:1912.09899*（2019）。[20] Zelun Kong、Junfeng Guo、Ang Li 和 Cong Liu。2020。Physgan：生成能够抵御物理世界攻击、适用于自动驾驶的对抗样本。载于*CVPR*。14254–14263。[21] Alexey Kurakin、Ian J Goodfellow 和 Samy Bengio。2018。物理世界中的对抗样本。载于人工智能安全与保障。99–112。[22] Jiadong Lin、Chuanbiao Song、Kun He、Liwei Wang 和 John E Hopcroft。2020。用于对抗攻击的 Nesterov加速梯度与尺度不变性。载于*ICLR*。[23] Yuyang Long、Qilong Zhang、Boheng Zeng、Lianli Gao、Xianglong Liu、Jian Zhang 和 Jingkuan Song。2022。用于对抗攻击的频域模型增强。载于*ECCV*。549–566。[24] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。2017。迈向能够抵御对抗攻击的深度学习模型。*arXiv* 预印本 *arXiv:1706.06083*（2017）。

[25] Duan Mingxing、Kenli Li、Lingxi Xie、Qi Tian 和 Bin Xiao。2021。通过对抗样本生成网络实现针对多个黑盒的攻击。载于*ACM MM*。264–272。[26] Mohammed Bany Muhammad 和 Mohammed Yeasin。2020。Eigen-CAM：基于主成分的类激活图。载于*IJCNN*。1–7。[27] Muzammal Naseer、Salman Khan、Munawar Hayat、Fahad Shahbaz Khan 和 Fatih Porikli。2020。一种提升对抗鲁棒性的自监督方法。载于*CVPR*。262–271。[28] Muhammad Muzammal Naseer、Salman H Khan、Muhammad Haris Khan、Fahad Shahbaz Khan 和 Fatih Porikli。2019。对抗扰动的跨域可迁移性。*NeurIPS*（2019）。[29] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。2015。Faster R-CNN：借助区域提议网络实现实时目标检测。*NeurIPS*（2015）。[30] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。2015。U-Net：用于生物医学图像分割的卷积网络。载于*MICCAI*。234–241。[31] Christian Szegedy、Sergey Ioffe、Vincent Vanhoucke 和 Alexander A Alemi。2017。Inception-v4、Inception-ResNet 及残差连接对学习的影响。载于*AAAI*。[32] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。2016。重新思考面向计算机视觉的 Inception 架构。载于*CVPR*。2818–2826。[33] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。2013。神经网络的奇妙特性。*arXiv* 预印本 *arXiv:1312.6199*（2013）。[34] Anil Thomas 和 Oguz Elibol。2017。对抗攻击防御——第三名。

[35] Xiaosen Wang 和 Kun He。2021。通过方差调节增强对抗攻击的可迁移性。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集。1924–1933。[36] Xiaosen Wang、Xuanran He、Jingdong Wang 和 Kun He。2021。Admix：增强对抗攻击的可迁移性。载于 *IEEE/CVF*国际计算机视觉会议论文集。16158–16167。[37] Yuxuan Wang、Jiakai Wang、Zixin Yin、Ruihao Gong、Jingyi Wang、Aishan Liu 和 Xianglong Liu。2022。针对视觉变换器生成可迁移的对抗样本。载于 *ACM MM*。5181–5190。[38] Zhibo Wang、Hengchang Guo、Zhifei Zhang、Wenxin Liu、Zhan Qin 和 Kui Ren。2021。特征重要性感知的可迁移对抗攻击。载于 *ICCV*。7639–7648。[39] Weibin Wu、Yuxin Su、Xixian Chen、Shenglin Zhao、Irwin King、Michael R Lyu 和 Yu-Wing Tai。2020。通过注意力增强对抗样本的可迁移性。载于 *CVPR*。1161–1170。[40] Cihang Xie、Jianyu Wang、Zhishuai Zhang、Zhou Ren 和 Alan Yuille。2017。通过随机化缓解对抗效应。*arXiv*预印本 *arXiv:1711.01991*（2017）。[41] Cihang Xie、Zhishuai Zhang、Yuyin Zhou、Song Bai、Jianyu Wang、Zhou Ren 和 Alan L Yuille。2019。通过输入多样性提升对抗样本的可迁移性。载于 *CVPR*。2730–2739。[42] Chaoning Zhang、Philipp Benz、Adil Karjauv 和 In So Kweon。2021。无数据通用对抗扰动与黑盒攻击。载于 *ICCV*。7868–7877。[43] Jianping Zhang、Weibin Wu、Jen-tse Huang、Yizhan Huang、Wenxuan Wang、Yuxin Su 和 Michael R Lyu。2022。通过基于神经元归因的攻击提升对抗迁移性。载于 *CVPR*。14993–15002。[44] Zhengyu Zhao、Zhuoran Liu 和 Martha Larson。2021。论成功与简洁：重新审视可迁移定向攻击。*NeurIPS*（2021），6115–6128。