

2023-2024学年秋季学期



# AI安全与伦理

## AI Security and Ethics

韦星星

---

副教授，博士生导师  
北京航空航天大学人工智能研究院  
[xxwei@buaa.edu.cn](mailto:xxwei@buaa.edu.cn)



**北京航空航天大学**  
**人工智能研究院**



# AI安全与伦理

## 第二章 对抗攻击方法

---

### 2.1 绪论

2.1.1 对抗攻击的基本术语

2.1.2 对抗攻击的类型

### 2.2 数字攻击

2.2.1 基于梯度的攻击

2.2.2 基于优化的攻击

2.2.3 基于迁移的攻击

2.2.4 基于分数的攻击

2.2.5 基于决策的攻击

### 2.3 物理攻击

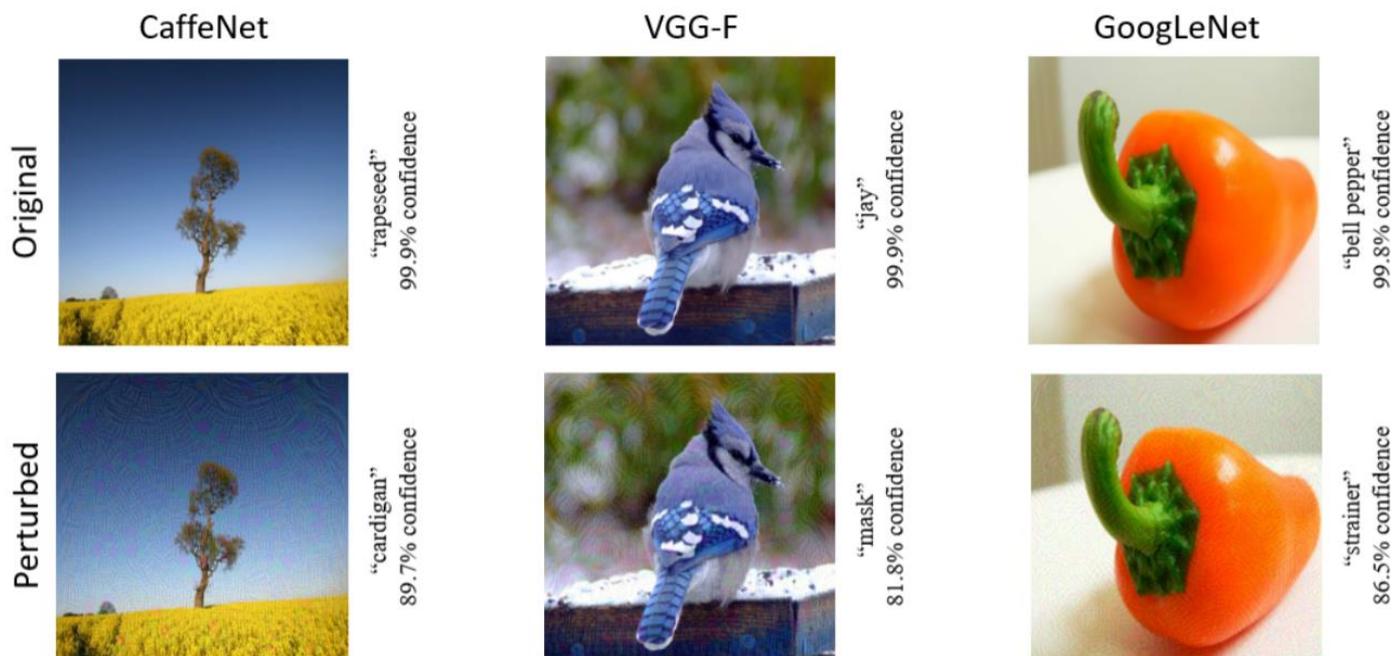
2.3.1 对抗性补丁攻击

2.3.3 基于自然变换攻击

2.3.2 基于其他媒介攻击

## 第二章 对抗攻击方法

### 2.1 绪论



**对抗攻击 (adversarial attack)**：即对输入样本故意添加一些人类无法察觉的**细微干扰**，导致模型以高置信度给出**错误的输出**。

## 2.1.1 对抗攻击中的基本术语

- **对抗样本 (Adversarial Example)** : 指被恶意扰动后的图像样本, 能够误导机器学习模型 (如深度神经网络)
- **对抗性扰动 (Adversarial perturbation)** : 指使得干净样本变为对抗样本所添加的微小噪声
- **白盒攻击 (White-box Attacks)** : 白盒攻击假设目标模型的完整知识可知, 包括模型的参数、架构、训练方法以及训练数据
- **黑盒攻击 (Black-box Attacks)** : 黑盒攻击假设目标模型的知识不可知, 或仅能够获得有限的模型知识, 如训练方法或模型结构, 但模型的参数未知
- **受害模型 (Victim Model)** : 指被攻击的模型, 如VGG、ResNet等
- **可转移性 (Transferability)** : 指的是对抗样本即使对于用于生成它的模型以外的模型仍然有效的能力

## 对抗攻击的类型：

- 根据攻击针对的环境：

### 数字攻击

假定攻击者可以完全访问受害模型的实际数字输入。攻击的行为发生在数字空间内。现有的对抗性攻击大多是此类

### 物理攻击

攻击行为发生在物理空间中，通过修改真实物体的视觉特征，如纹理、形状、光照等，用于攻击受害模型



Clean image

"Fast";  $L_\infty$  distance to clean image = 32

数字攻击

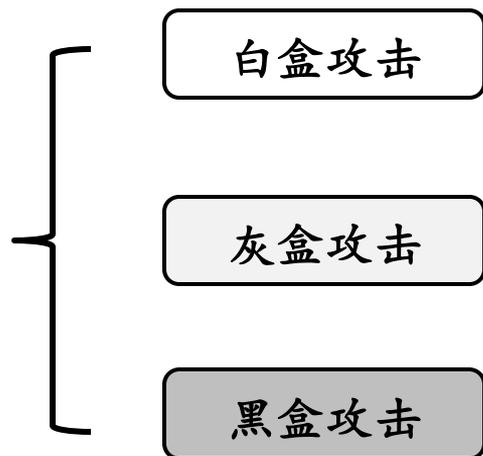


物理攻击

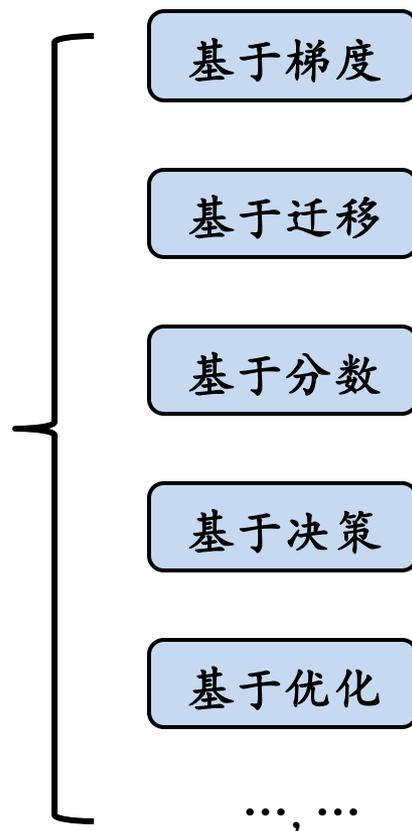
与数字世界的攻击相比，物理世界的攻击更具挑战性，这是因为物理世界约束和条件复杂（如：光线、距离、相机等），这些因素都会影响生成对抗扰动的攻击能力

## 对抗攻击的类型：

- 根据对受害模型的了解程度：



- 根据攻击所利用的信息：



- 根据攻击要达到的目的：



## 2.1 Reference

---

- [1] Chakraborty A, Alam M, Dey V, et al. Adversarial attacks and defences: A survey.
- [2] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey
- [3] Chakraborty A, Alam M, Dey V, et al. A survey on adversarial attacks and defences
- [4] CV|对抗攻击领域综述 (adversarial attack) <https://zhuanlan.zhihu.com/p/104532285>



# AI安全与伦理

## 第二章 对抗攻击方法

---

### 2.1 绪论

2.1.1 对抗攻击的基本术语

2.1.2 对抗攻击的类型

### 2.2 数字攻击

2.2.1 基于梯度的攻击

2.2.2 基于优化的攻击

2.2.3 基于迁移的攻击

2.2.4 基于分数的攻击

2.2.5 基于决策的攻击

### 2.3 物理攻击

2.3.1 对抗性补丁攻击

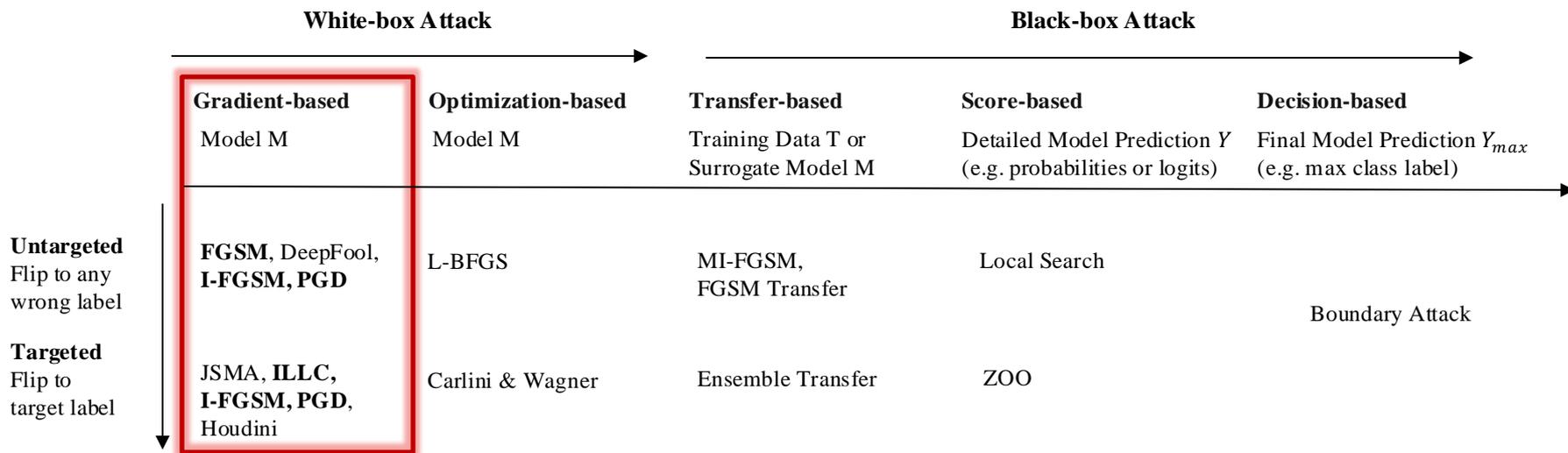
2.3.3 基于自然变换攻击

2.3.2 基于其他媒介攻击

## 2.2 数字攻击

### 2.2.1 基于梯度的攻击

基于梯度的攻击（Gradient-Based Attack）在已知模型的参数和结构前提下，计算损失对图像像素的梯度方向生成对抗样本。



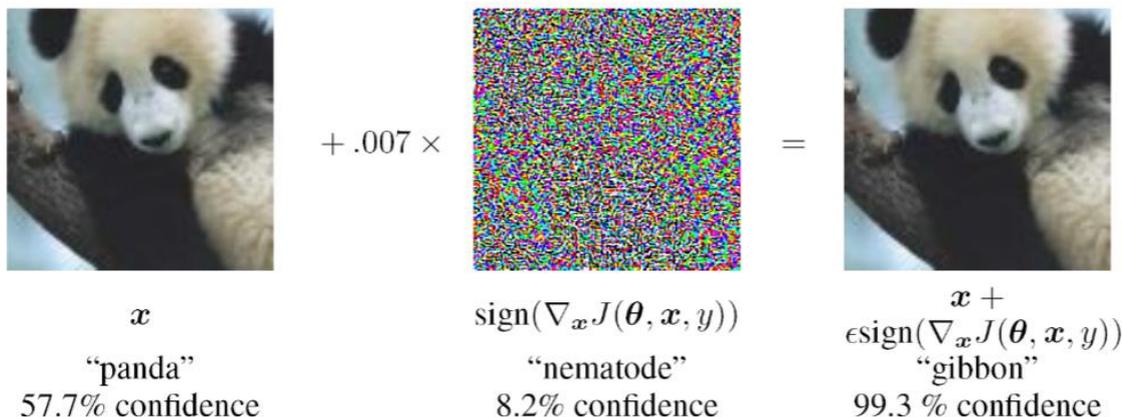
图：数字对抗攻击方法划分

## 2.2.1 基于梯度的攻击

### 1、快速梯度符号法（Fast Gradient Sign Method, FGSM）

**2014年**由Goodfellow et al. 提出，是最早的针对深度模型的对抗攻击方法之一。它将图像的各点像素值沿着交叉熵损失对输入的梯度方向进行一步迭代，寻找到 $L_\infty$ 约束下最大扰动的对抗样本。利用梯度信息最大化了对抗样本到正确标签之间的误差：

$$x' = x + \epsilon \cdot \text{sgn}(\nabla_x J(x, y_{true}))$$



其中， $\epsilon$ 表示对抗扰动的大小， $J$ 表示交叉熵损失函数， $\text{sgn}$ 是符号函数。FGSM具有计算量小、一步到位的优势。它是一种无目标的、非迭代的攻击方法。

## 2.2.1 基于梯度的攻击

### 2、基本迭代法 (Basic Iterative Method, BIM)

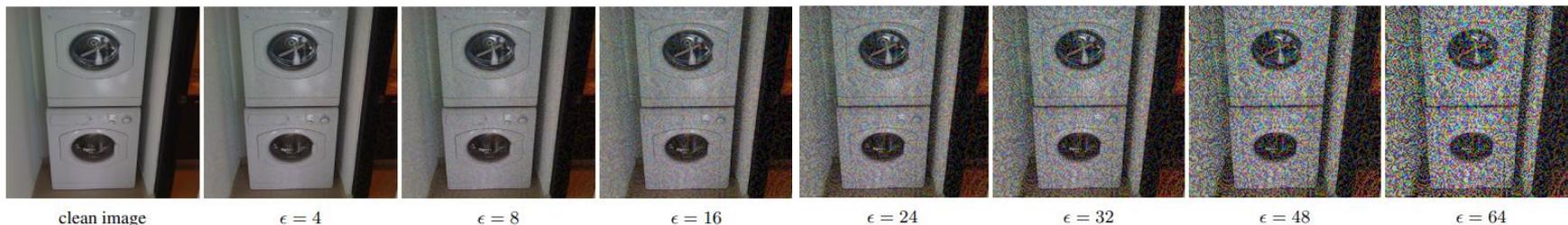
FGSM只利用梯度进行了一步迭代，因此所产生对抗样本的攻击性能有限。为了生成更为精细、攻击性更强的对抗样本。Kurakin et al. 在 FGSM 的基础上扩展了其迭代形式，因此，BIM也被称为I-FGSM (Iterative-FGSM)

$$x'_0 = x$$

$$x'_{N+1} = \text{clip}_{x, \epsilon} \{x'_N + \alpha \cdot \text{sgn}(\nabla_x J(x'_N, y_{true}))\}$$

$$\text{clip}_{x, \epsilon} \{x'\} = \min\{255, x + \epsilon, \max\{0, x - \epsilon, x'\}\}$$

如上式所示，BIM把 FGSM 的一次步长分为了 N 步进行，并将每次迭代攻击后的值约束在给定的扰动范围内。通过多步迭代的方式产生对抗样本具有更好的可视化效果，在相同的扰动大小下能够产生更好的攻击效果



## 2.2.1 基于梯度的攻击

### 3、迭代最小可能类攻击 (Iterative Least-Likely Class Method, ILLC)

上述的方法，都属于无目标攻击。并没有规定需要被错分到某一类。与之对应的，有目标攻击旨在最小化网络输出与某个错误标签的损失，使得对抗样本将被分为**特定的类**

无目标攻击

即只要攻击成功就好，对抗样本的最终属于哪一类不做限制

有目标攻击

不仅要求攻击成功，还要求生成的对抗样本属于特定的类

$$x'_0 = x$$

$$x'_{N+1} = \text{clip}_{x, \epsilon} \{x'_N - \alpha \cdot \text{sgn}(\nabla_x J(x'_N, y_{LL}))\}$$

其中 $y_{LL}$ 是指定分错的类别标签，注意到 ILLC 和 BIM 的区别就在于迭代公式由加上梯度改为了**减去梯度**

## 2.2.1 基于梯度的攻击

### 4、投影梯度下降法（Projected Gradient Descent, PGD）

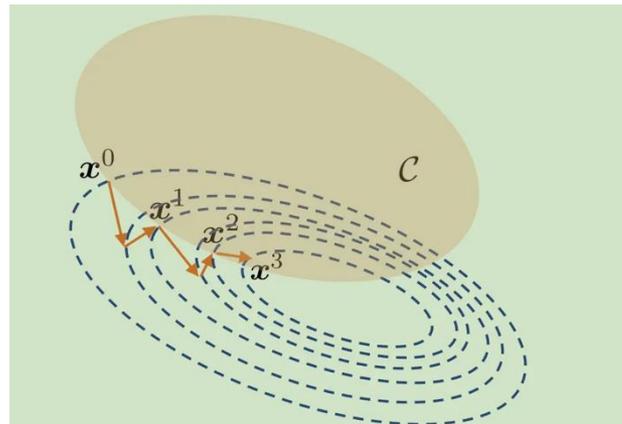
BIM的缺陷在于其迭代的次数取决于给定的扰动大小和规定的步长，因此迭代的次数是受到限制的。

Madry et al. 所提出的PGD算法通过每次迭代后将扰动投影回约束范围，确保扰动不会导致输入样本离原始数据分布过远，并可以实现任意多的迭代次数。因此PGD被证明是最强的一阶攻击方法

如果一个网络对PGD攻击鲁棒，则它对目前所有的一阶攻击方法都鲁棒

$$x'_{N+1} = \prod_{x+S} (x'_N + \alpha \cdot \text{sgn}(\nabla_x J(x'_N, y_{true})))$$

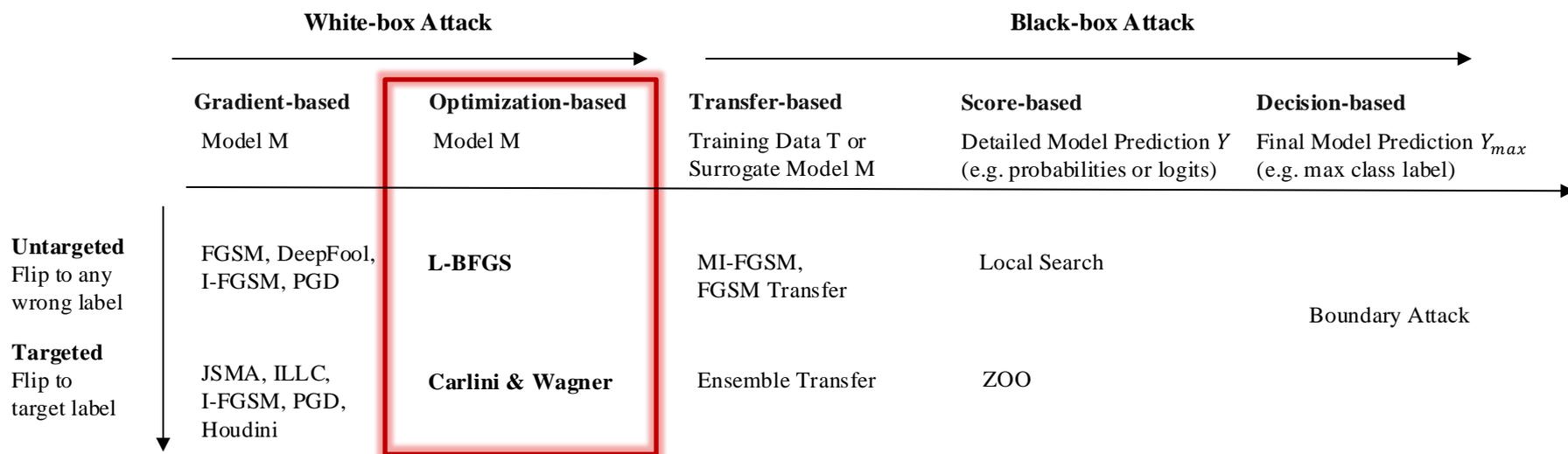
其中的 $\Pi_{x+S}$ 表示将每次迭代生成的对抗样本投影至 $x+S$ 空间内，即扰动的无穷范数不能超过 $S$



投影梯度下降示意

## 2.2.2 基于优化的攻击

基于优化的攻击（Optimization-Based Attack）旨在优化某个精心设计的目标函数（通常包含对抗性和样本间距离相关的目标）以实现攻击目标。



图：数字对抗攻击方法划分

## 2.2.2 基于优化的攻击

### 1、拟牛顿法优化 (Limited Memory Broyden-Fletcher-Goldfarb-Shanno, L-BFGS)

L-BFGS攻击是最早被设计攻击DNN模型的对抗攻击算法，由Szegedy et al.提出，也是最早的基于优化的对抗攻击方法。将对抗样本生成看作是一个优化问题，L-BFGS认为该优化问题可以表示为：

$$\begin{aligned} \min_{x'} \quad & \|x' - x\|_p && \text{对抗样本和干净样本的距离} \\ & && \text{(以} L_p \text{范数度量) 越小越好} \\ \text{s.t.} \quad & f(x') = y' && \text{对抗样本需要被网络分类错误} \\ & f(x) = y \\ & y' \neq y \\ & x' \in [0, 1]^n && \text{对抗样本的归一化像素值需} \\ & && \text{要在允许范围内} \end{aligned}$$

L-BFGS首次将生成对抗样本的过程建模为一个优化的问题来处理

## 2.2.2 基于优化的攻击

### 1、拟牛顿法优化 (Limited Memory Broyden-Fletcher-Goldfarb-Shanno, L-BFGS)

上述优化问题属于 NP-hard 问题，直接优化十分困难。为此，可以使用拉普拉斯乘子法，用额外的损失项代替约束条件，将该问题转化为盒约束 (box-constrained) 优化问题

$$\min_{x'} \quad \|x' - x\|_p$$

$$s.t. \quad f(x') = y'$$

$$f(x) = y$$

$$y' \neq y$$

$$x' \in [0, 1]^n$$

$$\min_{x'} \quad c \cdot \|\delta\|_p + J_\theta(x', y')$$

$$s.t. \quad x' \in [0, 1]^n$$

目标模型的真实损失函数的 (比如交叉熵损失函数)

令  $\delta = x' - x$ ， $\delta$  代表施加在干净样本上的**对抗性扰动**， $c$  是正则化系数，对距离损失和交叉熵损失的重要性进行加权。以上的优化过程是通过迭代的形式进行优化，并且参数  $c$  通过线性查找的方式逐渐变大直到对抗样本被发现

## 2.2.2 基于优化的攻击

### 2、C&W攻击 (Carlini & Wagner Attack)

L-BFGS的局限性：仅使用交叉熵的优化目标效果较差 / 盒约束问题难以优化

为了解决优化的难题，Carlini et al.基于L-BFGS中对抗样本的原始优化目标，重新定义了7种不同的目标函数展开实验，并最终选择了优化效果最好的那一种目标函数。首先将对抗样本的优化问题定义如下：

$$\begin{aligned} \min \quad & \|\delta\|_p + c \cdot f(x + \delta) \\ \text{s.t.} \quad & x + \delta \in [0, 1]^n \end{aligned}$$

$$\begin{aligned} f_1(x') &= -\text{loss}_{F,t}(x') + 1 \\ f_2(x') &= (\max_{i \neq t} (F(x')_i) - F(x')_t)^+ \\ f_3(x') &= \text{softplus}(\max_{i \neq t} (F(x')_i) - F(x')_t) - \log(2) \\ f_4(x') &= (0.5 - F(x')_t)^+ \\ f_5(x') &= -\log(2F(x')_t - 2) \\ f_6(x') &= (\max_{i \neq t} (Z(x')_i) - Z(x')_t)^+ \\ f_7(x') &= \text{softplus}(\max_{i \neq t} (Z(x')_i) - Z(x')_t) - \log(2) \end{aligned}$$

若将目标函数替换为交叉熵损失即为 L-BFGS 的求解形式。经过不同优化目标的尝试，C&W攻击将上式具体转化为以下形式：

$$\begin{aligned} \min \quad & \left\| \frac{1}{2} (\tanh(w) + 1) - x \right\|_2^2 + c \cdot f\left(\frac{1}{2} (\tanh(w) + 1)\right) \\ \text{where} \quad & f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\mathcal{K}) \end{aligned}$$

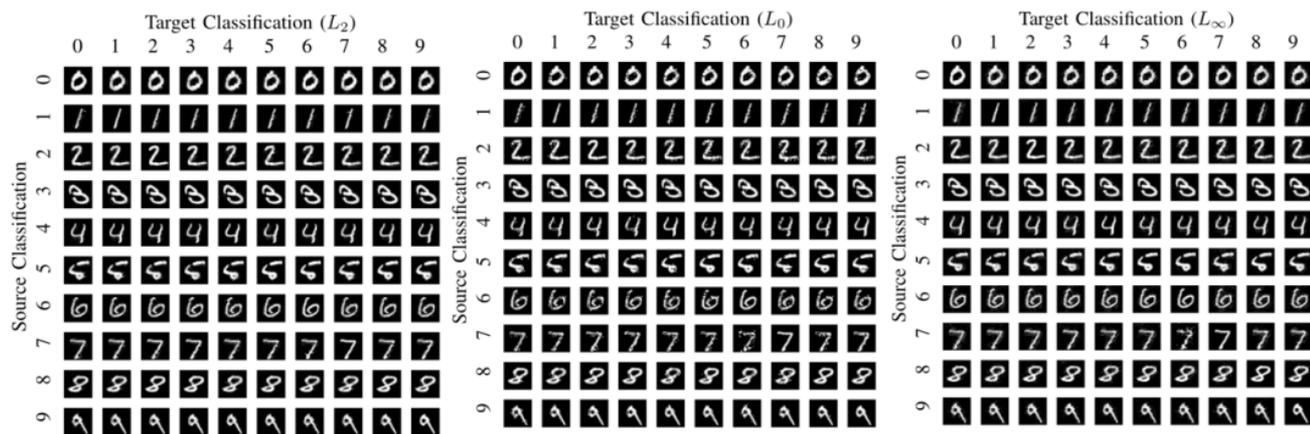
## 2.2.2 基于优化的攻击

### 2、C&W攻击 (Carlini & Wagner Attack)

$$\min \left\| \frac{1}{2} (\tanh(w) + 1) - x \right\|_2^2 + c \cdot f\left(\frac{1}{2} (\tanh(w) + 1)\right)$$

$$\text{where } f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\mathcal{K})$$

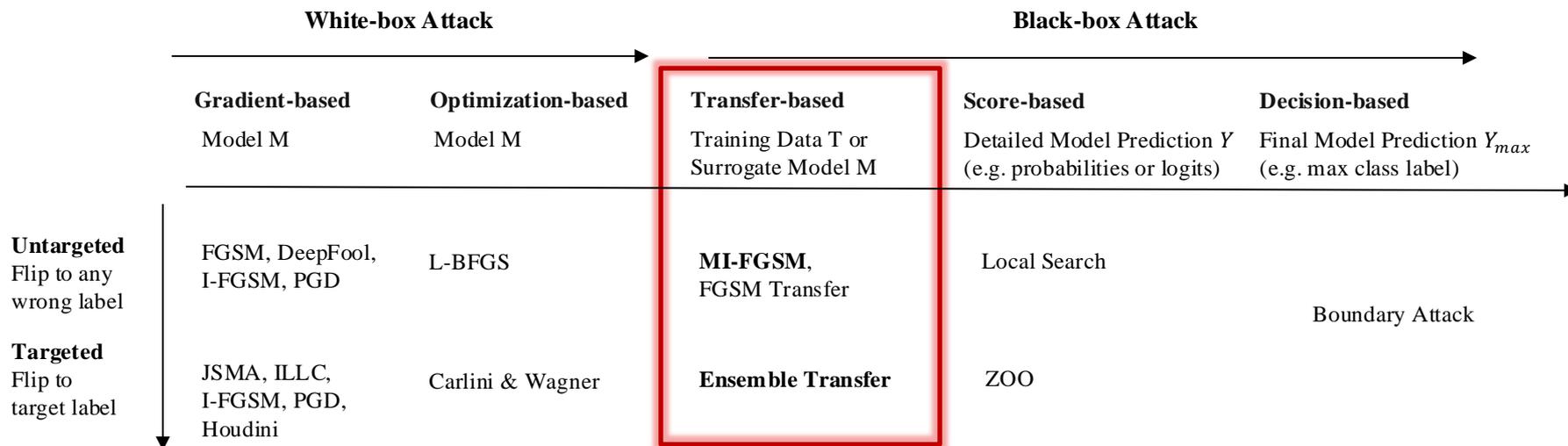
其中 $Z$ 表示神经网络最后一层输出的 softmax 向量。参数 $K$ 越大，生成样本被分类成 $t$ 类的置信度越高。此外，C&W将扰动映射至tanh空间内，由于tanh函数的特性，可以在实数域内优化扰动以应用各种先进的优化器（如 Adam）进行优化



与先前的方法相比，无论是可视化质量还是攻击成功率，C&W攻击都表现出了优越性

## 2.2.3 基于迁移的攻击

基于迁移的攻击 (Transfer-based Attack) 旨在不依赖受害模型的参数和结构信息, 通过梯度平滑、集成代理模型等方式, 获得具有可迁移性的对抗样本, 实现在黑盒模型上的攻击。在迁移攻击的设置下, 通常要知道受害模型的训练数据, 利用训练数据在本地训练代理模型。



图：数字对抗攻击方法划分

## 2.2.3 基于迁移的攻击

### 1、集成模型迁移攻击 (Ensemble-based Approaches)

如果一个对抗样本能够对多个模型保持对抗性，则它也更有可能会迁移到其他黑盒模型上

基于这一假设，Liu et al提出了一个利用多个模型生成对抗样本的技术。基本思想是使用k个白盒模型的softmax输出寻找能够成功攻击多个模型的对抗样本

$$\operatorname{argmin}_{x^*} -\log\left(\sum_{i=1}^k \alpha_i J_i(x^*)\right) \cdot \mathbf{1}_{y^*} + \lambda d(x, x^*)$$

$J_1, J_2, \dots, J_k$ 代表白盒模型的softmax输出，这里套用了C&W攻击的优化方法生成对抗样本，经过实验对比，对抗样本的迁移性得到了显著提升

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

无目标攻击的对抗样本在不同黑盒模型下的accuracy  
上：基于优化的方法，下：集成方法



	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

## 2.2.3 基于迁移的攻击

### 2、动量迭代法/集成动量迭代法 (MI-FGSM / Ensemble MI-FGSM)

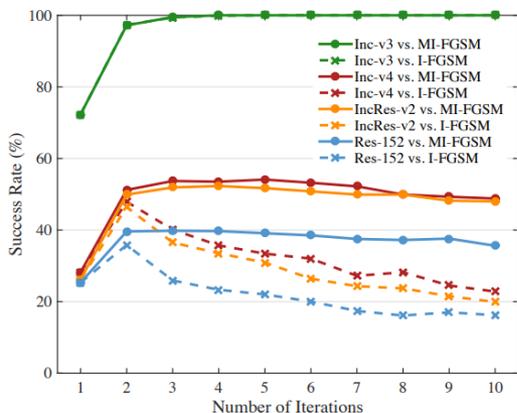
基于梯度的方法（如FGSM/I-FGSM）容易陷入局部最优值和对模型“过拟合”，会导致对抗样本的攻击性对其他的模型不太好，即缺乏可迁移性。借鉴深度神经网络中动量加速训练的技巧，MI-FGSM在I-FGSM的基础上引入了动量（Momentum）修正原本的梯度方向，加速了收敛，并且提升了对抗样本的可迁移性

$$\mathbf{g}_{t+1} = \mu \cdot \mathbf{g}_t + \frac{\nabla_{\mathbf{x}} J(\mathbf{x}_t^*, y)}{\|\nabla_{\mathbf{x}} J(\mathbf{x}_t^*, y)\|_1};$$

在原本的梯度上累加其速度矢量（动量），获得带动量的梯度

$$\mathbf{x}_{t+1}^* = \mathbf{x}_t^* + \alpha \cdot \text{sign}(\mathbf{g}_{t+1});$$

使用带动量的梯度进行迭代更新，获得对抗样本



与I-FGSM相比，引入动量后，对抗样本的收敛速度提升，且黑盒迁移攻击的成功率（右图Inc-v4、IncRes-v2、Res-152）取得了较高提升

## 2.2.3 基于迁移的攻击

### 2、动量迭代法/集成动量迭代法 (MI-FGSM / Ensemble MI-FGSM)

为了进一步提升可迁移性，Ensemble MI-FGSM集成多个模型的动量梯度方向，沿这个梯度方向生成对抗样本可以实现更强大的黑盒攻击。

Ensemble MI-FGSM对不同网络的logit输出进行融合，并定义集成的损失函数：

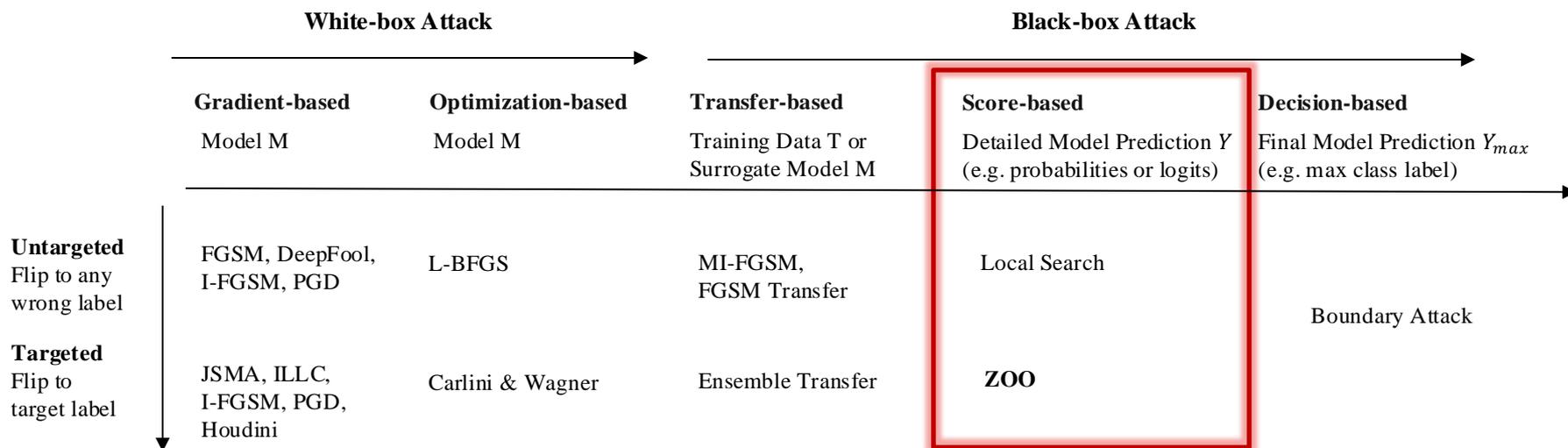
$$l(\mathbf{x}_t^*) = \sum_{k=1}^K w_k l_k(\mathbf{x}_t^*);$$

$$J(\mathbf{x}, y) = -\mathbf{1}_y \cdot \log(\text{softmax}(l(\mathbf{x}))),$$

接着使用MI-FGSM的梯度计算公式并更新对抗样本，最终能够获得可迁移性更强的对抗样本

## 2.2.4 基于分数的攻击

基于分数的攻击（Score-based Attack）只依赖于模型的预测分数（如类别概率或对数），属于黑盒攻击方法的一类。此类攻击通常使用预测分数来估计梯度，以对未知参数和结构的黑盒模型开展攻击



图：数字对抗攻击方法划分

### 1、零阶优化攻击（Zeroth Order Optimization, ZOO）

在黑盒设置下，模型的梯度信息不可获取，因此，ZOO采用了较为暴力的方式，利用模型的输出计算伪梯度（即数学上的无导数优化、零阶优化）

## 2.2.4 基于分数的攻击

### 1、零阶优化攻击（Zeroth Order Optimization, ZOO）

ZOO受到C&W攻击（基于优化的方法，白盒攻击）的启发，沿用了C&W的优化目标，不同的是，优化过程中ZOO不借助真实的梯度信息进行反向传播，而是利用估计的一阶和二阶梯度指导优化

首先回顾C&W攻击的目标：

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_0\|_2^2 + c \cdot f(\mathbf{x}, t) \\ & \text{subject to } \mathbf{x} \in [0, 1]^p, \\ & f(\mathbf{x}, t) = \max_{i \neq t} \{\max \log[F(\mathbf{x})]_i - \log[F(\mathbf{x})]_t, -\kappa\}, \end{aligned}$$

ZOO使用了对称差商（symmetric difference quotient）估计 $f(\mathbf{x})$ 的一阶梯度：

$$\hat{g}_i := \frac{\partial f(\mathbf{x})}{\partial x_i} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i)}{2h},$$

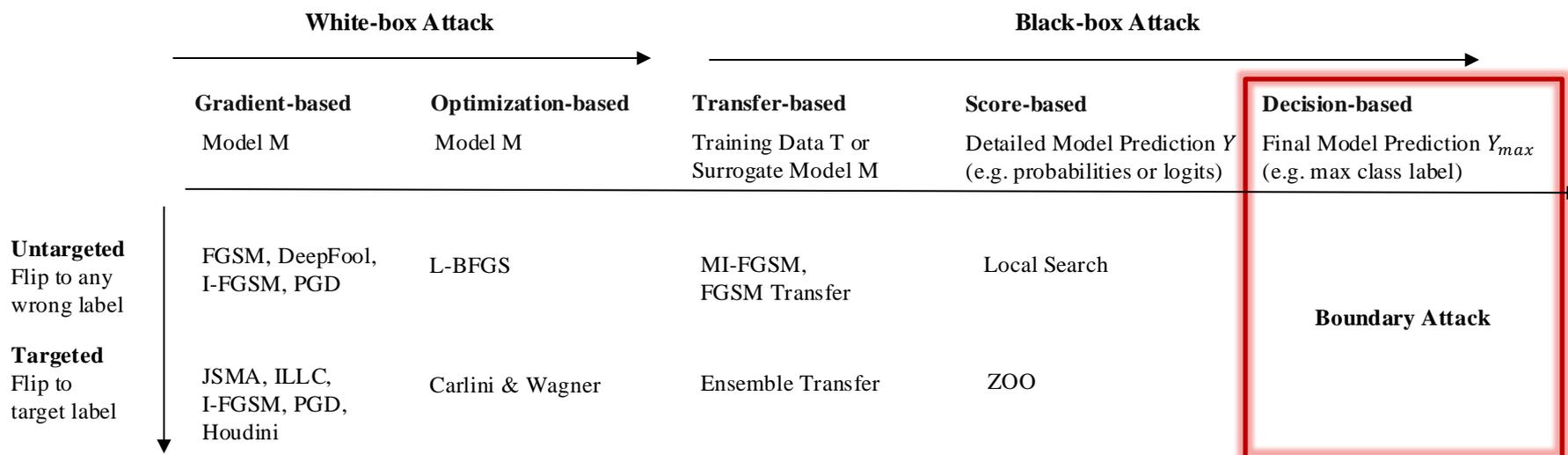
并进一步增加一次查询，可以计算二阶梯度：

$$\hat{h}_i := \frac{\partial^2 f(\mathbf{x})}{\partial x_{ii}^2} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - 2f(\mathbf{x}) + f(\mathbf{x} - h\mathbf{e}_i)}{h^2}.$$

根据这两类估计梯度，使用牛顿法便可以执行梯度下降操作优化对抗样本

## 2.2.4 基于决策的攻击

基于决策的攻击（Decision-based Attack）只依赖模型输出的分类结果（预测的标签），同样属于黑盒攻击方法。这种攻击最接近现实场景中的黑盒模型，对受害模型所需求的知识最少。



图：数字对抗攻击方法划分

## 2.2.4 基于决策的攻击

### 1、边界攻击 (Boundary Attack)

Boundary Attack从一个已经是对抗样本的点开始初始化，在优化过程中沿着对抗性和非对抗性区域之间的边界执行随机游走，逐渐收敛到和到目标图像的距离降低的那个对抗样本点。并且优化过程仅利用分类的结果开展攻击

**Data:** original image  $\mathbf{o}$ , adversarial criterion  $c(\cdot)$ , decision of model  $d(\cdot)$

**Result:** adversarial example  $\tilde{\mathbf{o}}$  such that the distance  $d(\mathbf{o}, \tilde{\mathbf{o}}) = \|\mathbf{o} - \tilde{\mathbf{o}}\|_2^2$  is minimized

initialization:  $k = 0$ ,  $\tilde{\mathbf{o}}^0 \sim \mathcal{U}(0, 1)$  s.t.  $\tilde{\mathbf{o}}^0$  is adversarial;

**while**  $k < \text{maximum number of steps}$  **do**

    draw random perturbation from proposal distribution  $\boldsymbol{\eta}_k \sim \mathcal{P}(\tilde{\mathbf{o}}^{k-1})$ ;

**if**  $\tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}_k$  is adversarial **then**

        set  $\tilde{\mathbf{o}}^k = \tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}_k$ ;

**else**

        set  $\tilde{\mathbf{o}}^k = \tilde{\mathbf{o}}^{k-1}$ ;

**end**

$k = k + 1$

**end**

初始化: Boundary attack的优化是从对抗样本开始的

- 对于Untargeted攻击，根据最大熵分布采样一个正态分布噪声的图像
- 对于Targeted攻击，直接使用攻击目标的图像

## 2.2.4 基于决策的攻击

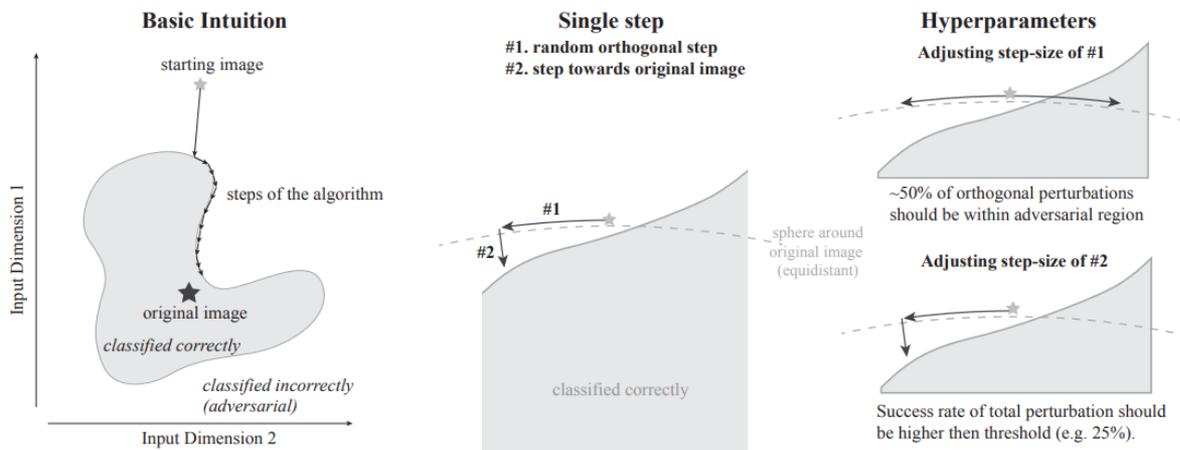
### 1、边界攻击 (Boundary Attack)

采样: idea就是从每次迭代采样一个扰动 $\eta_k$ , 扰动要满足:

- 施加该扰动的对抗样本要在值域内  $\tilde{o}_i^{k-1} + \eta_i^k \in [0, 255]$
- 扰动要和距离成相对关系  $\|\eta^k\|_2 = \delta \cdot d(o, \tilde{o}^{k-1})$
- 扰动需要减少对对抗样本和原始样本的距离  $d(o, \tilde{o}^{k-1}) - d(o, \tilde{o}^{k-1} + \eta_i^k) = \epsilon \cdot d(o, \tilde{o}^{k-1})$

超参数调整: Boundary Attack会在优化的过程动态调整算法的超参数

主要的超参数是扰动总长度 $\delta$ 和到原图距离的步长 $\epsilon$ 。如下图所示, Boundary Attack设计了一种正交扰动策略, 就是先水平扰动一下, 然后垂直靠近边界, 动态调整参数。



## 2.2 Reference

---

- [1] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples
- [2] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world
- [3] Carlini N, Wagner D. Towards evaluating the robustness of neural networks
- [4] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks
- [5] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks
- [6] Chen P Y, Zhang H, Sharma Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models
- [7] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum
- [8] Liu Y, Chen X, Liu C, et al. Delving into transferable adversarial examples and black-box attacks
- [9] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models



# AI安全与伦理

## 第二章 对抗攻击方法

---

### 2.1 绪论

2.1.1 对抗攻击的基本术语

2.1.2 对抗攻击的类型

### 2.2 数字攻击

2.2.1 基于梯度的攻击

2.2.2 基于优化的攻击

2.2.3 基于迁移的攻击

2.2.4 基于分数的攻击

2.2.5 基于决策的攻击

### 2.3 物理攻击

2.3.1 对抗性补丁攻击

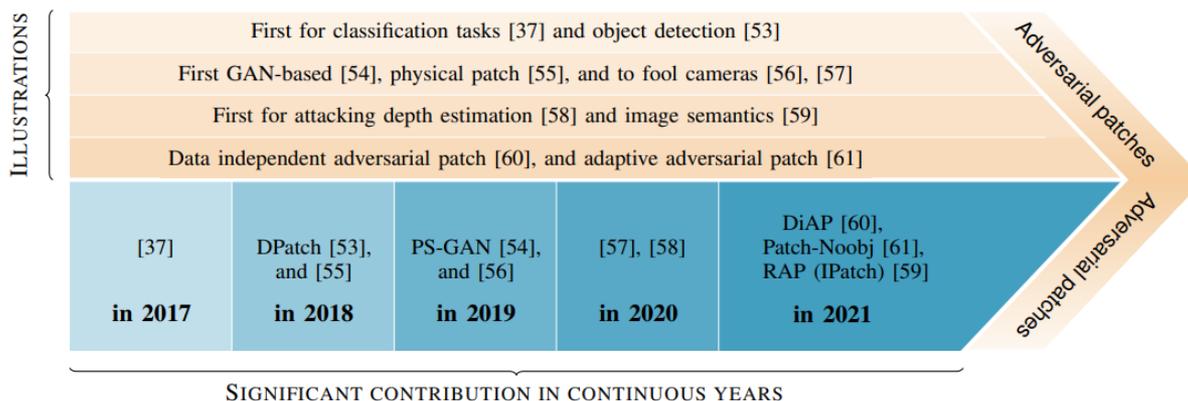
2.3.3 基于自然变换攻击

2.3.2 基于其他媒介攻击

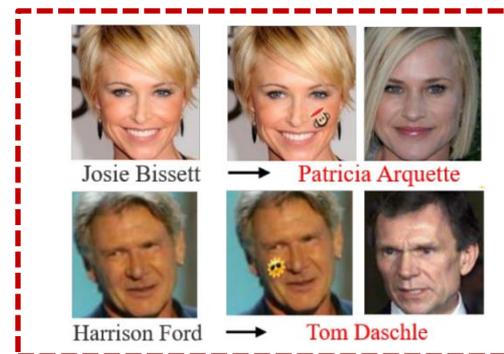
## 2.3 物理攻击

### 2.3.1 对抗补丁攻击

对抗补丁（Adversarial Patch）是一类区别于对抗扰动的对抗攻击方法，通常被用于物理世界的对抗攻击。在对抗补丁中，攻击者**不再将自己限制在难以察觉的变化中**。通常会放置一个与图像无关的补丁，让模型输出错误的或是指定的类别



无意义对抗补丁



有意义对抗补丁

## 2.3.1 对抗补丁攻击

### 1、Adversarial Patch

Adversarial Patch是最早提出并将对抗补丁用于物理世界的工作

下图展示了优化后的对抗补丁在物理世界的攻击示例：将对抗补丁打印后放置于桌面，并拍摄放置前后的图像，对抗补丁成功误导了VGG16的分类结果



99%的置信度认为是“toaster”

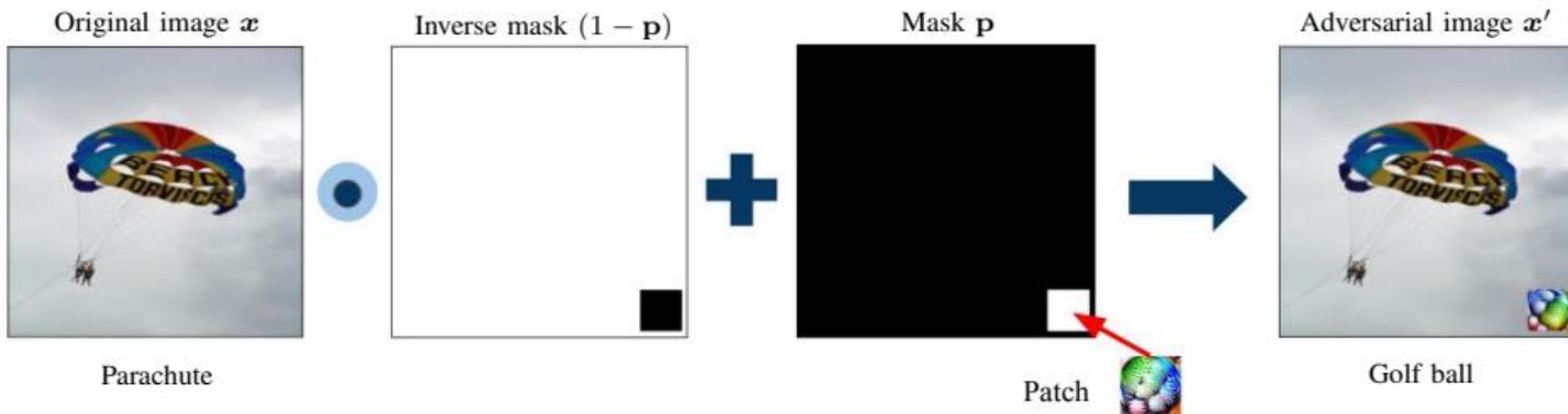
## 2.3.1 对抗补丁攻击

### 1、Adversarial Patch

使用Adversarial Patch制作的对抗样本通常可以表示为：

$$x' = (1 - \mathbf{p}) \odot \mathbf{x} + \mathbf{p} \odot \delta$$

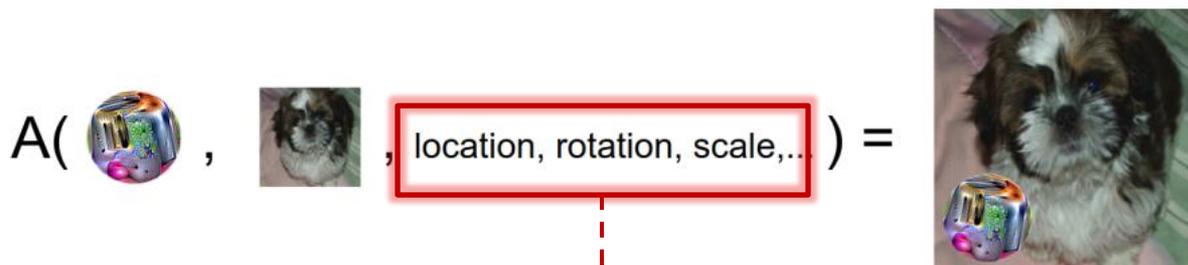
扰动的掩膜    干净样本    扰动的噪声



## 2.3.1 对抗补丁攻击

### 1、Adversarial Patch

考虑到物理世界中的各种细微扰动，如旋转平移以及补丁放置的误差，可能影响对抗补丁的效果，Adversarial Patch在优化阶段使用了Expectation over Transformation (EOT)技术来增强补丁的攻击鲁棒性



$$\hat{p} = \arg \max_p \mathbb{E}_{x \sim X, t \sim T, l \sim L} [\log \Pr(\hat{y} | A(p, x, l, t))]$$

T: 一组变换参数分布 (旋转/尺度变换)

L: 位置变换参数分布 (Patch的放置位置)

## 2.3.1 对抗补丁攻击

### 2、Robust Physical Perturbations (PR2)

PR2设计了一种针对物理世界中**交通牌识别**的对抗补丁，并能够抵抗物理世界中对视角相机的距离和角度变化

补丁的形式为黑白纯色的贴纸，仿照现实中交通牌可能出现的涂鸦和遮挡



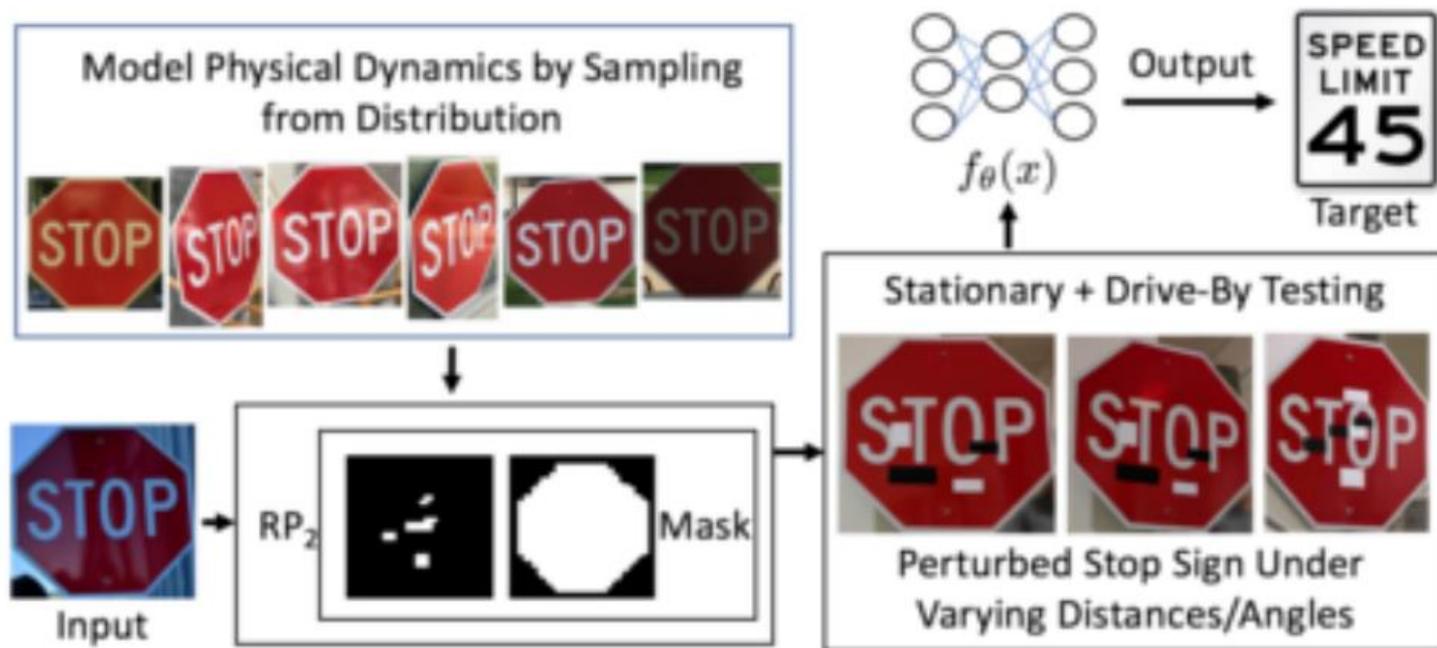
左图是一个停车标志上的真实涂鸦、大多数人不会认为这是可疑的。右图显示的是我们对停车标志进行的物理扰动。我们设计扰动来模仿涂鸦、从而将其“隐藏在人类的心灵深处”。

## 2.3.1 对抗补丁攻击

### 2、Robust Physical Perturbations (PR2)

PR2的Pipeline如下图所示：

- **抵御物理世界中的相机波动：**PR2考虑了一组相机旋转/平移/光照强度的变换对干净样本进行增强，使得对抗样本能够在各种波动情况下有效
- **空间限制：**将对抗贴纸的优化范围限定在交通牌区域，不针对考虑背景扰动
- **降低制造误差：**由于颜色的不可打印和混淆性，PR2采用最简单的黑白色贴纸来制造较大的视觉误差，以实现更有效的物理攻击



## 2.3.1 对抗补丁攻击

### 3、对抗性贴纸 (Adversarial sticker)

Adversarial sticker是一类**针对位置优化**的对抗补丁，它使用日常生活中常见的贴纸作为补丁纹理，仅依靠粘贴位置的优化，成功实现在针对人脸识别、交通牌识别、图像检索等多个任务上的对抗攻击



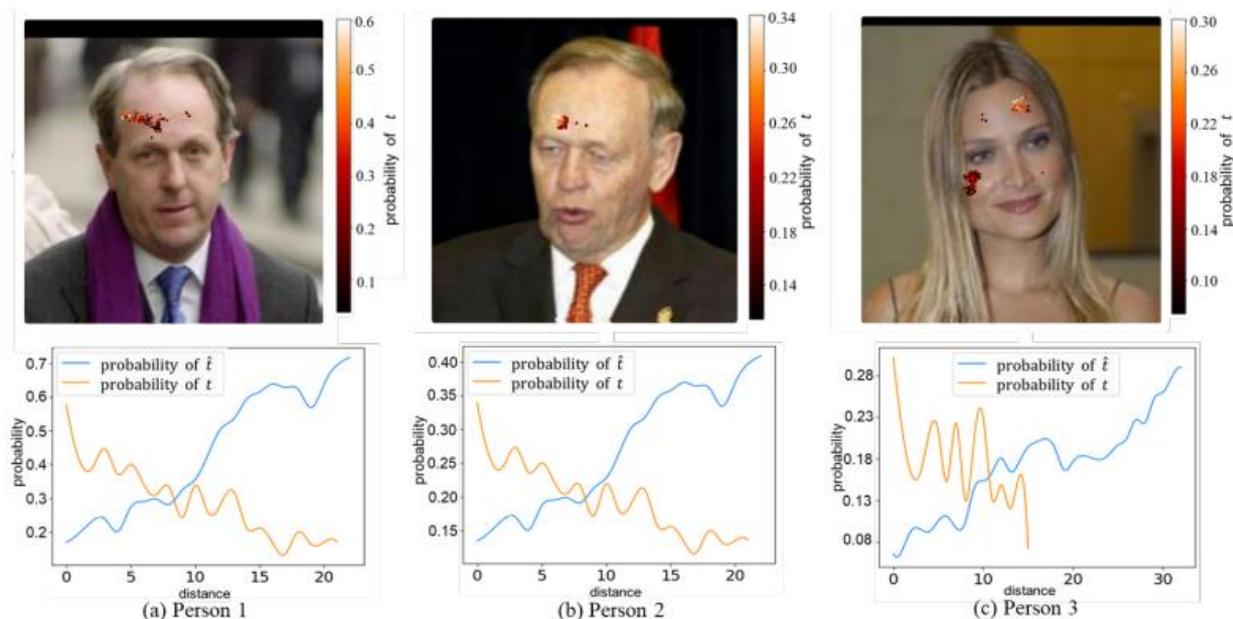
在日常生活中，人脸和交通标志牌存在贴纸的案例

## 2.3.1 对抗补丁攻击

### 3、对抗性贴纸 (Adversarial sticker)

$$\mathbf{x}^{adv} = (1 - \mathcal{A}) \odot \mathbf{x} + \mathcal{A} \odot \tilde{\mathbf{x}}$$

一般的对抗补丁方法旨在优化Patch纹理以实现其对抗性，并固定掩膜与之相反，Adversarial sticker使用常见的贴纸固定补丁纹理，并优化掩膜（即贴纸的粘贴位置）为了证明位置信息足够产生对抗性，做了以下探究：



- 依靠位置能够实现攻击
- 对抗性位置存在区域聚集性

## 2.3.1 对抗补丁攻击

### 3、Adversarial sticker

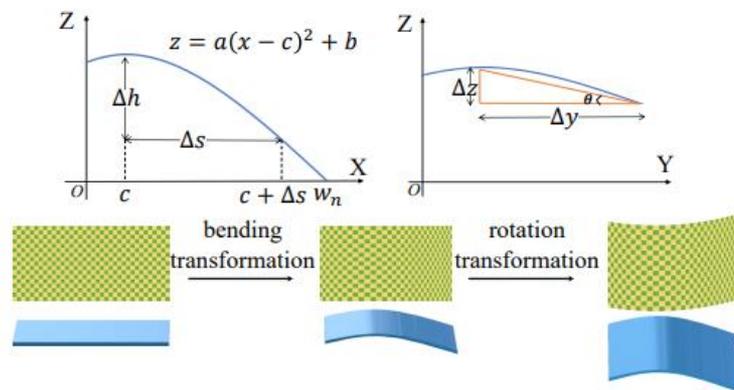
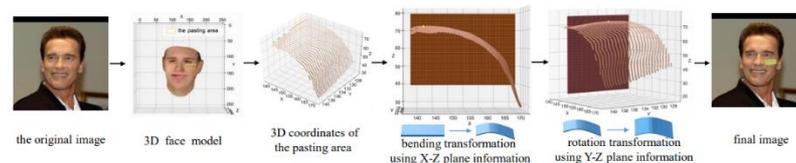
根据以上发现，本文设计了一种**基于区域**的差分进化算法（Region based Heuristic Differential Evolution, RHDE）优化对抗性贴纸的位置。并根据面部的3D建模对贴纸进行3D变换以获得更自然的对抗样本

#### Algorithm 1 Region based Heuristic Differential Evolution Algorithm

**Input:** Network  $f(\cdot)$ , face image  $x$  and label  $\hat{t}$ , the attack objective function  $\mathcal{L}(\theta)$ , the number of parameters  $d$ , value range  $(\theta^L, \theta^U)$ , population size  $P$ , maximum number of iterations  $T$ , hyperparameter  $l, r, \alpha, \rho, \delta$

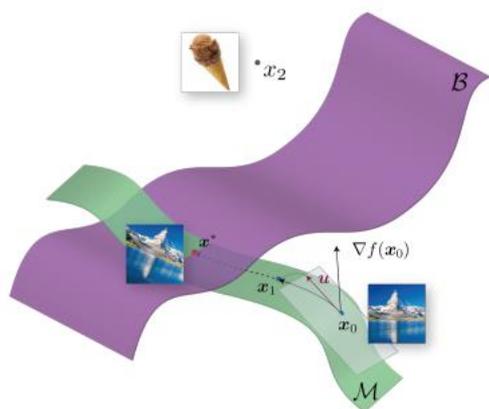
**Output:**  $\theta^*$

- 1: Initialize  $\mathbf{X}(0)$  randomly in  $[\theta_i^L, \theta_i^U]$  ( $1 \leq i \leq d$ ),  
 $\mathcal{J}(\theta) = \mathcal{L}(\theta)$ ,  $flag = 0$ ,  $stop = T$ ;
- 2: **for**  $k = 0$  to  $T - 1$  **do**
- 3:   Sort  $\mathbf{X}(k)$  in ascending order according to  $\mathcal{J}(\theta)$ ;
- 4:   **if**  $\mathbf{X}_0(k)$  makes the attack successful **then**
- 5:      $stop = k$ ; **break**;
- 6:   **end if**
- 7:   Generate candidate population  $\mathbf{C}(k)$   
    **if**  $i \in [1, \mu * P]$   $\mathbf{C}_i(k) \leftarrow$  according to Eq. (6)  
    **if**  $i \in [\mu * P + 1, P]$   $\mathbf{C}_i(k) \leftarrow$  according to Eq. (5)
- 8:   **if**  $(t_1(\mathbf{C}_{\gamma^*}(k)) == \hat{t}$  and  $flag == 0)$  **then**
- 9:      $bound \leftarrow$  according to Eq. (7)
- 10:    **if**  $bound \leq \delta$  **then**
- 11:      $flag = 1$ ,  $\tau = t_2$ ; Update  $\mathcal{J}(\theta)$  according to Eq. (8)
- 12:    **end if**
- 13:    **end if**
- 14:     $\mathbf{X}_i(k + 1) \leftarrow$  the better one between  $\mathbf{X}_i(k)$  and  $\mathbf{C}_i(k)$
- 15: **end for**
- 16: Sort  $\mathbf{X}(stop)$  in ascending order according to  $\mathcal{J}(\theta)$ ;
- 17: **return**  $\mathbf{X}_0(stop)$



## 2.3.2 基于自然变换攻击

基于自然变换 (Natural Transformation) 的对抗样本与人为恶意制造的对抗样本不同。它们是在物理环境中可能出现的分布外 (Out-of-Distribution, OOD) 样本, 如几何形变 / 仿射变换 / 图像损坏 / 视角变换等。通过操控变换参数实现对攻击的目的。



Natural



“revolver”

Adversarial



“mousetrap”



Keyboard: 47.80%



Street sign: 99.55%



Mouse: 37.44%



Cinema: 58.45%

- 仿射变换

- 旋转/平移

- 视角变换

Gaussian Noise



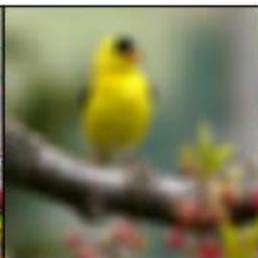
Shot Noise



Impulse Noise



Defocus Blur



Frosted Glass Blur



- 不同类型的图像损坏

## 2.3.2 基于自然变换攻击

### 1、ViewFool

ViewFool设计了一种系统性生成对抗性视角图像的攻击方法，它基于NeRF对物体进行三维表示，使其具备对物理世界的物体生成对抗性视角渲染的能力，并提出ImageNet-V数据集，用于视觉模型的视角鲁棒性评估



Chair: 78.66%



Keyboard: 47.80%



Street sign: 99.55%



Traffic light: 97.94%



Board: 63.50%



Mouse: 37.44%



Cinema: 58.45%



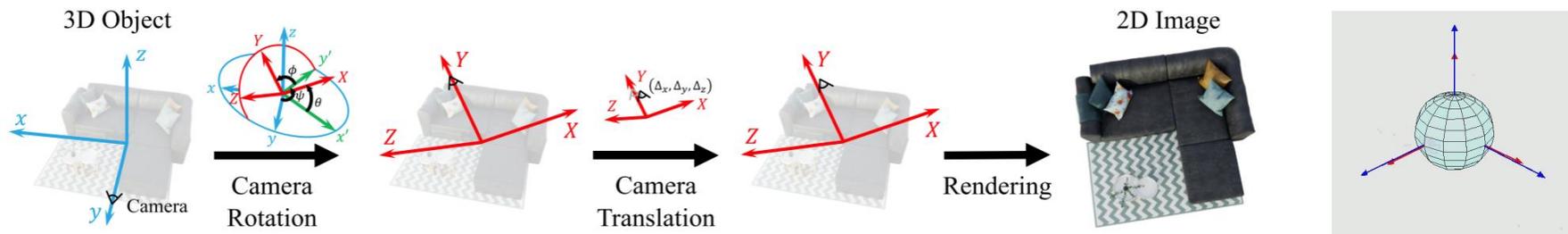
Canoe: 41.01%



## 2.3.2 基于自然变换攻击

### 1、ViewFool

**视角变换参数化：** ViewFool将物体的视角表示为一组6维向量  $\mathbf{v} = [\theta, \varphi, \gamma, \Delta_x, \Delta_y, \Delta_z]$ ，由三个遵循泰勒-布莱恩约定的z-y-X轴欧拉旋转角，和三个轴向的偏移量组成，这组参数能够通过输入NeRF获得物体在该视角参数下的渲染图像



**优化目标：** 优化每个视角参数的独立高斯分布  $p(\mathbf{v}) \sim N(\mathbf{u}, \sigma^2)$

$$\max_{p(\mathbf{v})} \left\{ \mathbb{E}_{p(\mathbf{v})} [\mathcal{L}(f(\mathcal{R}(\mathbf{v})), y)] + \lambda \cdot \mathcal{H}(p(\mathbf{v})) \right\}$$

- 优化目标的第一项是**分类损失**：即交叉熵损失期望，这是为了保证对抗视角分布的攻击性
- 优化目标的第二项是**熵正则损失**：即分布熵，保持对抗视角分布始终在一个较大的范围，这是为了保证优化到的结果是一个较大的对抗视角区域，在该区域下具有抗相机波动能力

## 2.3.2 基于其他媒介攻击

### 1、针对3D人脸识别的光照攻击

该工作采用光学噪声实现物理环境的3D人脸识别攻击

- 将结构光三维扫描仪作为攻击目标，通过投影仪将对抗光照打在人脸实现攻击
- 引入3D不变损失提升物理变换下攻击的鲁棒性
- 针对基于点云的模型和基于深度图的人脸识别模型展开实验

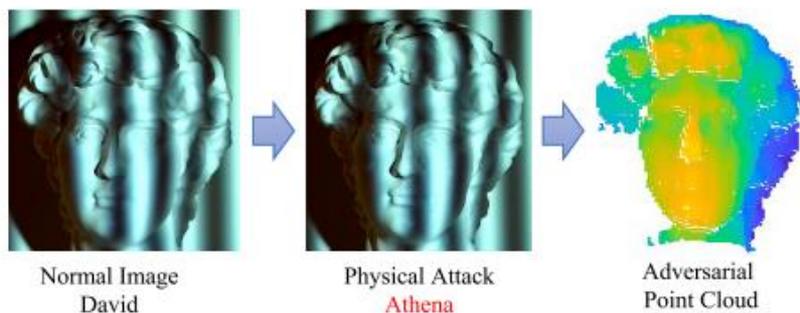


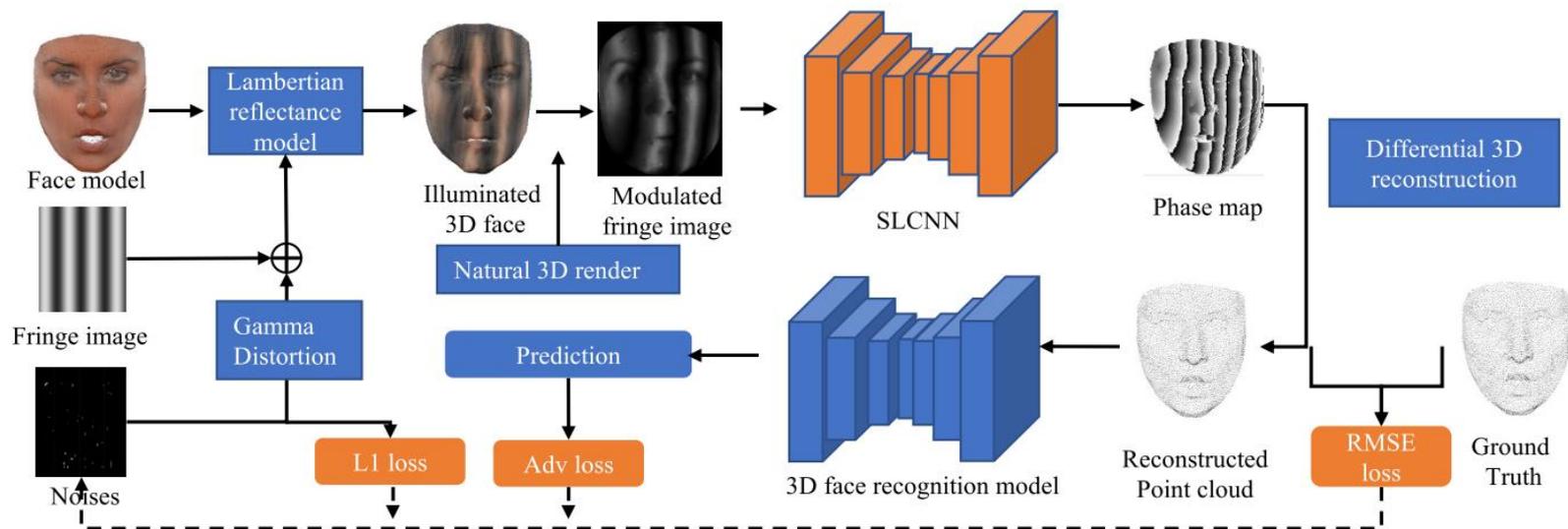
Figure 1. An demonstration of our attacks. We project additional noises on the fringe images to generate adversarial point clouds. Our attacks do not need the adversarial points strictly adjacent to the 3D surface and therefore need to modifier fewer points than state-of-the-art physical 3D adversarial attacks.



Figure 4. The physical settings of the phase superposition attack. The adversary uses an additional projector to add noises to the original fringe images for 3D reconstructions.

## 2.3.2 基于其他媒介攻击

### 1、针对3D人脸识别的光照攻击



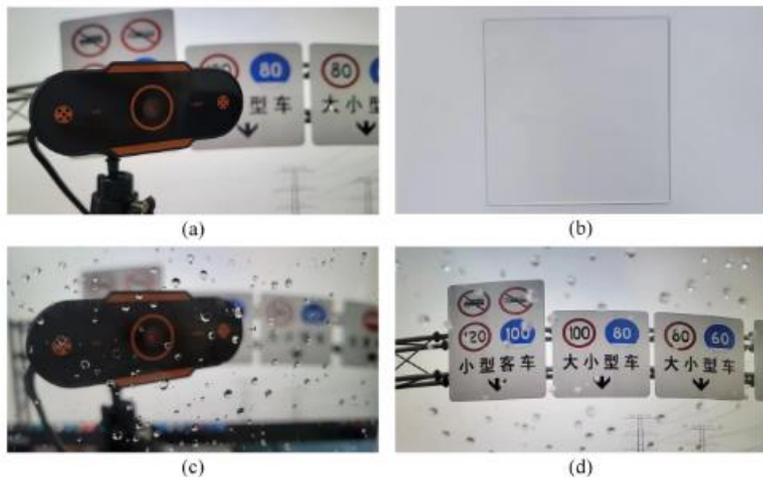
- 利用渲染器渲染光照下3D人脸的调制条纹图像，并输入SLCNN，并获得重建后的3D点云
- 将重建3D点云与GT的点云计算RMSE重建损失
- 输入3D人脸识别模型，计算对抗性损失(CE)
- 以及光学噪声和Gamma失真后的L1范数损失
- 采用三个损失共同指导光学噪声图的优化，得到对抗性的噪声扰动

## 2.3.2 基于其他媒介攻击

### 2、AdvRD: 针对图像识别的雨点模拟攻击

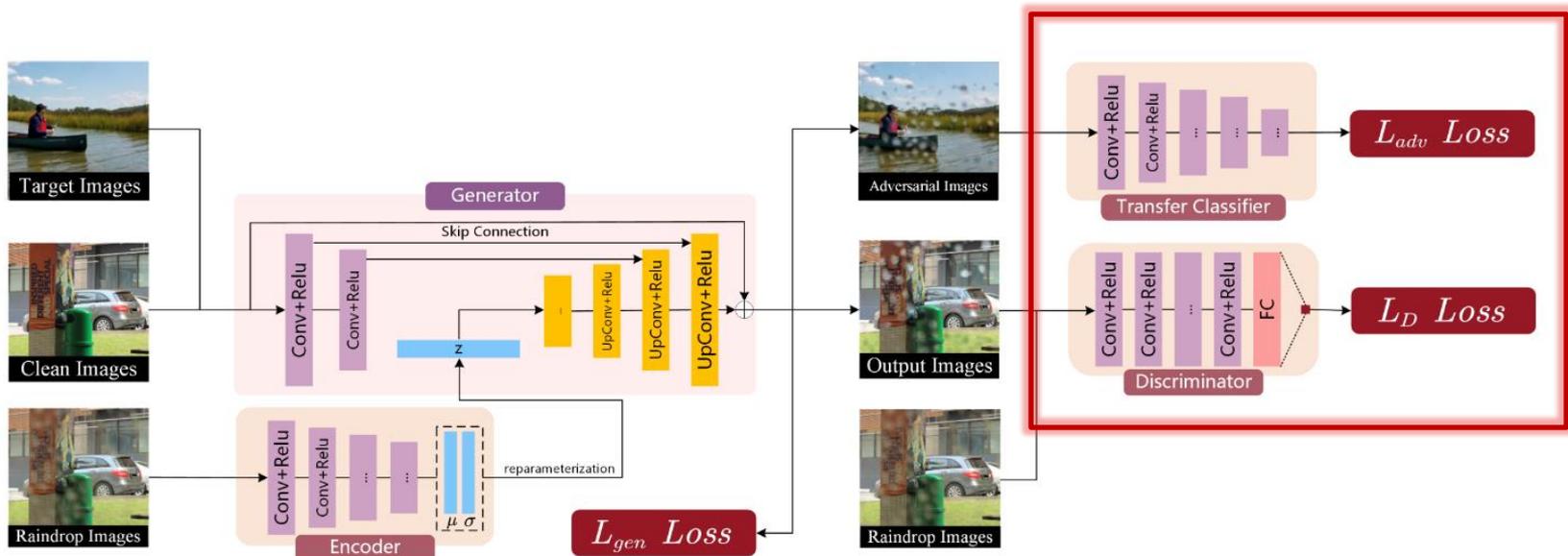
该工作采用GAN模拟对抗性的雨滴，作为攻击手段

- 针对自动驾驶场景可能出现的雨滴干扰，探究如何将雨滴作为攻击媒介，以及提升应对雨滴扰动的鲁棒性
- 提出一种在物理世界获取对抗性雨滴扰动的方法，并发现DNN对雨滴扰动的不鲁棒
- 提出AdvRD: 采用GAN作为风格转换器，将干净图像转换为逼真的雨滴场景图像，并引入对抗性损失优化对抗性雨滴图像



## 2.3.2 基于其他媒介攻击

### 2、AdvRD: 针对图像识别的雨点模拟攻击



#### 生成器的第一个训练目标：生成逼真的带雨滴图像

训练样本的浅层特征和雨滴图像计算的噪声向量 $z$ 进行融合，在鉴别器作用下使得输出接近真实雨滴，这一部分使用干净样本和对应雨滴样本的图像对进行训练，并额外添加了一个L1正则化损失（为了加快收敛）

#### 生成器的第二个训练目标：提升带雨滴图像的攻击能力

$L_c$ 为分类器 $c$ 的交叉熵损失，在此添加了一个超参数 $\eta$ ，平衡真实性和攻击成功率

## 2.3 Reference

---

- [1] Sharma A, Bian Y, Munz P, et al. Adversarial patch attacks and defences in vision-based tasks: A survey
- [2] Brown T B, Mané D, Roy A, et al. Adversarial patch
- [3] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification
- [4] Wei X, Guo Y, Yu J. Adversarial sticker: A stealthy attack method in the physical world
- [5] Dong Y, Ruan S, Su H, et al. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints
- [6] Engstrom L, Tran B, Tsipras D, et al. Exploring the landscape of spatial robustness
- [7] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum
- [8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations
- [9] Li Y, Li Y, Dai X, et al. Physical-World Optical Adversarial Attacks on 3D Face Recognition
- [10] Liu J, Lu B, Xiong M, et al. Adversarial Attack with Raindrops



北京航空航天大学  
BEIHANG UNIVERSITY

人工智能研究院  
Institute of Artificial Intelligence

# 感谢聆听！

Email: [xxwei@buaa.edu.cn](mailto:xxwei@buaa.edu.cn)

---