南京大学
NANJING UNIVERSITY

RINC

面向深度神经网络及其可迁移性的黑盒模型对抗攻击研究

- 答辩人：刘晓涛 MG1937017
- 导　师：申富饶 教授

南京大学
NANJING UNIVERSITY

目录
CONTENTS

南京大学
NANJING UNIVERSITY

第一部分

Research Background

研究背景

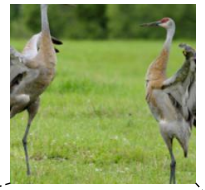
对抗攻击 在干净样本上添加人眼不可察觉的扰动后，使得模型以高置信度分类错误；

干净样本：跑车

对抗扰动η

对抗样本：牛蛙

黑盒攻击 攻击者不了解黑盒模型的结构、参数设置和模型权重时，仅根据模型的输出信息生成对抗样本



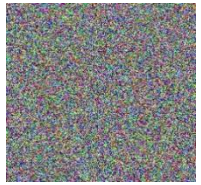
输入

黑盒模型

输出

Label: 狗

黑盒攻击 $x_{adv} = x + \alpha\,\eta$



研究对抗攻击意义

隐私保护

信息加密，网站保护

促进增强神经网络鲁棒性

军事战略意义

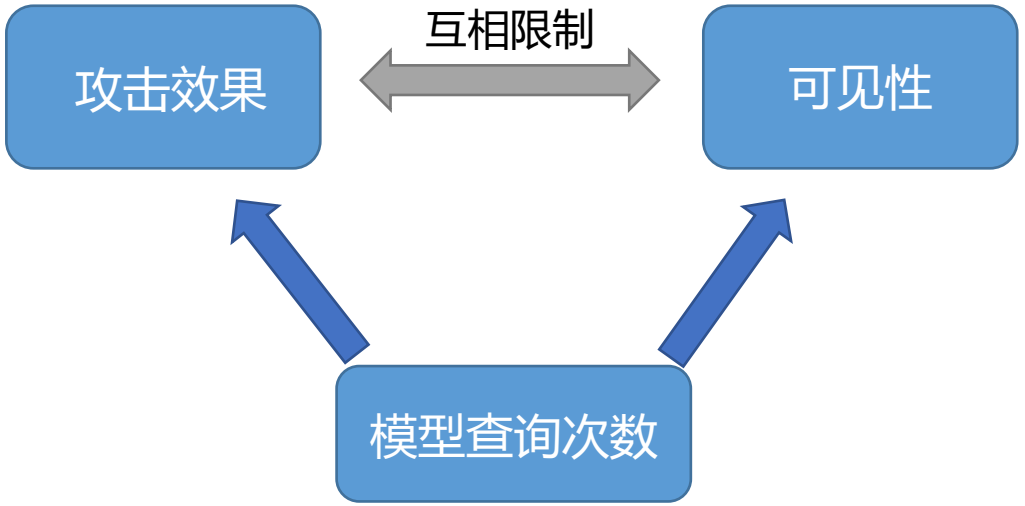
黑盒攻击场景

- Soft-label 场景: 可查询黑盒模型，获得模型关于各个类别的置信度输出信息
- **Hard-label 场景**: 可查询黑盒模型，获得模型的预测标签（Top-1标签）
- **无模型查询场景**: 无法查询黑盒模型

困难和挑战

攻击成功率

- 无目标：1 – 模型准确率
- 有目标：目标匹配率



L_p范数

p可取值为$0, 2, \infty$

黑盒攻击的目标：用**更少黑盒模型查询**次数实现**高攻击成功率**和**低可见性**

南京大学
NANJING UNIVERSITY

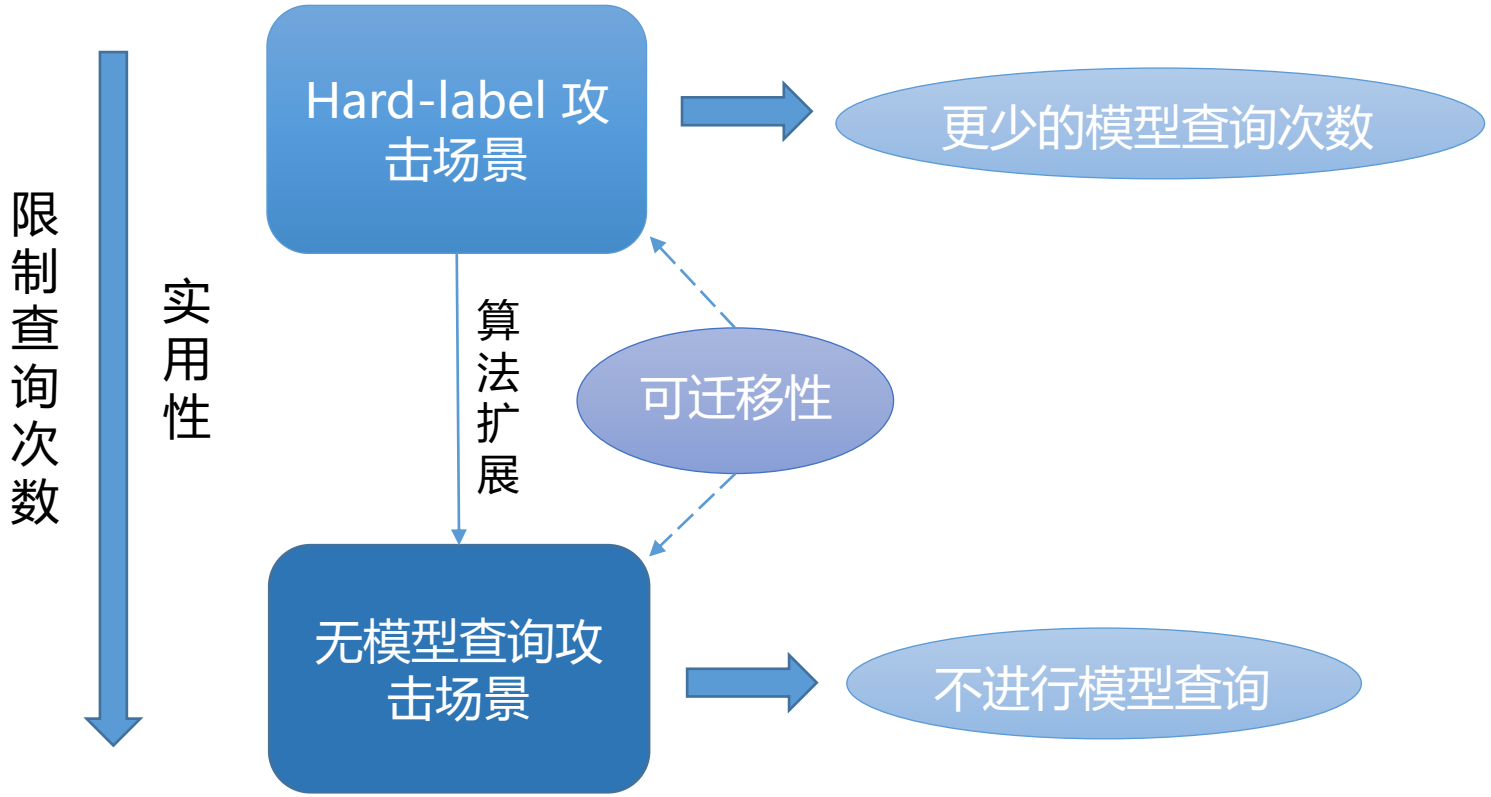
第二部分

Research Content

研究内容

- 结合可迁移性的hard-label黑盒攻击方法
- 基于对抗变换的无模型查询的黑盒攻击方法

工作的联系


限制查询次数
实用性
Hard-label 攻击场景
更少的模型查询次数
算法扩展
可迁移性
无模型查询攻击场景
不进行模型查询

结合可迁移性的hard-label黑盒攻击方法

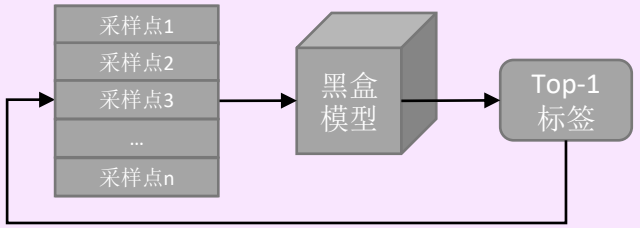
研究动机

可迁移性

对抗样本存在特性：**可迁移性**

A模型和B模型可能具有不同的模型结构或在不同数据集上训练获得

当前的hard-label攻击



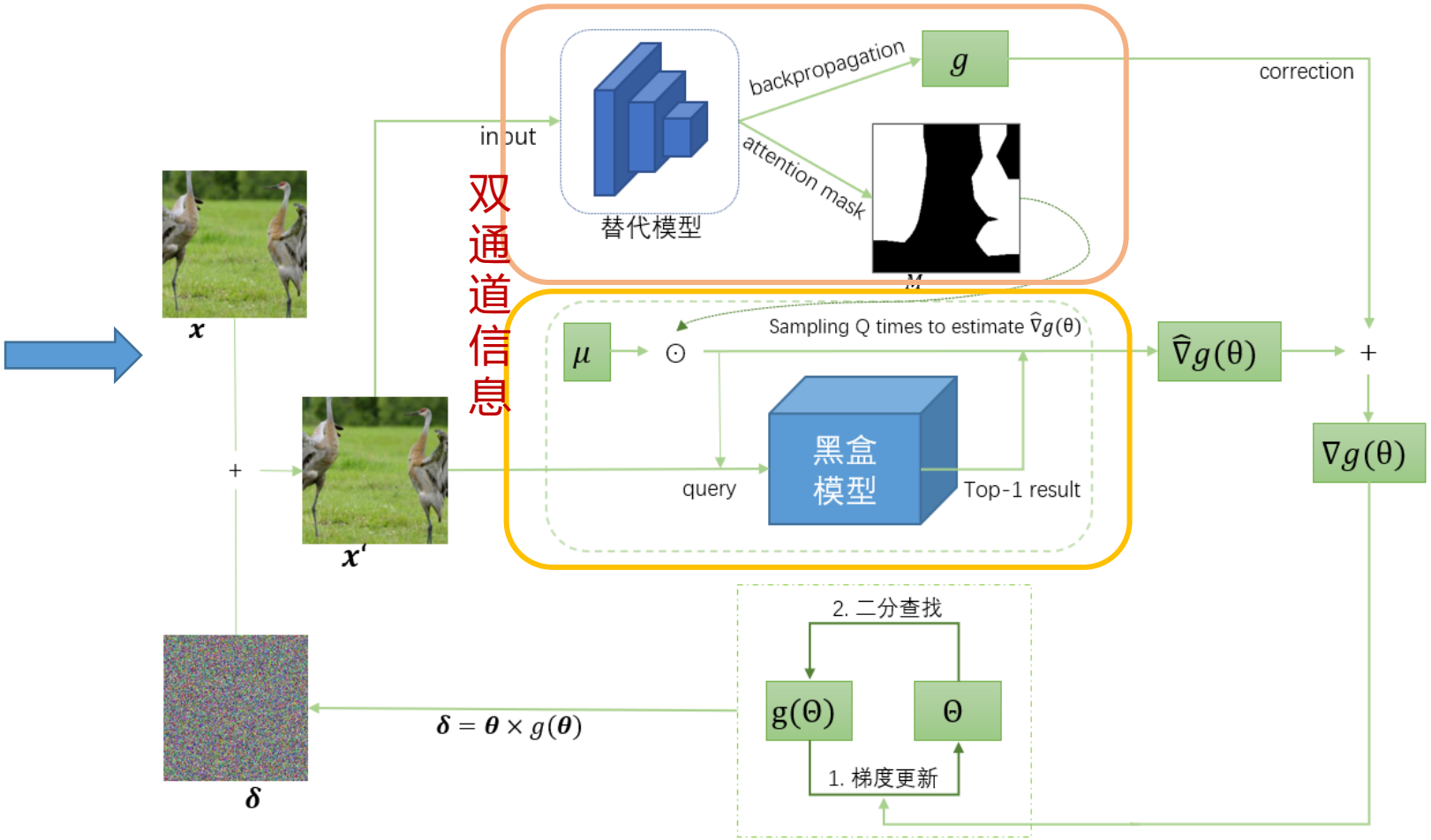
需要**大量采样**样本点进而**查询**黑盒模型
→消耗了大量资源，引起模型的攻击检测

- 当前的黑盒攻击最大的问题：可利用信息匮乏，存在过量查询
- 而基于可迁移性的方法则是：从无信息→引入可迁移信息
- 能否使用替代模型的可迁移信息弥补hard-label信息不足的缺点

结合可迁移性的hard-label黑盒攻击方法

整体结构

设计结合可迁移信息和hard-label模型查询双通道信息的攻击结构



结合可迁移性的hard-label黑盒攻击方法

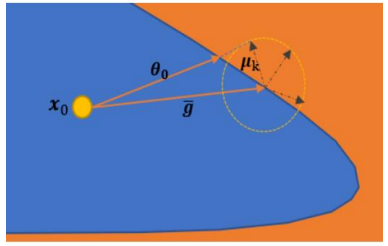
关键部分

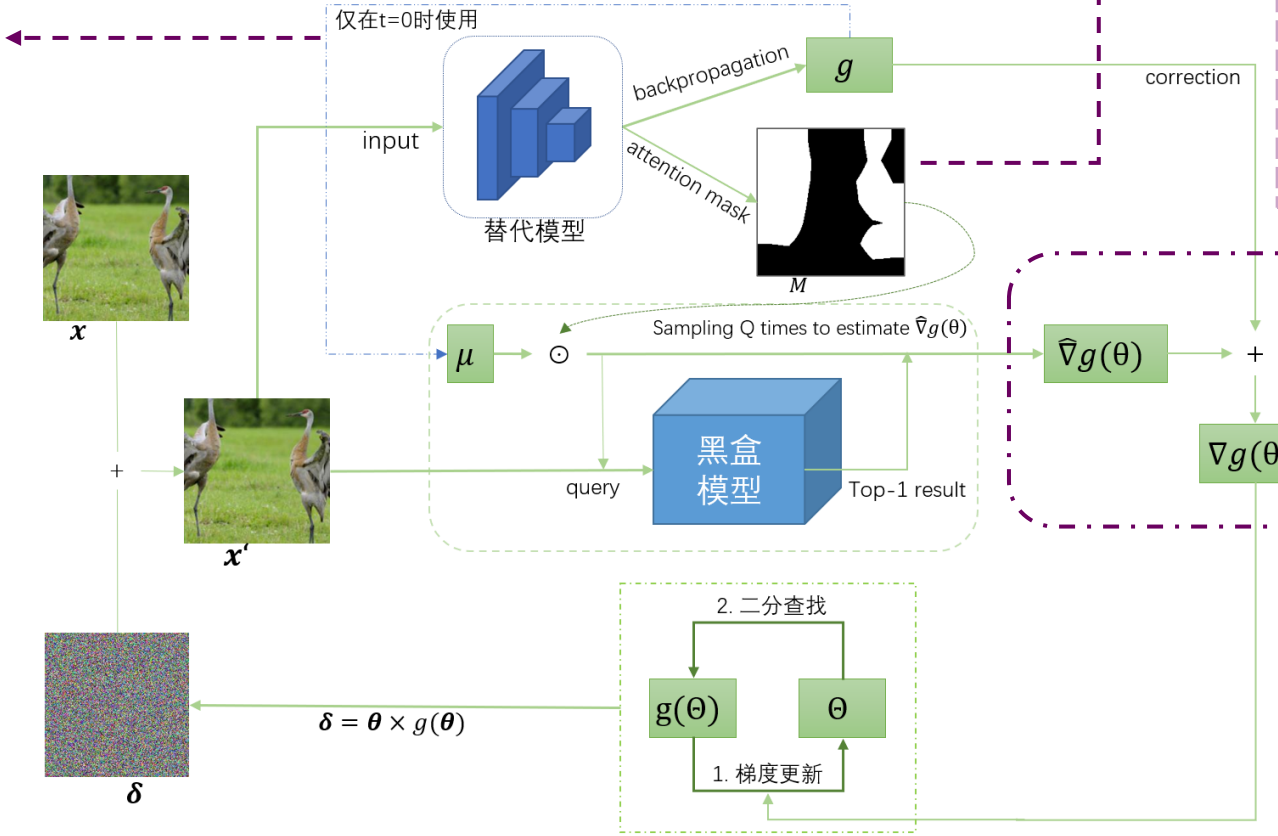
(1) 引导

快速确定扰动起始方向，避免盲目采样

$$\min_{\theta_0} \quad g(\theta_0)$$

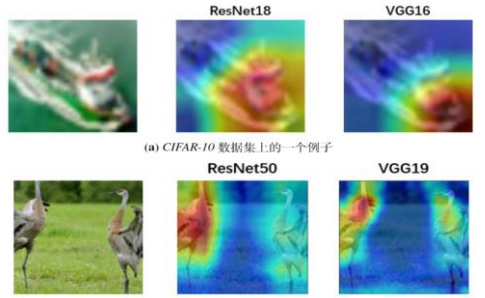
$$s.t. \quad \theta_0 = \frac{\mu_k + \alpha\bar{g}}{\|\mu_k + \alpha\bar{g}\|_2}, \quad for \quad k \in [1, n].$$





(3) 类相关的注意力掩码

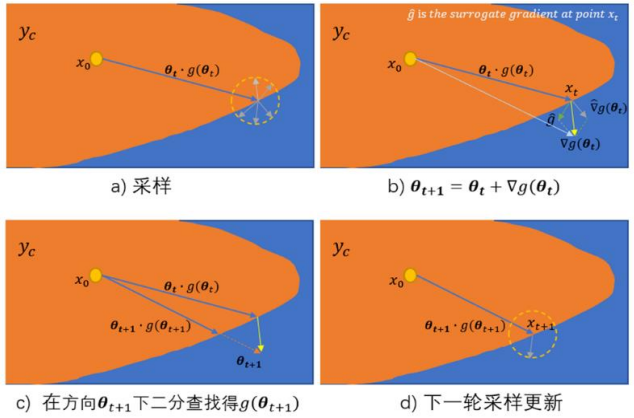
- 不同模型对相同目标可能具有类似的关注点
- 利用掩码过滤无关扰动，提高攻击方向性



(2) 修正

利用采样进行蒙特卡洛估计梯度更新，结合替代模型实时梯度调整更新值

$$L_{surr}(\mathbf{x}') = -CrossEntropy(S(\mathbf{x}'), y) + \lambda\|\mathbf{x}' - \mathbf{x}_0\|_2$$



实验验证

与 hard-label 攻击方法的实验验证

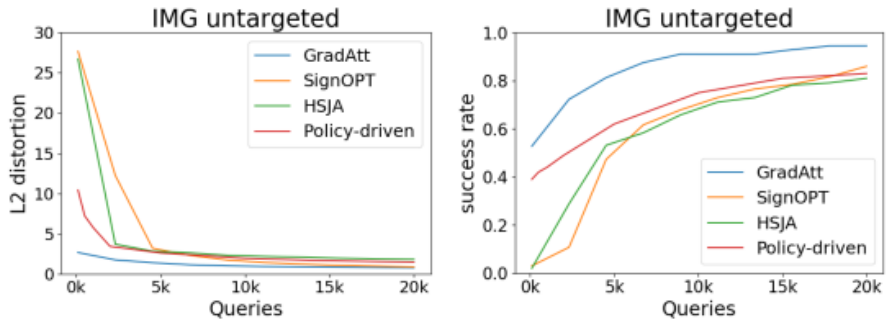
表 3.1 无目标攻击: 不同模型查询次数下的平均 L_2 扰动的对比

数据集	被攻击模型	攻击方法	Avg. L_2 @1k	Avg. L_2 @5k	Avg. L_2 @10k	Avg. L_2 @20k
MNIST	CNN	BA	20.322	4.441	2.120	1.824
		OPT attack	12.412	4.850	2.812	2.210
		Sign-OPT	10.221	1.652	1.350	1.292
		GradAtt-OPT	**4.354**	**1.220**	**1.162**	**1.160**
CIFAR-10	VGG16	BA	6.328	1.385	0.485	0.392
		OPT attack	2.504	1.400	0.742	0.655
		Sign-OPT	2.387	0.422	0.390	0.251
		GradAtt-OPT	**0.822**	**0.202**	**0.182**	**0.171**
ImageNet	ResNet50	BA	60.618	39.359	13.887	5.675
		Sign-OPT	23.876	3.130	1.541	0.853
		HJSA	8.295	2.816	2.236	1.800
		Rays	6.990	2.985	2.395	1.924
		Policy-driven	4.751	1.593	1.473	1.351
		GradAtt-OPT	**2.647**	**1.248**	**0.965**	**0.759**

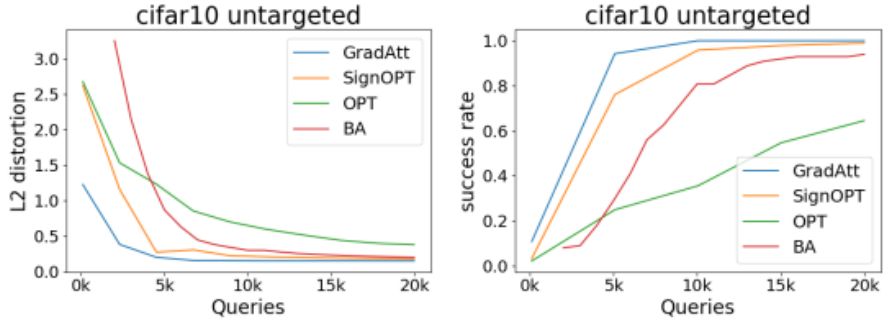
表 3.2 无目标攻击: 不同模型查询次数下的 ASR 的对比

数据集	被攻击模型	攻击方法	ASR @1k	ASR @5k	ASR @10k	ASR @20k
MNIST	CNN	BA	0.0%	3.2%	32%	41.3%
		OPT attack	1.5%	5.6%	8.2%	30.0%
		Sign-OPT	5.5%	45.3%	57.4%	58.7%
		GradAtt-OPT	**10.2%**	**68.2%**	**72.5%**	**73.0%**
CIFAR-10	VGG16	BA	0.0%	28.5%	78.2%	91.5%
		OPT attack	8.1%	26.8%	38.5%	69.4%
		Sign-OPT	17.7%	68.6%	92.2%	95.6%
		GradAtt-OPT	**22.1%**	**82.3%**	**97.7%**	**98.9%**
ImageNet	ResNet50	Sign-OPT	8.8%	51.6%	71.4%	85.2%
		HJSA	15.0%	56.2%	68.0%	79.4%
		Rays	16.6%	50.4%	57.8%	67.6%
		Policy-driven	44.4%	62.4%	75.8%	83.2%
		GradAtt-OPT	**61.4%**	**82.4%**	**90.8%**	**94.2%**

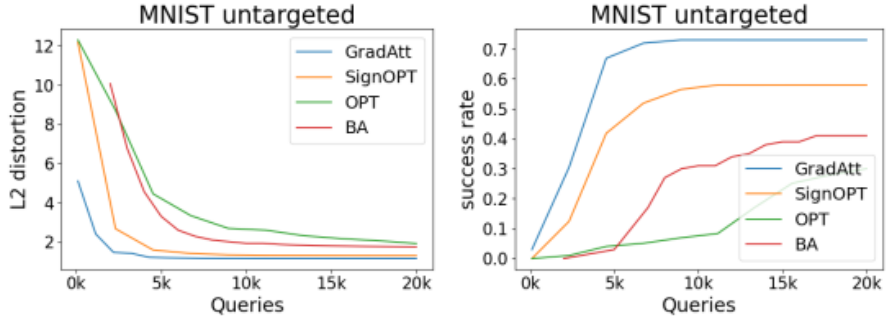
1. 消耗相同的模型查询次数，本方法具有**更小的平均L2扰动**
2. 消耗相同的模型查询次数，本方法具有**更高的攻击成功率**（ASR）
3. 实现相同的攻击效果，本方法消耗的**模型查询次数更少**



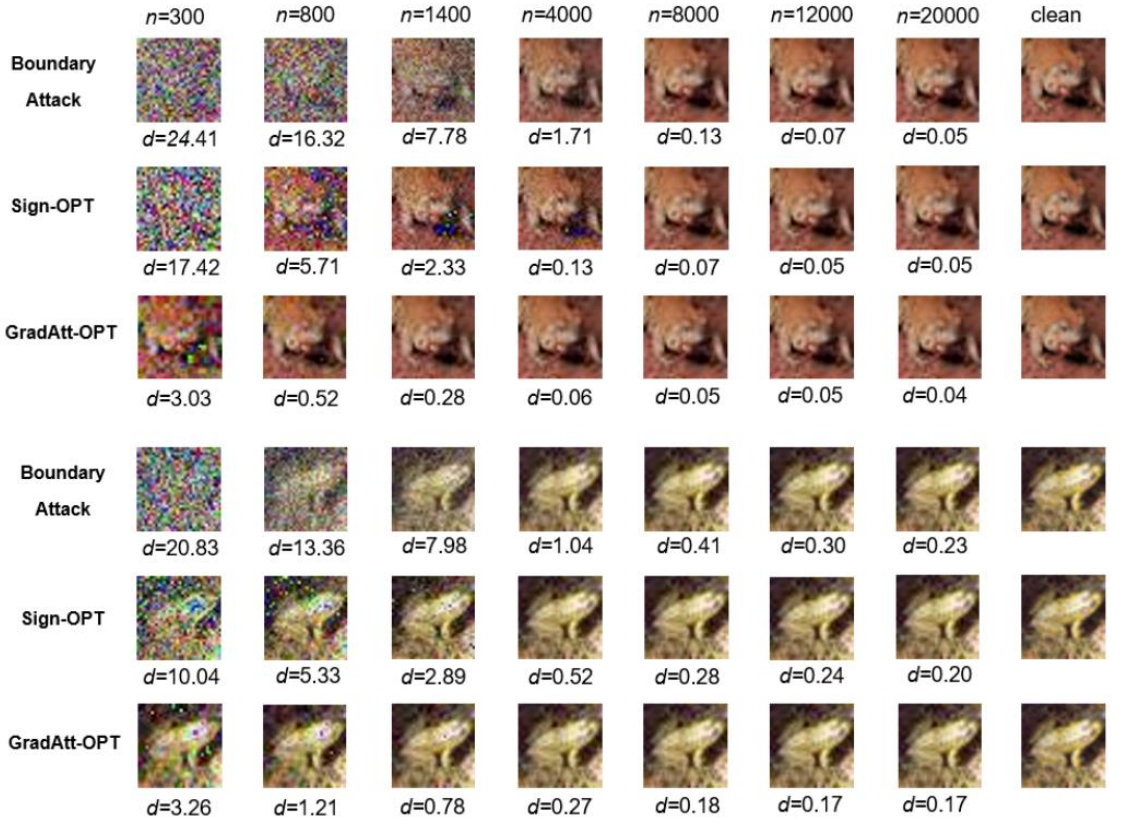
(a) 在 *ImageNet* 数据集上，不同方法的平均 L_2 扰动和攻击成功率的比较



(b) 在 *CIFAR-10* 数据集上，不同方法的平均 L_2 扰动和攻击成功率的比较



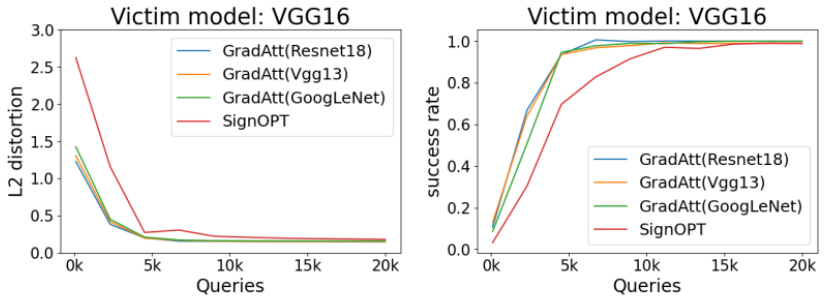
(c) 在 *MNIST* 数据集上，不同方法的平均 L_2 扰动和攻击成功率的比较



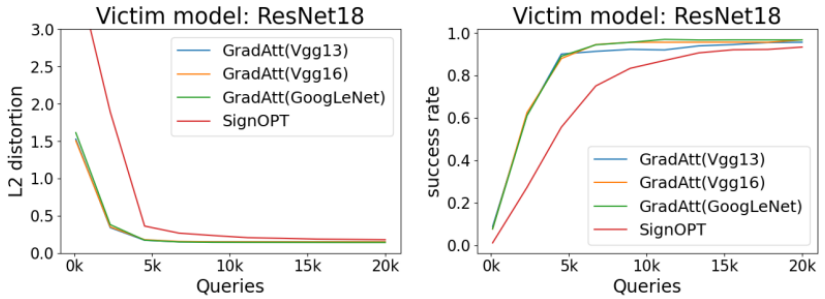
1. 在攻击起始阶段就可以快速降低对抗扰动，主要得益于**引导**作用
2. 本方法的**查询效率更高，攻击成功率也更高**，可以快速生成扰动更小的对抗样本

消融实验

1. 使用不同替代模型并攻击不同victim model



(a) 使用不同替代模型攻击 VGG16 时的平均 L_2 扰动和 ASR 对比



(b) 使用不同替代模型攻击 ResNet18 时的平均 L_2 扰动和 ASR 对比

图 3.7 使用不同替代模型攻击不同 Victim Model 的对比 (括号表示所使用的替代模型)

1. 使用其他替代模型下，本方法具有稳定的攻击效果，说明其**泛化性**
   攻击不同的黑盒模型，本方法优于hard-label方法，说明其**通用性**
2. 在**相似任务**下训练的替代模型同样可以本方法提高查询效率
3. 本方法效果提升在于**可迁移信息的启发作用**，而不是可迁移信息本身

2. 使用不同数据集训练的替代模型

表 3.7 在不同数据集下训练的替代模型攻击相同 victim model 的攻击成功率 ASR

攻击方法	替代模型	@2k	@4k	@10k
Sign-OPT	-	18.2%	59.1%	92.2%
GradAtt-OPT	**ResNet18**	**53.7%**	**82.7%**	**97.5%**
	VGG13-CF100	51.0%	80.1%	94.6%
	GoogLeNet-CF100	50.2%	78.3%	94.5%
	ResNet50-CF100	51.6%	81.4%	95.9%

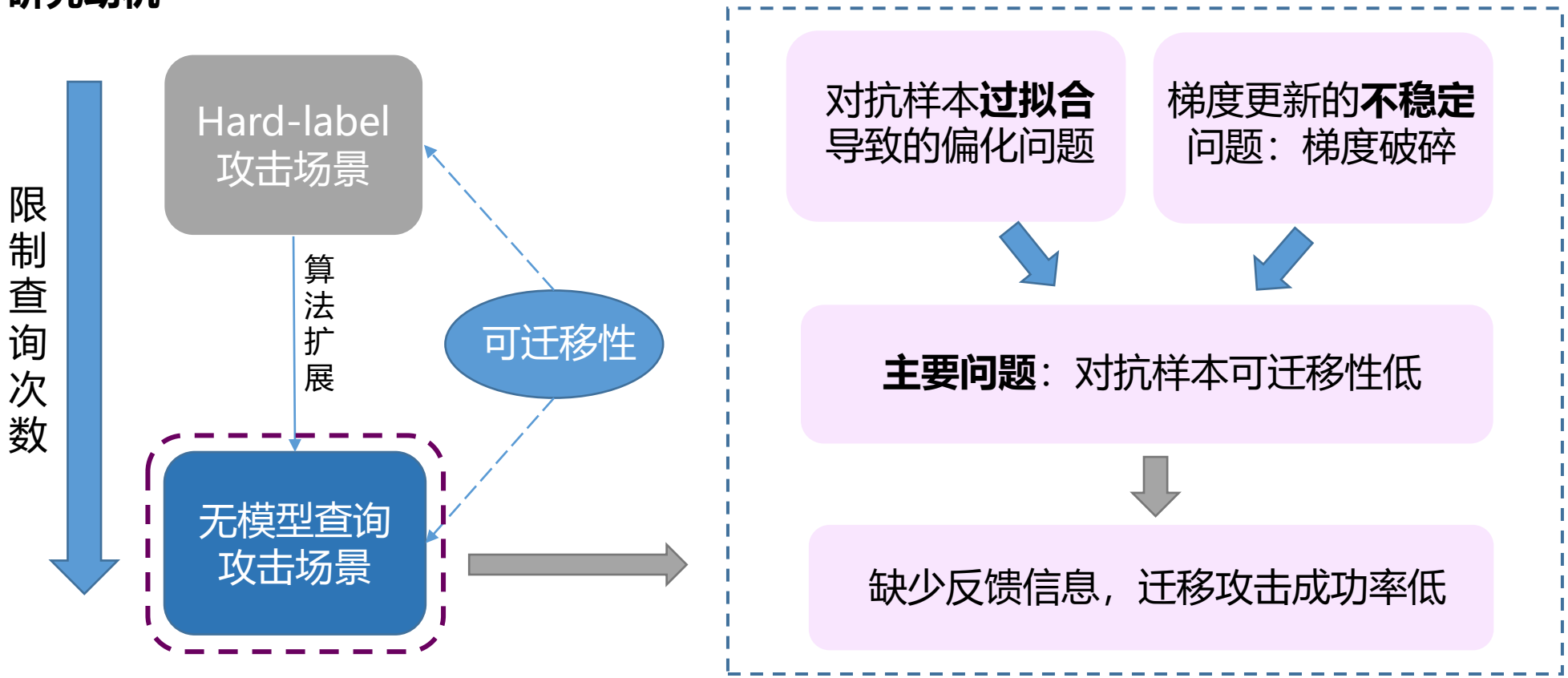
3. 与基于可迁移性的方法的对比

表 3.5 在不同扰动阈值下直接利用替代模型生成对抗样本的攻击成功率

攻击方法	替代模型	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$
PGD	VGG13	15.8%	49.2%	76.2%
	GoogLeNet	15.0%	44.0%	69.3%
	ResNet18	13.1%	35.9%	65.9%
MI-FGSM	VGG13	17.0%	50.9%	74.9%
	GoogLeNet	15.9%	49.8%	72.7%
	ResNet18	12.9%	38.7%	62.7%
GradAtt-OPT	ResNet18	**98.9%**	-	-

基于对抗变换的无模型查询的黑盒攻击方法

研究动机

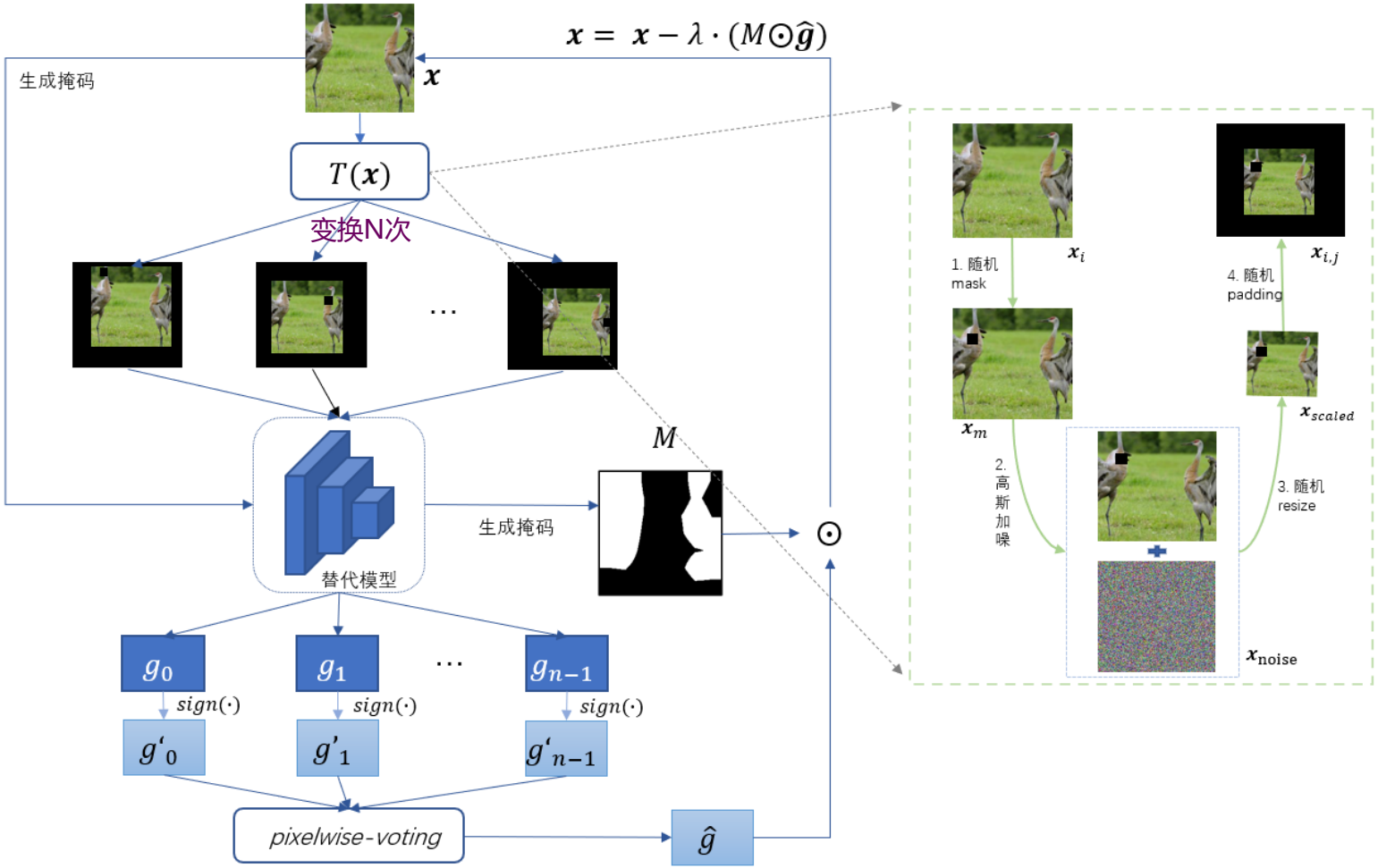


禁止模型查询是更实际的场景，提升对抗样本的可迁移性是关键

基于对抗变换的无模型查询的黑盒攻击方法

整体结构

1. 设计**自集成**的结合**对抗变换**的**无模型查询**攻击方法
2. 继承了利用可迁移信息的方式和基于类别相关的敏感像素的注意力掩码



基于对抗变换的无模型查询的黑盒攻击方法

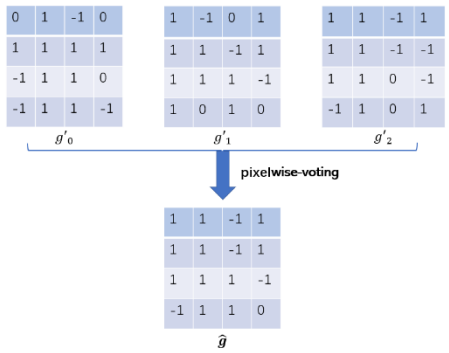
关键模块

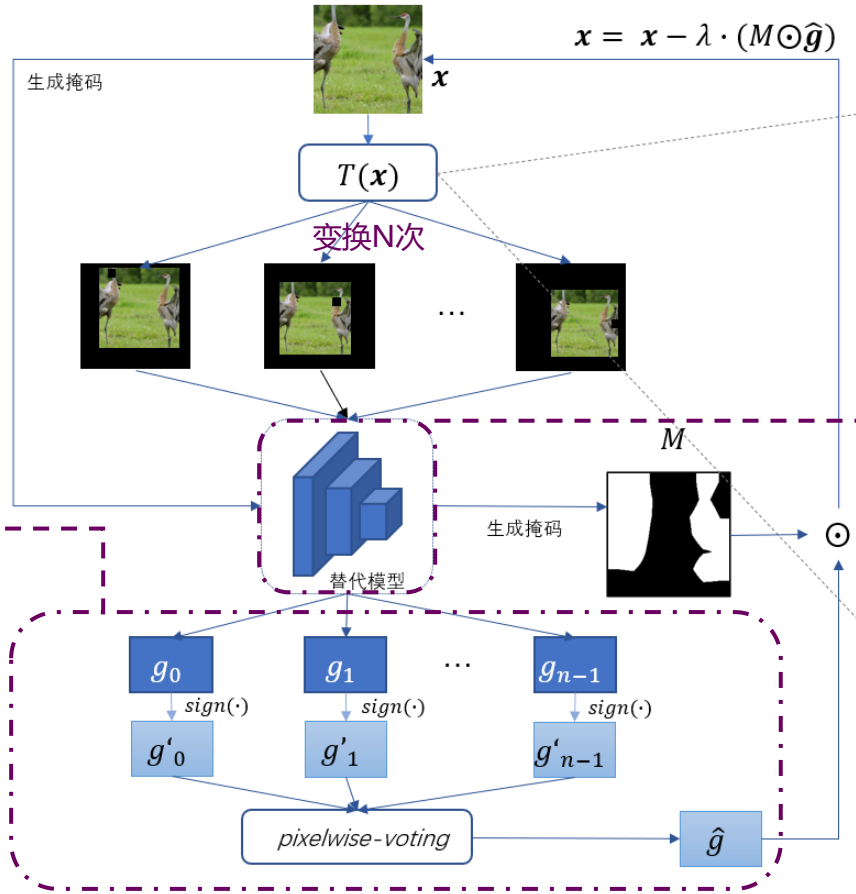
针对梯度不稳定问题

(2) 自集成 sign 梯度的 pixelwise-voting 平滑

- 自集成：集成的不是模型，而是集成对抗变换的样本点的sign梯度值

$$sign(\mathbf{x}_{i,j})\begin{cases}1, & \mathbf{x}_{i,j}>0\\ 0, & \mathbf{x}_{i,j}=0\\ -1, & \mathbf{x}_{i,j}<0\end{cases}$$

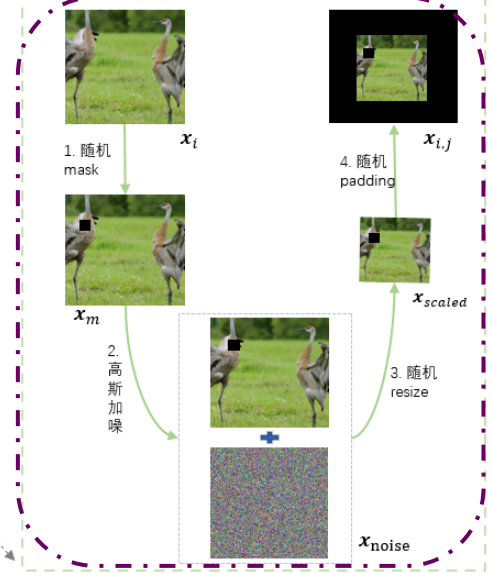




(1) 对抗变化处理

- 添加随机个数的0-1掩码
- 随机变换避免了样本单一性
- 随机mask和高斯加噪：防止扰动单方面发展

针对过拟合问题



(3) 损失函数

$$L(S(\mathbf{x}_{adv}); y) = -CrossEntropy(S(\mathbf{x}_{adv}), y) + \beta_1 \|\mathbf{x}_{adv} - \mathbf{x}\|_2 + \beta_2 L_{inf}(\mathbf{x}_{adv}, \mathbf{x})$$

$CrossEntropy(S(\mathbf{x}adv), y)$ 为误分类损失，越大越好；$\|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_2$用于防止生成的整体扰动太大；

$L_{inf}(x_{adv}, x) = \sum_{i=0}^{H}\sum_{j=0}^{W}(\max\left(abs\left(x_{adv_{ij}} - x_{ij}\right) - \tau, 0\right))$ 抑制单像素的扰动过大

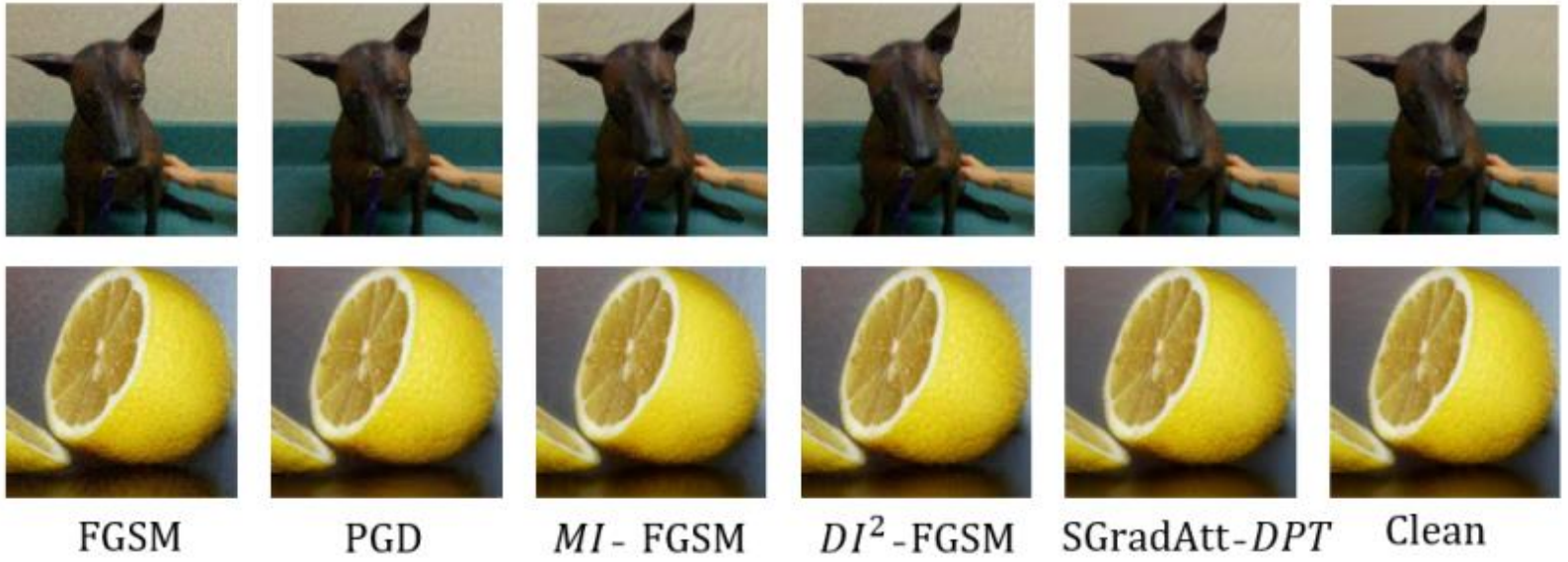
实验验证

表 4.3 在 *CIFAR-10* 数据集上 ResNet18 的对抗样本的迁移攻击成功率

被攻击模型	FGSM		PGD		DI^2-FGSM		MI-FGSM		**ours**	
	L_2	ASR	L_2	ASR	L_2	ASR	L_2	ASR	L_2	ASR
VGG13	2.0	68.0%	1.5	86.1%	1.5	85.9%	1.5	83.4%	**1.4**	**89.1%**
VGG16	2.1	54.5%	1.6	67.1%	1.5	66.0%	1.5	65.2%	**1.4**	**76.2%**
GoogLeNet	1.9	65.3%	1.6	86.0%	1.5	82.6%	1.5	81.6%	**1.4**	**89.2%**

表 4.4 在 *ImageNet* 数据集上 ResNet18 的对抗样本的迁移攻击成功率

被攻击模型	FGSM		PGD		DI^2-FGSM		MI-FGSM		**ours**	
	L_2	ASR	L_2	ASR	L_2	ASR	L_2	ASR	L_2	ASR
VGG19	13.8	77.8%	8.8	90.0%	9.1	86.4%	9.0	82.6%	**7.4**	**93.0%**
ResNet50	13.8	78.6%	8.8	80.2%	9.1	86.6%	9.0	84.6%	**7.3**	**93.6%**
DenseNet121	13.9	72.4%	8.8	62.8%	9.1	73.0%	9.0	73.8%	**7.3**	**85.4%**



1. 在不同数据集上，攻击不同黑盒模型，我们的方法能都取得**最高的迁移攻击成功率**
2. 本方法生成对抗样本比其他方法的**对抗扰动更小**

消融实验

1. 不同对抗变换次数N对迁移攻击成功率的影响

表 4.5 随机多样化变换次数 N 的不同时候的 ASR

victim model	ASR@ $N=1$	ASR@ $N=2$	ASR@ $N=4$	ASR@ $N=8$	ASR@ $N=16$
VGG19	86.2%	88.6%	93.0%	94.4%	**95.2%**
ResNet50	87.8%	90.2%	93.6%	94.8%	**95.2%**
DenseNet121	76.0%	79.0%	85.4%	87.6%	**88.6%**

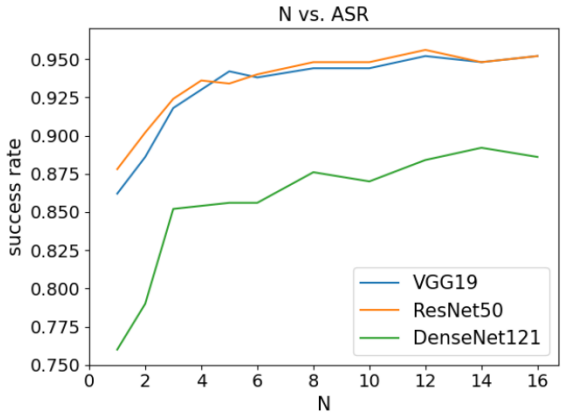


图 4.6 不同模型下随着 N 次数增加的攻击准确率变化曲线

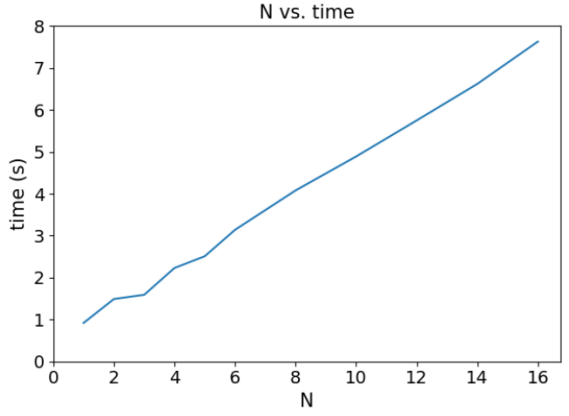


图 4.7 随着 N 次数增加的平均时间变化曲线

2. 随机掩码和高斯加噪对迁移攻击成功率的影响

表 4.6 引入随机掩码和高斯加噪前后的 ASR 对比

victim model	无加入	随机掩码	高斯加噪	随机掩码 + 高斯加噪
VGG19	89.6%	91.4%	91.8%	**93.0%**
ResNet50	90.2%	92.8%	91.6%	**93.6%**
DenseNet121	83.2%	84.4%	83.8%	**85.4%**

3. 对比攻击对抗训练模型的效果

表 4.7 不同方法攻击 IAT 对抗训练的 ResNet18 的识别准确率

	FGSM	PGD	DI^2-FGSM	MI-FGSM	ours
模型未受到攻击	92.4%	92.4%	92.4%	92.4%	92.4%
攻击后 (原始 VGG16)	89.3%	90.9%	89.9%	89.4%	**84.2%**
攻击后 (VGG16-IAT)	59.5%	57.4%	61.3%	58.5%	**56.4%**

1. N越大，则ASR越高，平均耗时越大，如何选择**N值是攻击成功率和速度的权衡**
2. 随机掩码和高斯加噪可以**干预扰动的单方面发展**，提高对抗扰动的可迁移性
3. 本方法攻击对对抗训练后的模型具有更好的攻击成功率，更强的攻击效果

南京大学
NANJING UNIVERSITY

第三部分

Applications

实际应用

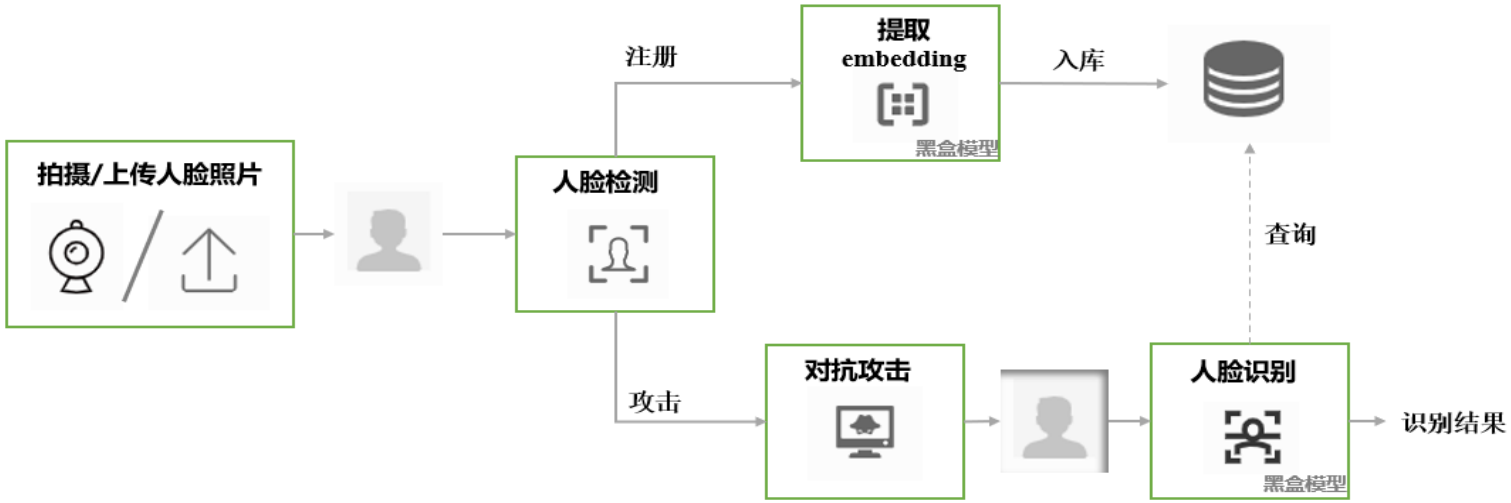
- 黑盒人脸对抗攻击系统

人脸对抗攻击系统：简述

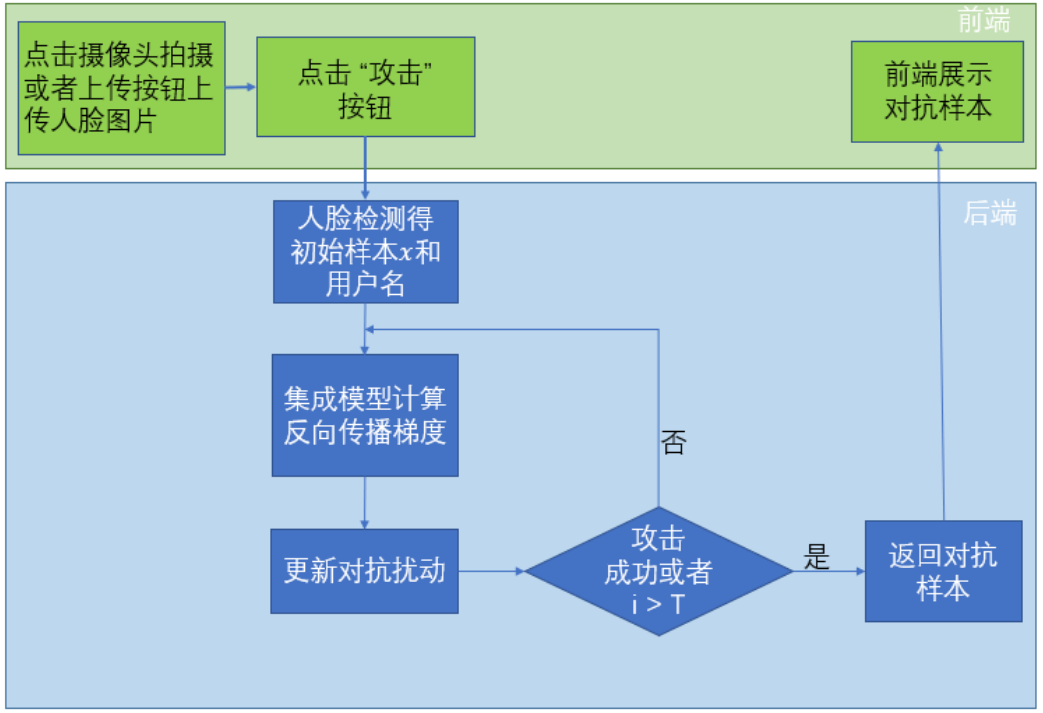
- 意义
  - 为个人提供隐私保护手段。
  - 为人脸识别模型提供增强鲁棒性的训练数据；

难点：**人脸多种多样**，难以通过**低可见的**对抗扰动攻击**黑盒条件**下的人脸识别模型

- 系统需求
  - 人脸照片采集（上传或拍摄）
  - 人脸注册
  - 人脸识别
  - 人脸攻击（生成对抗样本）



人脸对抗攻击模块介绍



实验效果:

能够跨数据集攻击不同结构的黑盒模型 → **攻击算法具有有效性**

关键技术

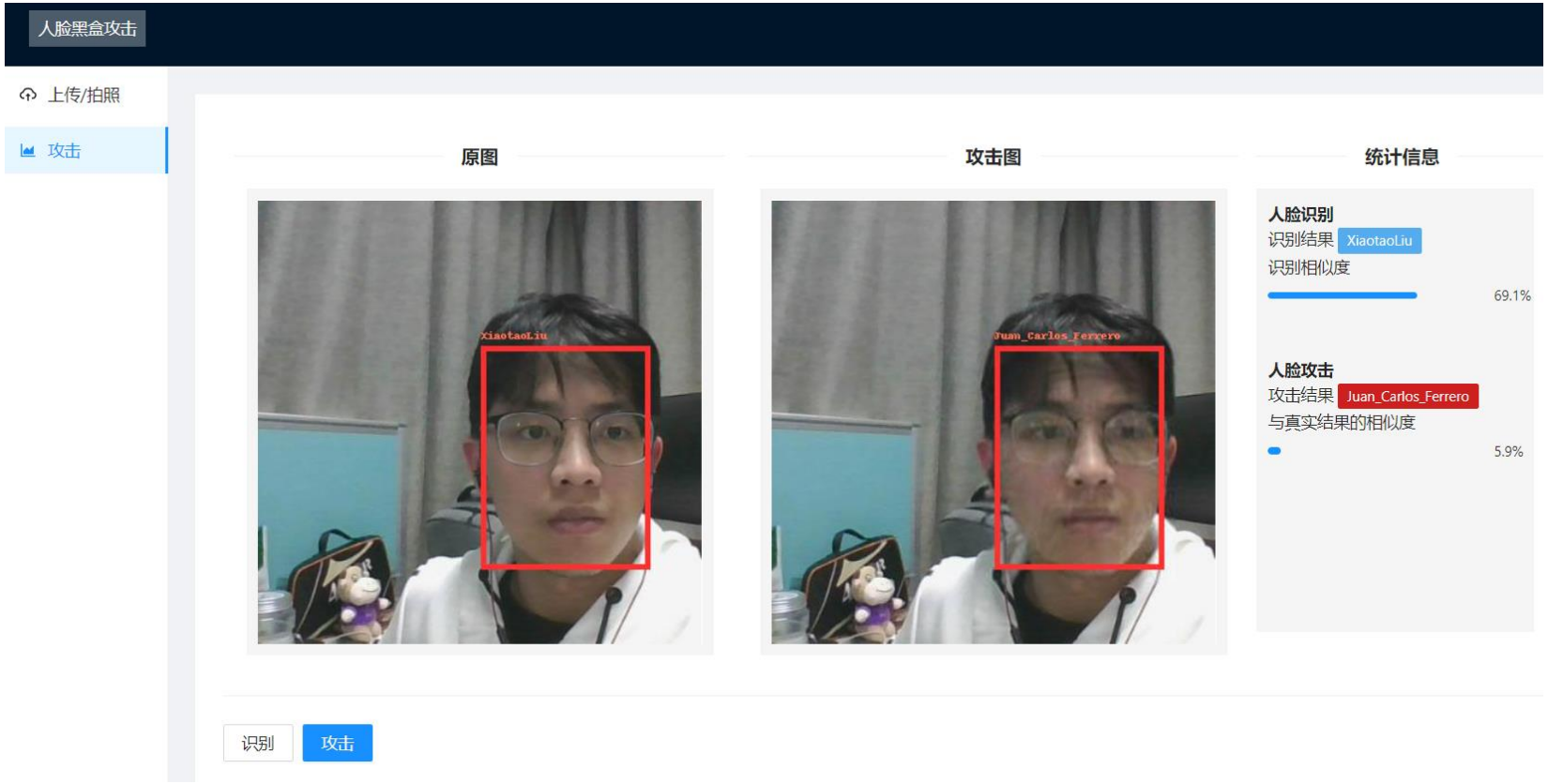
- 使用了前述**无模型查询**算法
- 引入**模型集成**的思想
- 设计**双扰动约束的相似度损失**函数

$$L_{total} = L_{cos} + \lambda_1 L_{mse} - \lambda_2 L_{dist}$$

表 5.1 白盒和黑盒的人脸识别模型分别对干净样本和对抗样本的识别准确率

模型	LFW_IR50 (LFW，白盒模型)	MS1M_IRSE102 (MS1M，黑盒模型)
干净样本	99.8%	96.7%
对抗样本	0.3%	13.6%

攻击系统：系统功能展示



系统效果：

1. 黑盒模型以高置信度识别正确干净样本；

2.黑盒模型将攻击后的对抗样本识别成其他人

本方法具有**足够的攻击能力**，系统具有**使用价值**

南京大学
NANJING UNIVERSITY

第四部分

Work and Research Progress

研究生期间工作成果

论文

- **Xiaotao Liu**, Jinqiao Li, Furao Shen, Jian Zhao. “GradAtt-OPT: A Query-Efficient Black-Box Attack”, under-review

专利

- 李俊，**刘晓涛**，严骅. 一种基于统计和桩定位视觉的快速检测和定位方法别方法。专利申请号：202110718272.5

项目

- 国家自然科学基金“基于深度感知增量式联想记忆神经网络的信息融合系统研究”项目

竞赛

- **刘晓涛**，李金桥，薛轲. FinTechathon 2020 微众银行第二届金融科技高校技术大赛 ，AI赛道top3

南京大学
NANJING UNIVERSITY

第五部分

Summary

工作总结

工作总结

结合可迁移性的hard-label黑盒攻击方法

- 提出了结合可迁移信息的查询高效的攻击方法
- 引入替代模型的梯度进行引导和修正
- 基于类别相关的注意力掩码提高采样方向性

算法拓展

基于对抗变换的无模型查询黑盒攻击方法

- 提出高可迁移性的攻击方法
- 设计随机多样性的对抗变化算法
- 设计自集成 sign 梯度的 pixelwise-voting 平滑机制

算法应用

黑盒人脸对抗攻击的应用

- 基于黑盒条件设计了人脸对抗攻击系统
- 引入SGradAtt-DAT算法并适应于人脸识别
- 构造了人脸注册，人脸识别到人脸攻击的完整流程

南京大学
NANJING UNIVERSITY

谢谢！