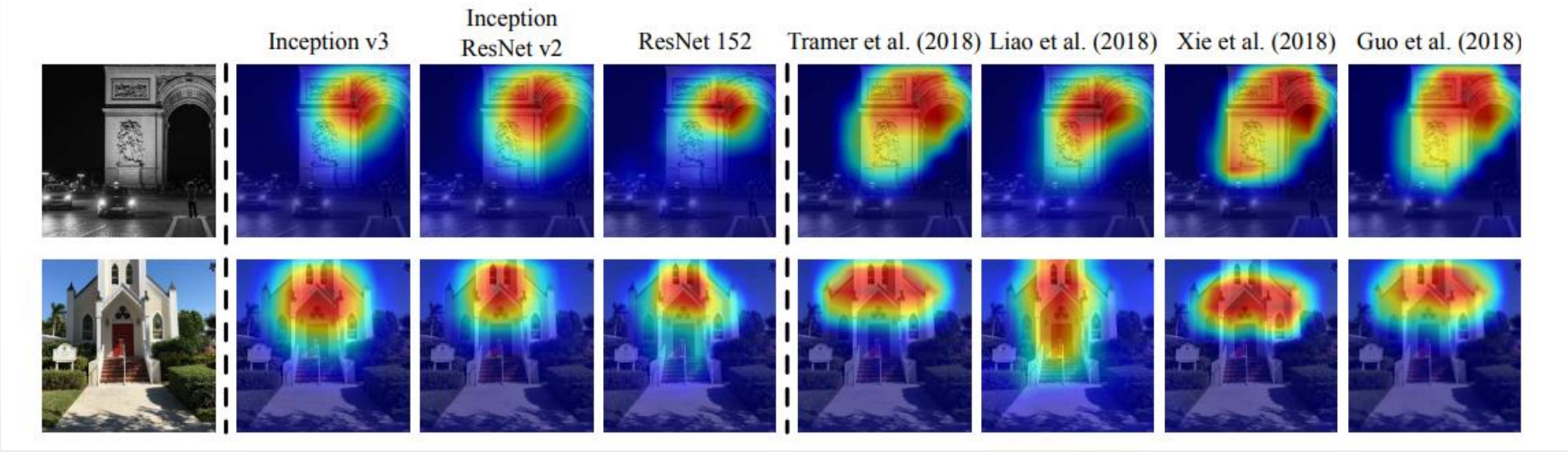
# Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks

CVPR-2019

Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu

Dept. of Comp. Sci. and Tech., BNRist Center, State Key Lab for Intell. Tech. & Sys.,
Institute for AI, THBI Lab, Tsinghua University, Beijing, 100084, China

# 动机




正常训练模型的注意力区域与防御模型的注意力区域有很大不同，当前基于可迁移的黑盒攻击通常使用单个输入生成白盒模型的对抗样本，生成的对抗样本与给定输入点的白盒模型的注意力区域或梯度高度相关，无法进行迁移攻击黑盒防御模型。

# 主要工作

提出了一种平移不变的攻击方法，通过源图像及其翻译版本组成的图像集合生成对抗样本，使得对抗样本对被攻击的白盒模型的注意力区域不太敏感，有更高概率欺骗另一个具有防御机制的黑盒模型。

# 约束条件

对抗样本的约束：$\arg\max_{\boldsymbol{x}^{adv}} J(\boldsymbol{x}^{adv}, y), \quad \text{s.t. } \|\boldsymbol{x}^{adv} - \boldsymbol{x}^{real}\|_\infty \leq \epsilon.$

平移对抗样本约束：

$$\arg\max_{\boldsymbol{x}^{adv}} \sum_{i,j} w_{ij} J(T_{ij}(\boldsymbol{x}^{adv}), y),$$
$$\text{s.t. } \|\boldsymbol{x}^{adv} - \boldsymbol{x}^{real}\|_\infty \leq \epsilon,$$

$i, j \in \{-k, ..., 0, ..., k\}$

其中，$T_{ij}(x)$代表将图像x沿二维方向分别平移i和j个像素，即如果对像素(a,b)，$T_{ij}(x)_{a,b} = x_{a-i,b-j}$

# 梯度计算

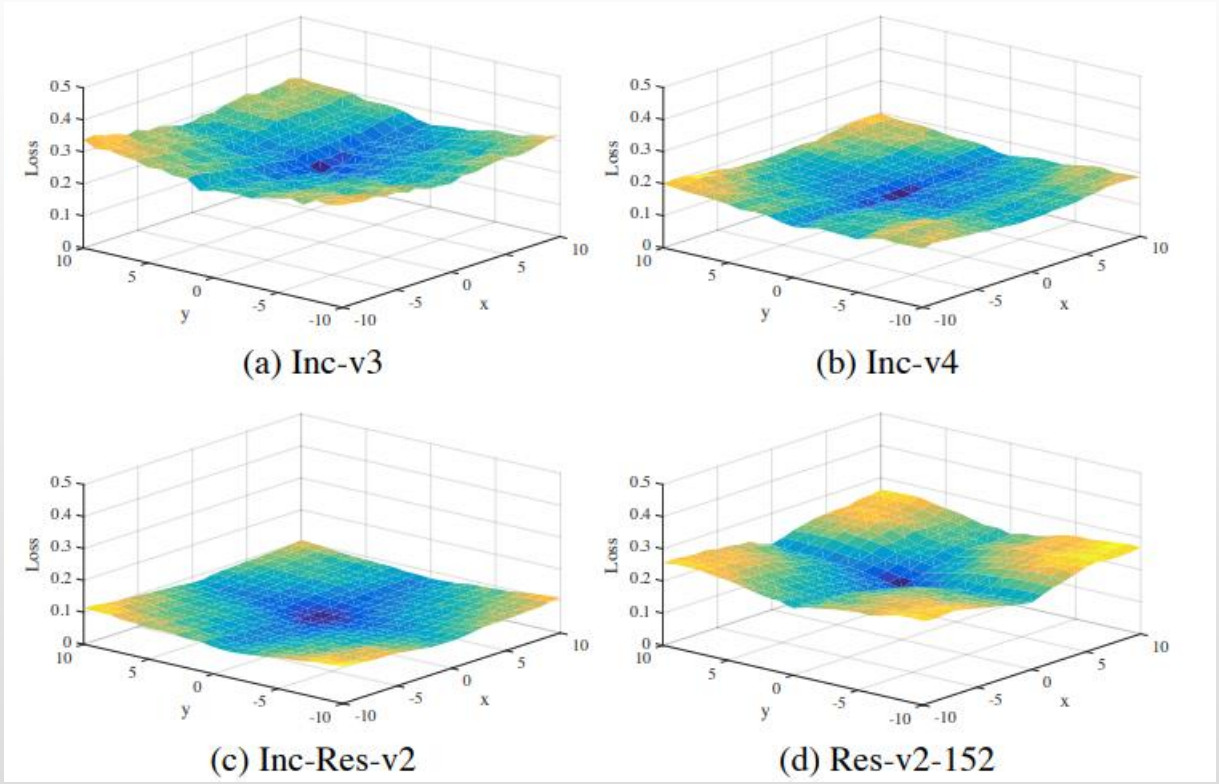
卷积神经网络具有平移不变性，故

$$\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=T_{ij}(\hat{\boldsymbol{x}})} \approx \nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}.$$

$$\begin{aligned}
&\nabla_{\boldsymbol{x}}\Big(\sum_{i,j} w_{ij} J(T_{ij}(\boldsymbol{x}), y)\Big)\Big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\
&= \sum_{i,j} w_{ij} \nabla_{\boldsymbol{x}} J(T_{ij}(\boldsymbol{x}), y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\
&= \sum_{i,j} w_{ij} \Big(\nabla_{T_{ij}(\boldsymbol{x})} J(T_{ij}(\boldsymbol{x}), y) \cdot \frac{\partial T_{ij}(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big)\Big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\
&= \sum_{i,j} w_{ij} T_{-i-j}\Big(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=T_{ij}(\hat{\boldsymbol{x}})}\Big) \\
&\approx \sum_{i,j} w_{ij} T_{-i-j}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}).
\end{aligned}$$

$$\sum_{i,j} w_{ij} T_{-i-j}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}) \Leftrightarrow \boldsymbol{W} * \nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}},$$

W是(2k+1)*(2k+1)大小的卷积核矩阵



# 设置卷积核矩阵W，优化梯度攻击

设计思路：较大的平移对应较小的权值；

三种设计：标准形式，线性形式，高斯形式；

(1) A uniform kernel that $W_{i,j} = {}^{1}/_{(2k+1)^2}$;

(2) A linear kernel that $\tilde{W}_{i,j} = (1 - {}^{|i|}/_{k+1}) \cdot (1 - {}^{|j|}/_{k+1})$, and $W_{i,j} = {}^{\tilde{W}_{i,j}}/_{\sum_{i,j} \tilde{W}_{i,j}}$;

(3) A Gaussian kernel that $\tilde{W}_{i,j} = \frac{1}{2\pi\sigma^2} \exp(-\frac{i^2+j^2}{2\sigma^2})$ where the standard deviation $\sigma = {}^{k}/_{\sqrt{3}}$ to make the radius of the kernel be $3\sigma$, and $W_{i,j} = {}^{\tilde{W}_{i,j}}/_{\sum_{i,j} \tilde{W}_{i,j}}$.

FGSM：$$\boldsymbol{x}^{adv} = \boldsymbol{x}^{real} + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{real}, y)),$$

TI-FGSM：$$\boldsymbol{x}^{adv} = \boldsymbol{x}^{real} + \epsilon \cdot \text{sign}(\boldsymbol{W} * \nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{real}, y)).$$

BIM：$$\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_{t}^{adv} + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_{t}^{adv}, y)),$$

TI-BIM：$$\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_{t}^{adv} + \alpha \cdot \text{sign}(\boldsymbol{W} * \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_{t}^{adv}, y)).$$

# 实验结果

MI-FGSM：

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y)}{\|\nabla_x J(x_t^{adv}, y)\|_1},$$

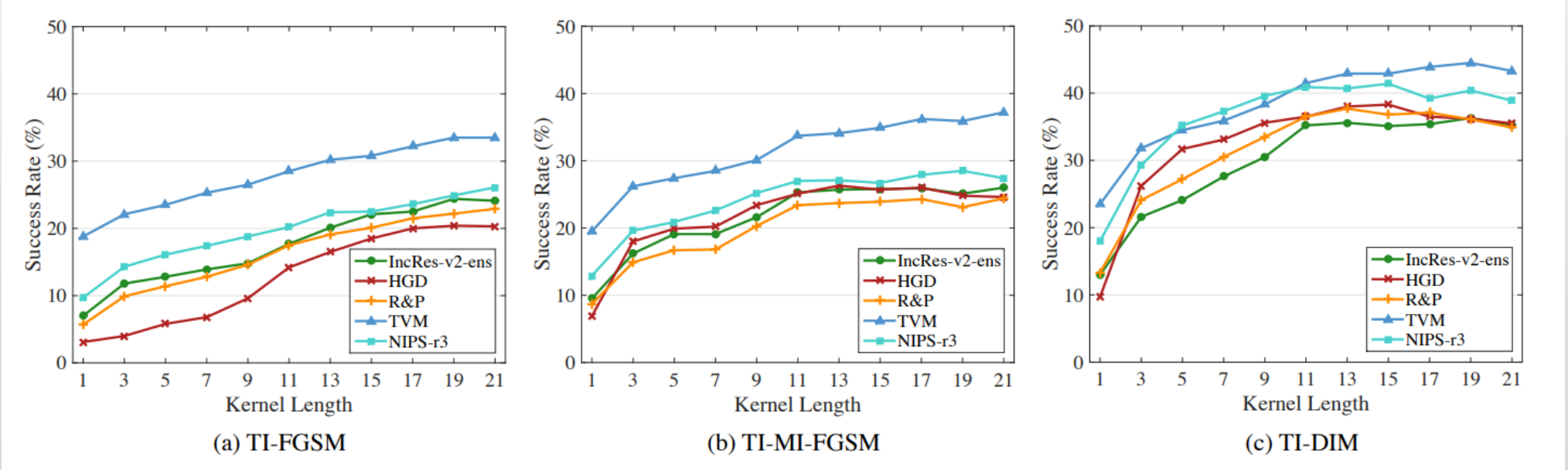
$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \mathrm{sign}(g_{t+1}),$$

DIM：利用梯度随机调整大小或者以给定概率填充

| | Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|---|
| TI-FGSM | Uniform | 25.0 | 27.9 | 21.1 | 15.7 | 19.1 | 24.8 | 32.3 | 21.9 |
| | Linear | **30.7** | **32.4** | **24.2** | **20.9** | **23.3** | **28.1** | **34.6** | **25.8** |
| | Gaussian | 28.2 | 28.9 | 22.3 | 18.4 | 19.8 | 25.5 | 30.7 | 24.5 |
| TI-MI-FGSM | Uniform | 30.0 | 32.2 | 22.8 | 21.7 | 22.8 | 26.4 | 32.7 | 25.9 |
| | Linear | **35.8** | 35.0 | **26.8** | 25.5 | 23.4 | **29.0** | **35.8** | **27.5** |
| | Gaussian | **35.8** | **35.1** | 25.8 | **25.7** | **23.9** | 28.2 | 34.9 | 26.7 |
| TI-DIM | Uniform | 32.6 | 34.6 | 25.6 | 24.1 | 27.2 | 30.2 | 34.9 | 28.8 |
| | Linear | 45.2 | 47.0 | 34.9 | 35.6 | 35.2 | **38.5** | 43.6 | 39.7 |
| | Gaussian | **46.9** | **47.1** | **37.4** | **38.3** | **36.8** | 37.0 | **44.2** | **41.4** |

# 实验结果

| | Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|---|
| Inc-v3 | FGSM | 15.6 | 14.7 | 7.0 | 2.1 | 6.5 | 19.9 | 18.8 | 9.8 |
| | TI-FGSM | **28.2** | **28.9** | **22.3** | **18.4** | **19.8** | **25.5** | **30.7** | **24.5** |
| Inc-v4 | FGSM | 16.2 | 16.1 | 9.0 | 2.6 | 7.9 | 21.8 | 19.9 | 11.5 |
| | TI-FGSM | **28.2** | **28.3** | **21.4** | **18.1** | **21.6** | **27.9** | **31.8** | **24.6** |
| IncRes-v2 | FGSM | 18.0 | 17.2 | 10.2 | 3.9 | 9.9 | 24.7 | 23.4 | 13.3 |
| | TI-FGSM | **32.8** | **33.6** | **28.1** | **25.4** | **28.1** | **32.4** | **38.5** | **31.4** |
| Res-v2-152 | FGSM | 20.2 | 17.7 | 9.9 | 3.6 | 8.6 | 24.0 | 22.0 | 12.5 |
| | TI-FGSM | **34.6** | **34.5** | **27.8** | **24.4** | **27.4** | **32.7** | **38.1** | **30.1** |




(a) TI-FGSM (b) TI-MI-FGSM (c) TI-DIM

# 实验结果

| Attack | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | HGD | R&P | JPEG | TVM | NIPS-r3 |
|---|---|---|---|---|---|---|---|---|
| FGSM | 27.5 | 23.7 | 13.4 | 4.9 | 13.8 | 38.1 | 30.0 | 19.8 |
| TI-FGSM | **39.1** | **38.8** | **31.6** | **29.9** | **31.2** | **43.3** | **39.8** | **33.9** |
| MI-FGSM | 50.5 | 48.3 | 32.8 | 38.6 | 32.8 | 67.7 | 50.1 | 43.9 |
| TI-MI-FGSM | **76.4** | **74.4** | **69.6** | **73.3** | **68.3** | **77.2** | **72.1** | **71.4** |
| DIM | 66.0 | 63.3 | 45.9 | 57.7 | 51.7 | 82.5 | 64.1 | 63.7 |
| TI-DIM | **84.8** | **82.7** | **78.0** | **82.6** | **81.4** | **83.4** | **79.8** | **83.1** |

# Feature Space Perturbations Yield More Transferable Adversarial Examples

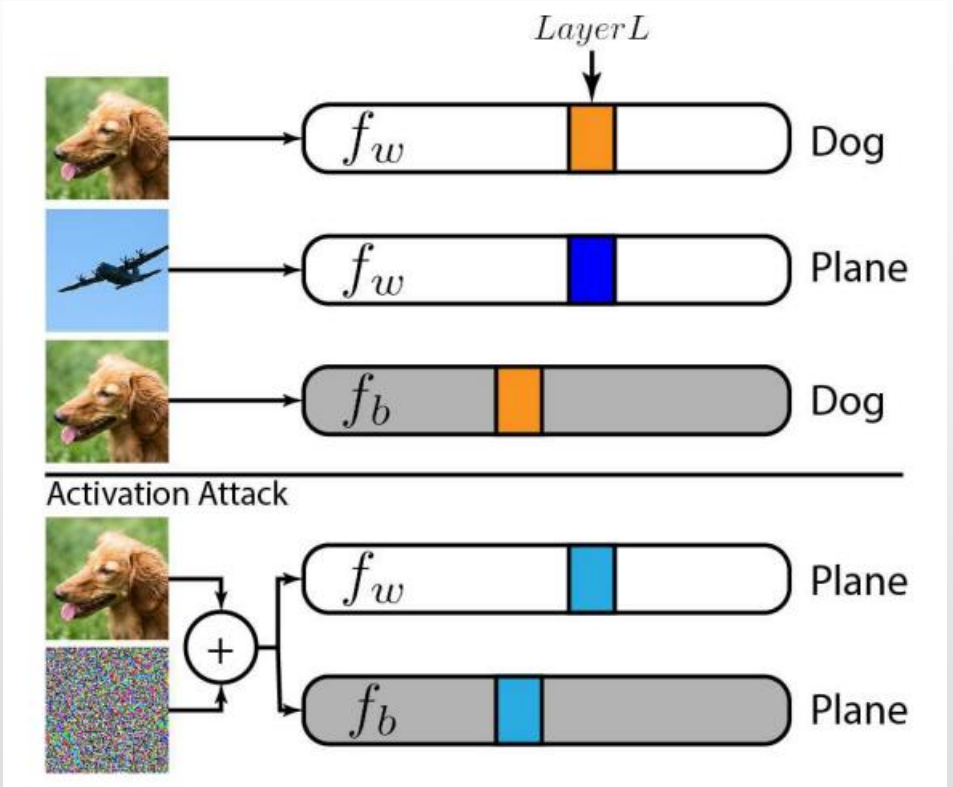
CVPR-2019

Nathan Inkawhich, Wei Wen, Hai (Helen) Li and Yiran Chen

Duke University

# 主要工作

利用模型的中间特征是可迁移的，设计了一种针对深度特征空间表示的基于传输的黑盒激活攻击方法（AA）。通过将源图像上白盒模型的L层特征空间表示驱动到目标图像的L层特征空间表示，生成了高度可迁移的对抗样本。



# LOSS计算

$I_s$：源图像

$\mathrm{I_t}$：目标图像

$A_s^L$：源图像L层激活

$A_\mathrm{t}^L$：目标图像L层激活

$$J_{AA}(I_t, I_s) = \|f_L(I_t) - f_L(I_s)\|_2 = \left\|A_t^L - A_s^L\right\|_2.$$

即L 层源图像激活和目标图像激活之间的欧几里德距离

$$m_{k+1} = m_k + \frac{\nabla_{I_k} J_{AA}(I_t, I_k)}{||\nabla_{I_k} J_{AA}(I_t, I_k)||_1},$$

$$I_{k+1} = Clip(I_k - \alpha * sign(m_{k+1}), 0, 1).$$

三种因素：

1、深度特征空间表示的调整对分类结果有相当大的影响；

2、由于深度模型的中间层特征已证明是可迁移的，特征空间中的显式攻击将产生可迁移的对抗样本；

3、假设使用相同分布的数据训练的两个不同模型学习的决策边界和类方向是相似的；

# 评价指标

错误率（error）

非目标迁移率（uTR）：

$$uTR = \frac{1}{|D_{uTR}|} \sum_{\substack{(x_{adv}, y_{true}) \\ \in D_{uTR}}} \mathbb{1}[(f_b(x_{adv})) \neq y_{true})],$$

目标成功率（tSuc）

目标迁移率（tTR）：

$$tTR = \frac{1}{|D_{tTR}|} \sum_{\substack{(x_{adv}, y_{target}) \\ \in D_{tTR}}} \mathbb{1}[(f_b(x_{adv})) = y_{target})].$$

# 实验结果

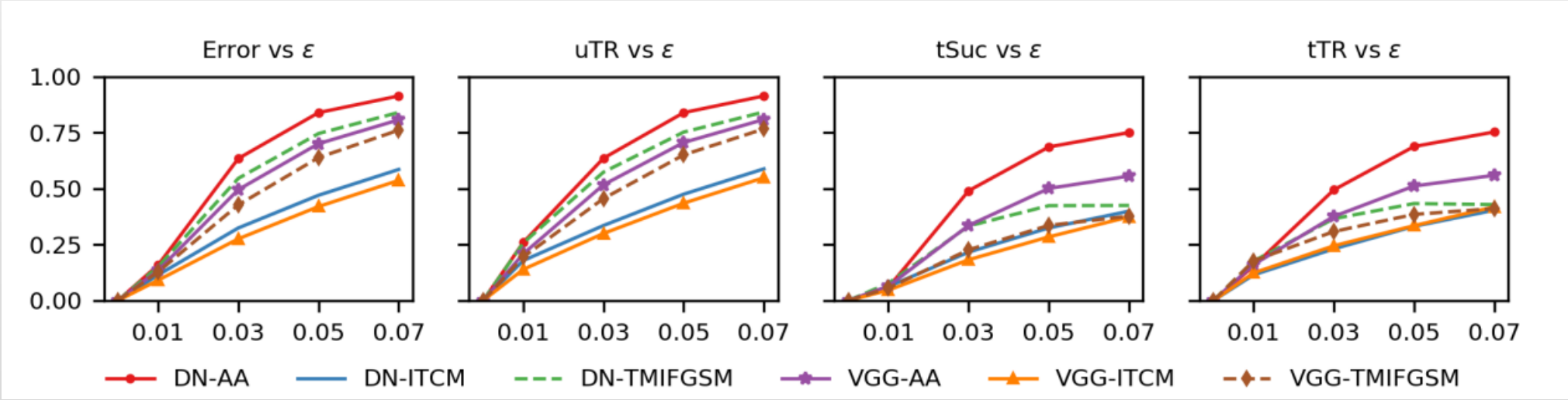
数据集：CIFAR-10，ImageNet(IN)

DN：DenseNet121

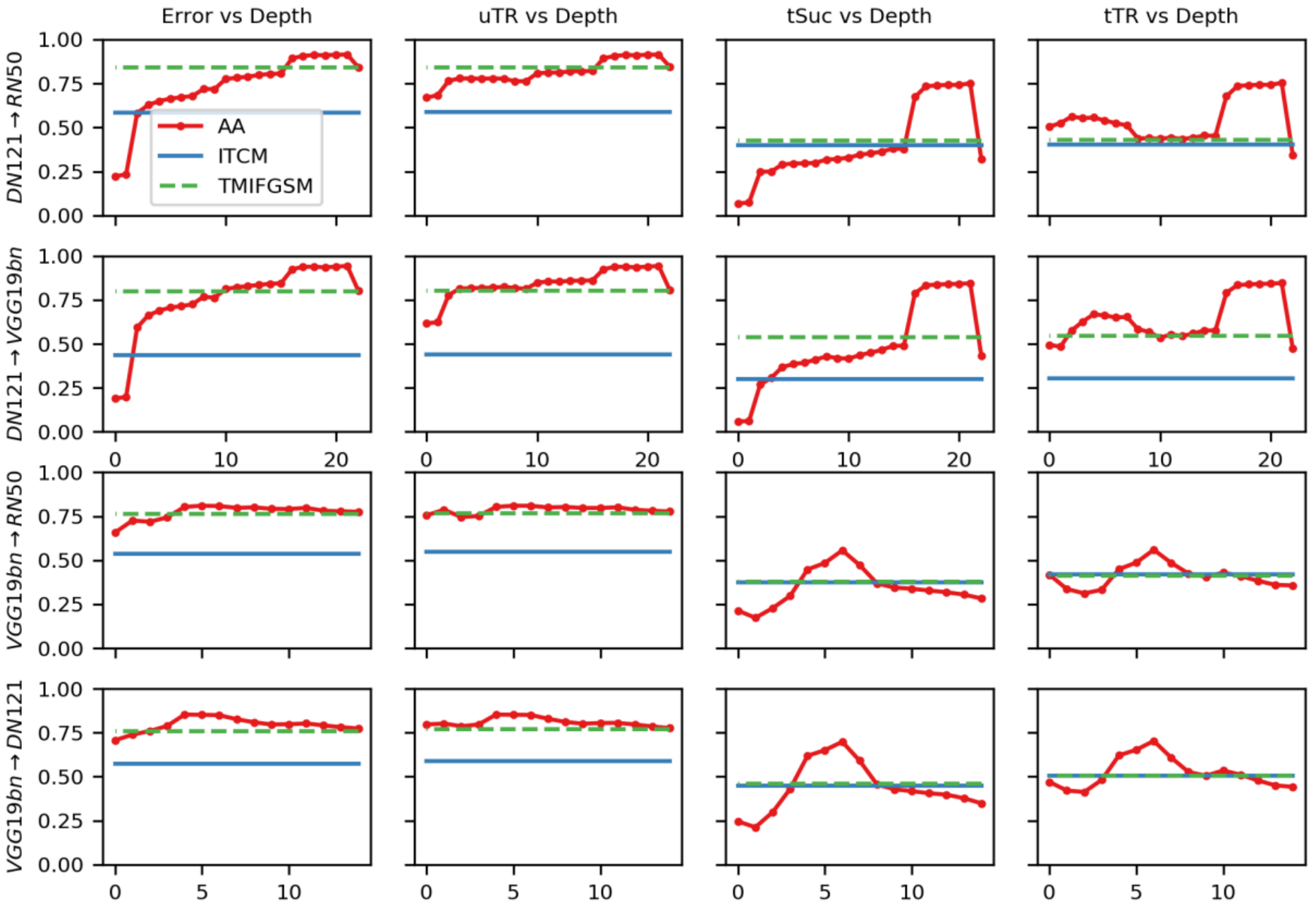
VGG：VGG19bn

黑盒模型：ResNet-50（RN50）

| Base | Attack | Error | uTR | tSuc | tTR |
|---|---|---|---|---|---|
| DN | ITCM | 58.62 | 58.88 | 39.97 | 40.39 |
| | TPGD | 61.33 | 61.52 | 35.84 | 36.15 |
| | TMIFGSM | 84.12 | 84.32 | 42.53 | 42.90 |
| | $AA_{L=21}$ | 91.49 | 91.48 | 75.13 | 75.38 |
| VGG | ITCM | 53.72 | 54.94 | 37.40 | 41.88 |
| | TPGD | 58.85 | 59.55 | 36.60 | 39.15 |
| | TMIFGSM | 76.19 | 76.75 | 37.72 | 41.14 |
| | $AA_{L=6}$ | 80.81 | 80.97 | 55.64 | 55.98 |
| $DN_{IN}$ | ITCM | 21.14 | 21.18 | 0.99 | 1.01 |
| | TPGD | 25.23 | 25.29 | 0.94 | 0.97 |
| | TMIFGSM | 47.75 | 47.76 | 2.25 | 2.25 |
| | $AA_{L=7}$ | 80.58 | 81.77 | 2.57 | 8.63 |



# 实验结果


Error vs Depth
uTR vs Depth
tSuc vs Depth
tTR vs Depth
DN121→RN50
DN121→VGG19bn
VGG19bn→RN50
VGG19bn→DN121
AA
ITCM
TMIFGSM
1.00
0.75
0.50
0.25
0.00
0
5
10
20

# 为什么不同层的攻击效果不同

特征图平均角距离