Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

偷走你的训练数据：模型反演攻击方法研究

硕士研究生 皮佳伟

2024年01月28日

问题回溯

- 相关内容
  - 2023.03.05 张辰龙《深度神经网络模型窃取检测》
  - 2022.10.16 程瑶《成员推理攻击》
  - 2022.07.17 丁扬《面向生成模型的模型窃取方法》

内容提要

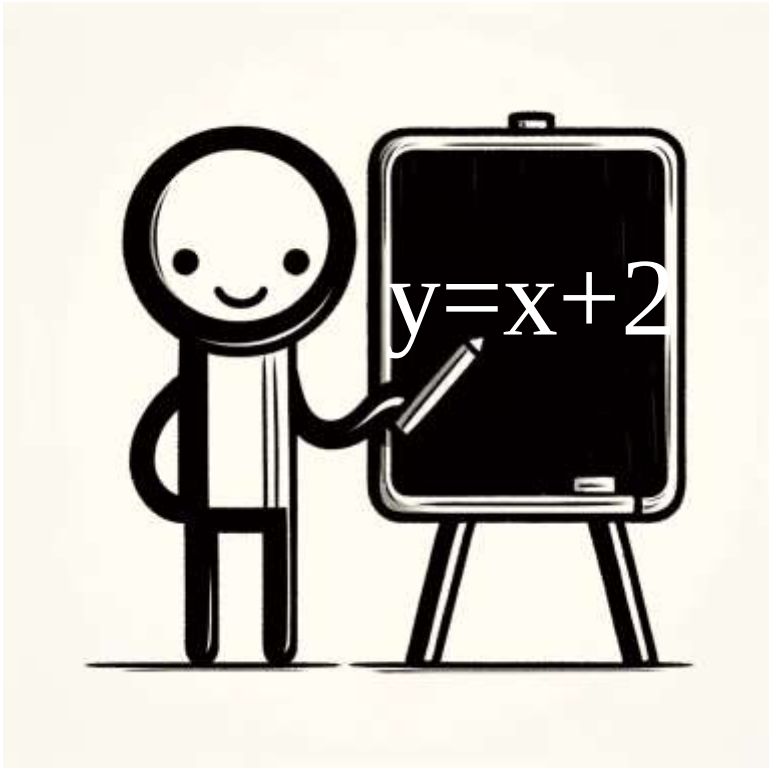
- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - GMI
  - SMI
- 特点总结与工作展望
- 参考文献

背景简介

- 预期收获
  - 掌握模型反演攻击的相关知识
  - 了解白盒模型反演攻击方法
  - 了解黑盒模型反演攻击方法

内涵解析与研究目标

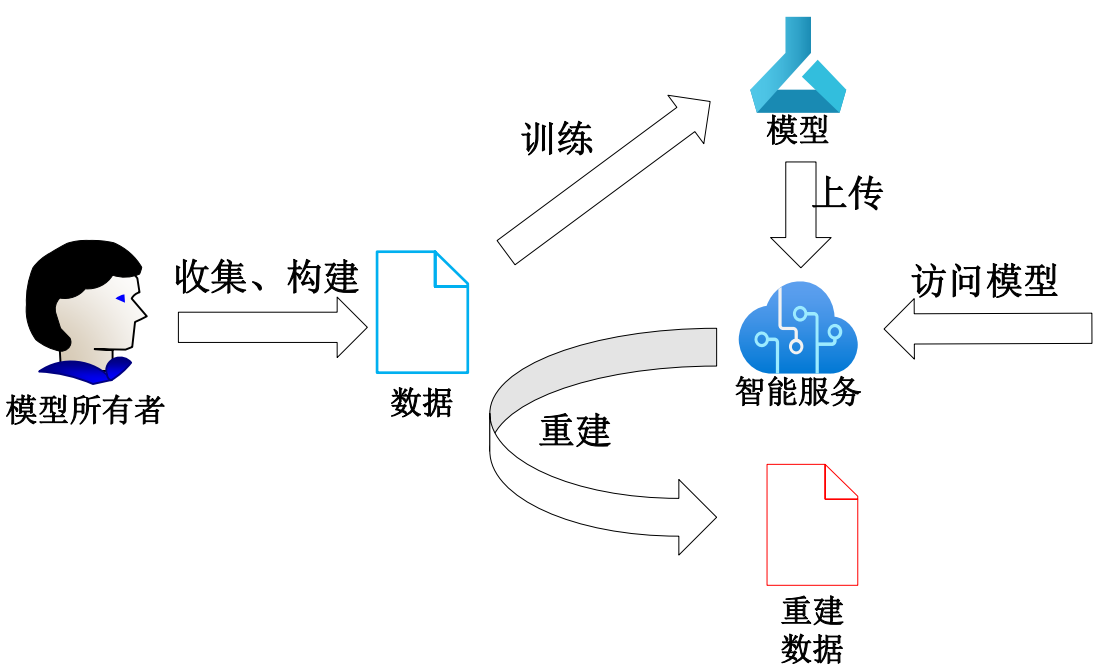
- 题目内涵解析（模型反演攻击方法研究）
  - 反演：即逆转，如何由输出得到输入
  - 模型反演：反演在深度学习模型上的体现
  - 模型反演攻击：通过特殊设计的算法，重建目标模型的私有训练样本，进而造成敏感信息的泄露



- 研究目标
  - 面向深度学习模型的隐私安全研究
  - 研究目标模型特征迁移、特征空间搜索优化、生成样本质量评估等关键问题
  - 结合对抗生成网络、扩散模型、梯度下降算法、黑盒优化策略等理论
  - 重建目标模型私有训练样本，揭示模型训练数据所面临的隐私安全问题

研究背景

- 研究背景
  - 深度学习技术发展迅速，在图像识别、自然语言处理等领域发挥重要作用
  - 特定任务的深度学习模型需要特定的数据样本进行训练
  - 数据样本通常包含各种敏感信息、或者所有者涉及知识产权不愿公开
  - 模型反演攻击能够针对目标模型重建训练数据样本，导致严重的隐私泄露



研究意义

- 研究意义
  - 验证数据泄露风险
    - 重建私有训练样本
    - 验证模型面临训练数据泄露风险
  - 促进防御方法发展
    - 揭示白盒、黑盒等各个场景模型面临的模型反演风险
    - 验证差分隐私等传统防御方式对模型反演攻击的防御无效性
    - 促进对各场景下模型反演防御方法的研究，保证人工智能隐私安全

重建私有训练样本，促进防御方法发展，保障模型隐私安全

研究历史

2014 Fredrikson等人首次提出了模型反演攻击的概念，利用最大后验估计器，对基因隐私相关的线性回归模型进行反演，获得了患者基因隐私数据

2020 Zhang等人针对已有的攻击方法难以应用于稍复杂的深度神经网络问题，提出白盒场景下的生成模型反演攻击；首次强调了辅助数据集的重要性，并取得了较好的实验结果

2021 Chen等人提出依据任务调整GAN进行训练，显式建模私有数据的特征分布方法，成功提高了攻击准确率，并保证重建的样本人眼真实性高

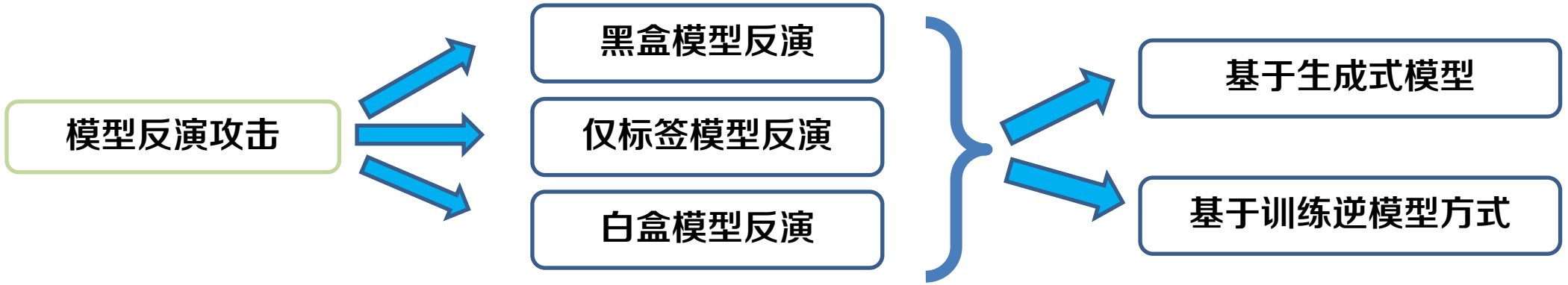
2021 Zhao等人针对已有的研究没有针对可解释模型的算法，提出多种利用模型揭示信息的模型的反演攻击算法，开拓了模型反演算法的应用场景

2022 Kahla等人针对没有适用于仅标签场景下的算法，提出一种利用置信度差异和梯度估计器的方法，验证了仅标签模型也面临数据泄露风险

2022 An等针对特征空间耦合，在特征空间中搜索优化潜在向量困难，导致重建样本质量不佳的问题，提出利用StyleGAN进行变换进而解耦合

2023 Nguyen等人针对性分析了已有的模型反演攻击算法面临的共同问题，提出更符合反演任务的优化目标以及避免MI过拟合问题的措施

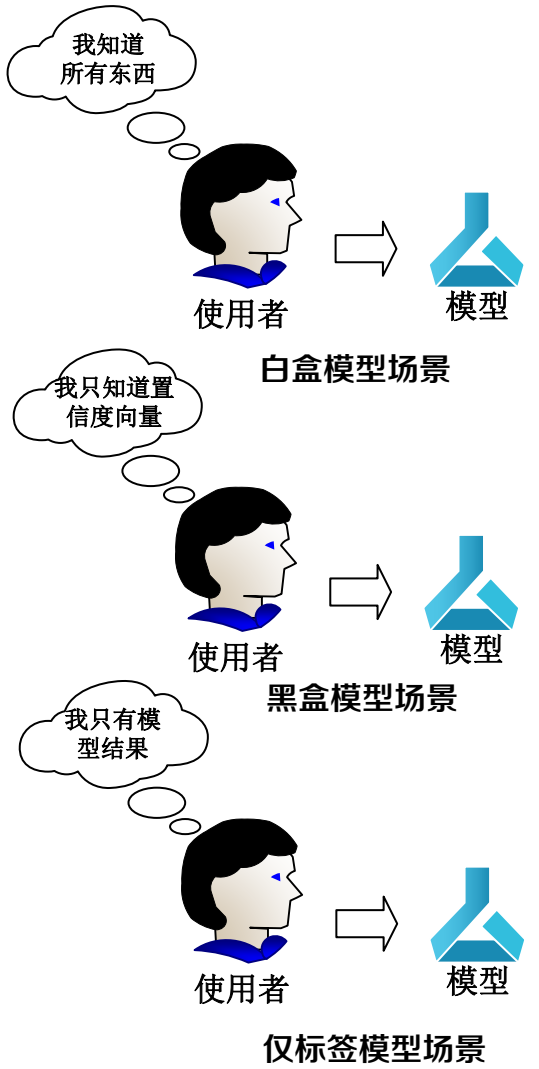
2023 Tian等人针对已有方法过度依赖先验信息的问题，提出充分利用类别信息所隐藏的内容实现模型反演



知识基础　三个模型场景

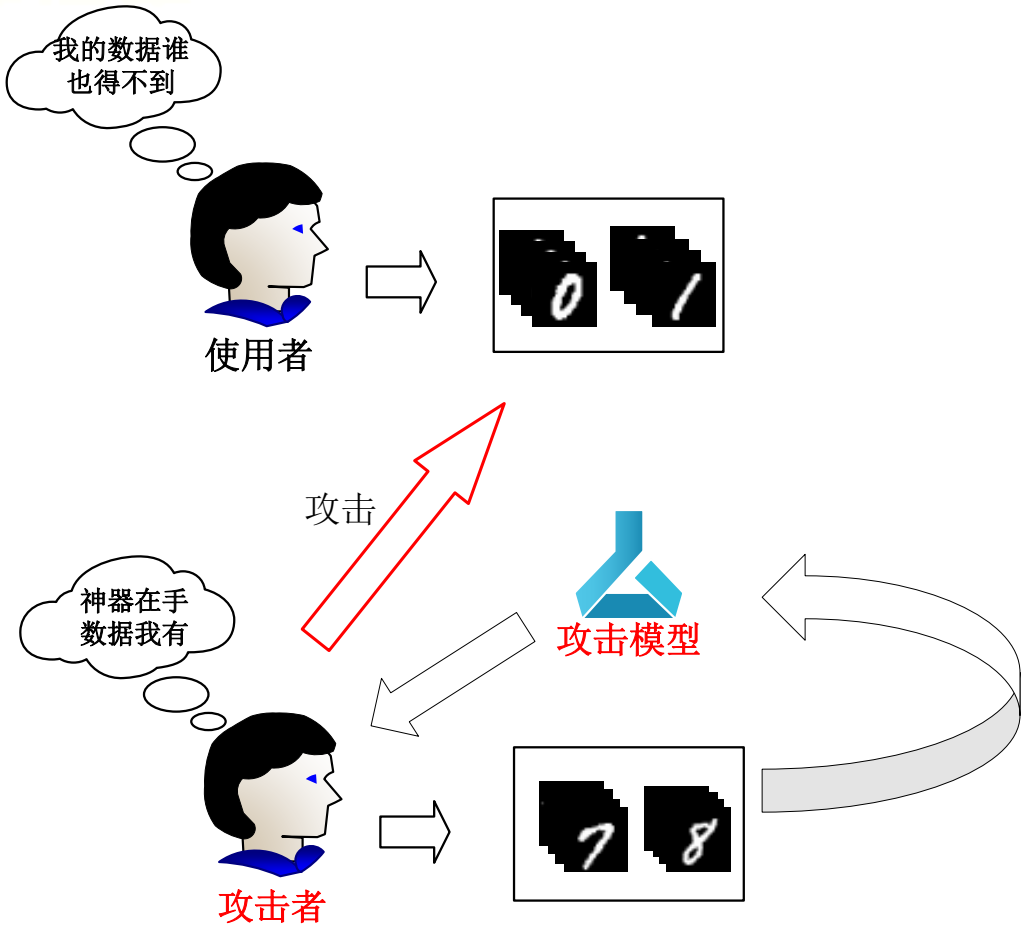
- 白盒模型场景
  - 攻击者能够获得模型的所有信息
  - 包括模型参数、结构、执行结果、中间向量等
- 黑盒模型场景
  - 攻击者不可获知模型的参数、结构等信息
  - 攻击者可访问模型，获得模型的置信度向量、执行结果
- 仅标签模型场景
  - 攻击者不可获知模型的参数、结构、置信度向量等信息
  - 攻击者可访问模型，获得模型最终执行结果

不同的场景代表攻击者所能够获得的信息丰富程度不同



知识基础　辅助数据集&私有训练数据集

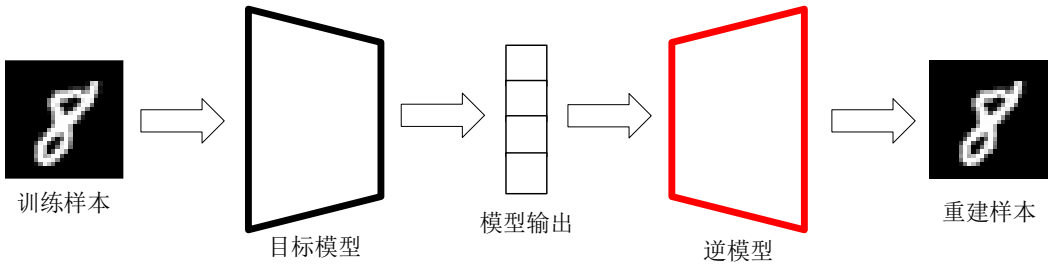
- 私有训练数据集
  - 模型所有者训练模型的训练样本
- 辅助数据集（公共数据集）
  - 攻击者在攻击过程中所采用的数据集
  - 目的：学习图像先验信息；缩小优化范围
  - 特点
    - 公共可获得
    - 与私有训练数据无重叠部分
    - 与私有训练数据拥有相似分布（不一定）



辅助数据是攻击者的帮凶！

研究现状

- 基于训练逆模型的方式
  - 逆模型：由目标模型的输出得到目标模型的输入
  - 可为目标类重建多样样本
  - 重建样本质量差（模型识别&人眼识别）



基于训练逆模型的方式

- 基于生成式模型的方式
  - 基于GAN等生成式模型
  - 重建目标类的代表性样本
  - 重建样本质量好（模型识别&人眼识别）

基于生成式模型的反演攻击方法是研究主流

【CVPR】
The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks

GMI TIPO

T	目标	提出具有高可信度的模型反演攻击方法
I	输入	目标模型*1，辅助数据集*1，先验信息*n
P	处理	1. 公共知识蒸馏：使用辅助数据集训练生成器和鉴别器 2. 敏感信息解密：利用生成器，解决优化问题来恢复图像的敏感区域
O	输出	目标样本*n

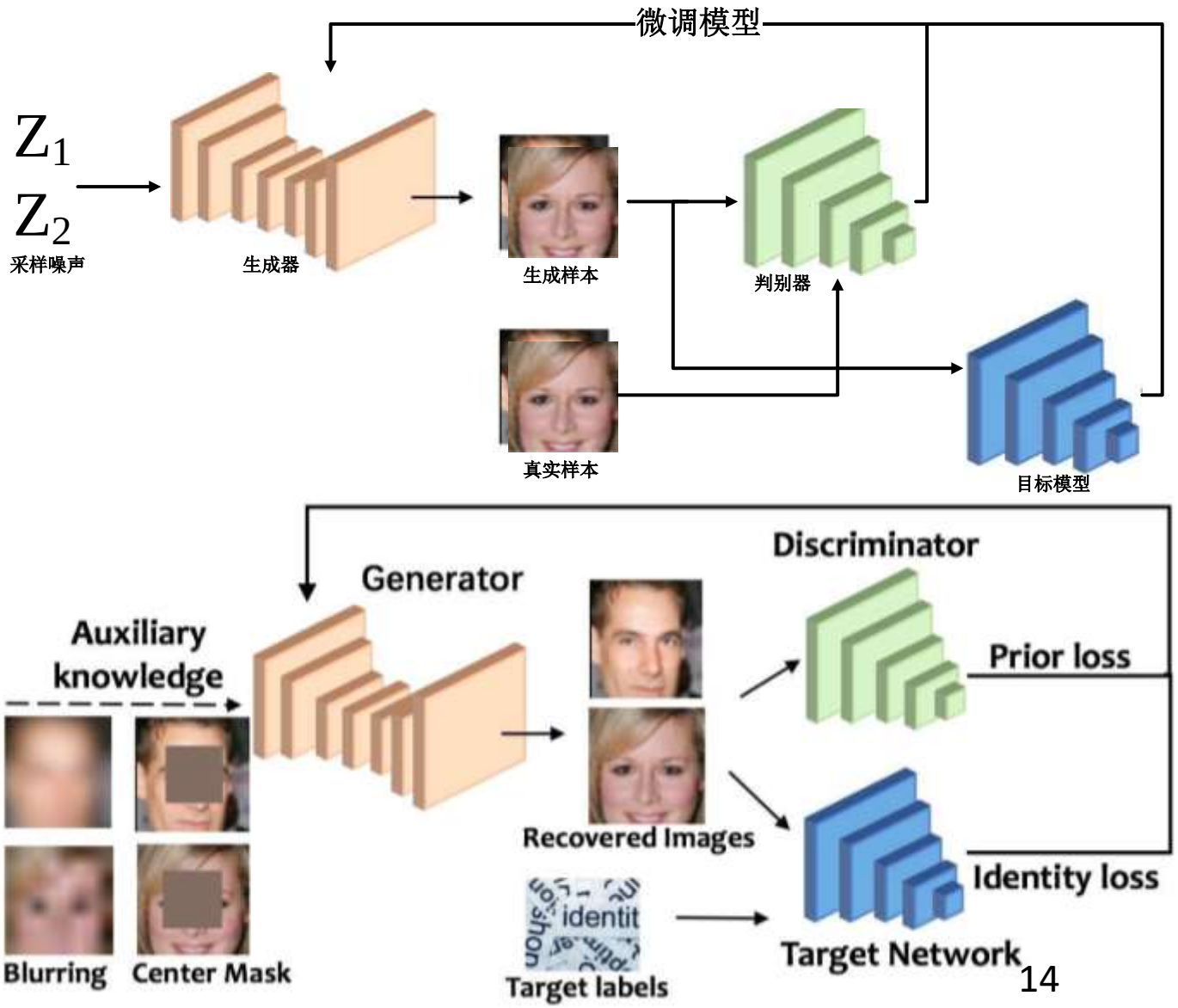
P	问题	现有模型反演攻击算法无法有效针对深度神经网络来重构私有训练数据
C	条件	目标模型白盒设置、拥有部分私有训练数据的先验信息
D	难点	1. 如何避免直接优化实现反演过程中容易陷入局部最优解的问题 2. 如何避免无约束情况下，算法生成不真实样本 3. 如何保证生成的样本在目标网络特征空间中具有多样性
L	水平	CVPR 2020（CCF A）

GMI 算法原理图

- 步骤一：公共知识蒸馏
  - 使用GAN在公共数据集上训练生成器G和判别器D
  - 保证生成图像的真实性和多样性
- 步骤二：敏感信息解密
  - 利用学习得到的生成器来重构私有训练数据
  - 寻找最优潜在向量重建私有训练样本

公共知识蒸馏

敏感信息解密



GMI 公共知识蒸馏

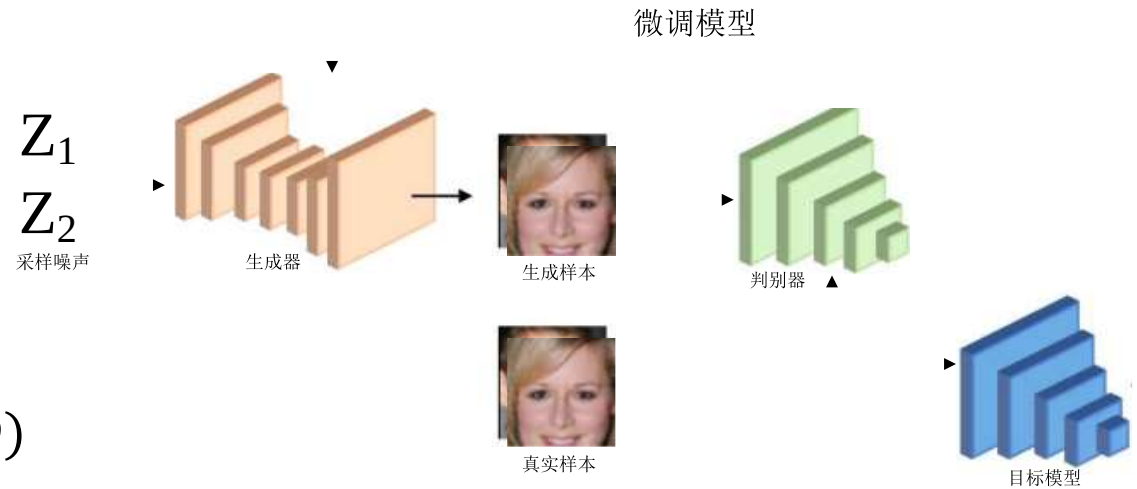
- 公共知识蒸馏
  - 目的：保证生成图像的真实性和多样性
  - 整体流程
    - 收集公开数据集，学习某任务的泛化知识
    - 定义并最小化GAN的损失函数$L_{wgan}(G,D)$

$$min_G max_D L_{wgan}(G,D) = E_x[D(x)] - E_z[D(G(z))]$$

    - 引入样本多样性损失$L_{div}(G)$
    - 得到综合优化目标

$$min_G max_D L_{wgan}(G,D) - \alpha_d L_{div}(G)$$

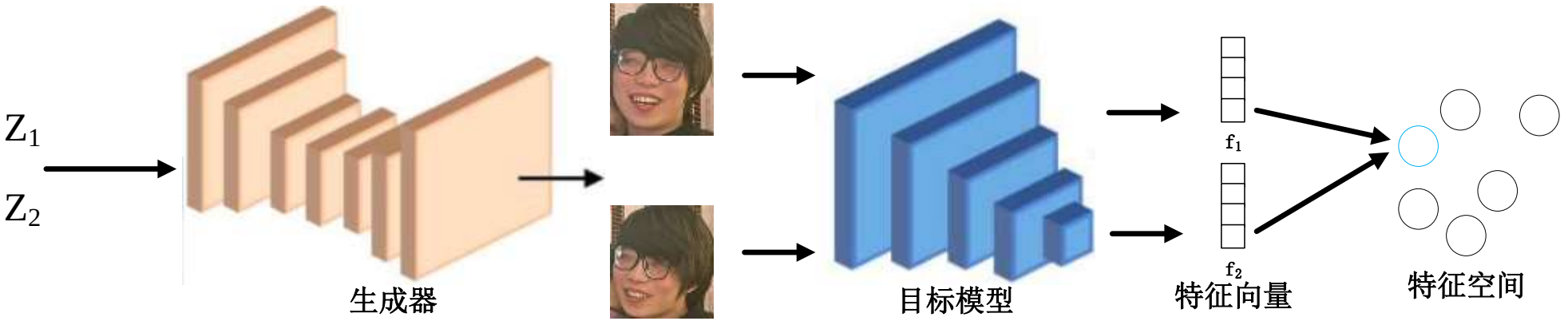
    - 训练GAN得到生成器G
    - 其中，$G(z)$为生成器，$D(x)$为判别器，$F(\cdot)$为目标网络特征输出，z_i为随机噪声，α_d为损失权重



- 公共知识蒸馏
  - 面临问题
    - 如何保证G生成图像的多样性
    - 生成图像特征集中在目标模型特征空间一点
  - 多样性损失$L_{div}(G)$

$$max_G L_{div}(G) = E_{z_1,z_2}[\frac{||F(G(z_1)) - F(G(z_2))||}{||z_1 - z_2||}]$$

  - 其中，$G(z)$为生成器，$F(\cdot)$为目标网络特征输出，z_i为潜在特征向量（随机噪声）



- 敏感信息解密
  - 面临问题
    - 如何解决找到最优生成图像的潜在向量
  - 解决方案
    - 提出目标函数，使得潜在向量生成的图像，在目标网络下具有最大可能性并保证图像的真实合理性
    - 定义先验损失$L_{prior}(z)$和身份损失$L_{id}(z)$

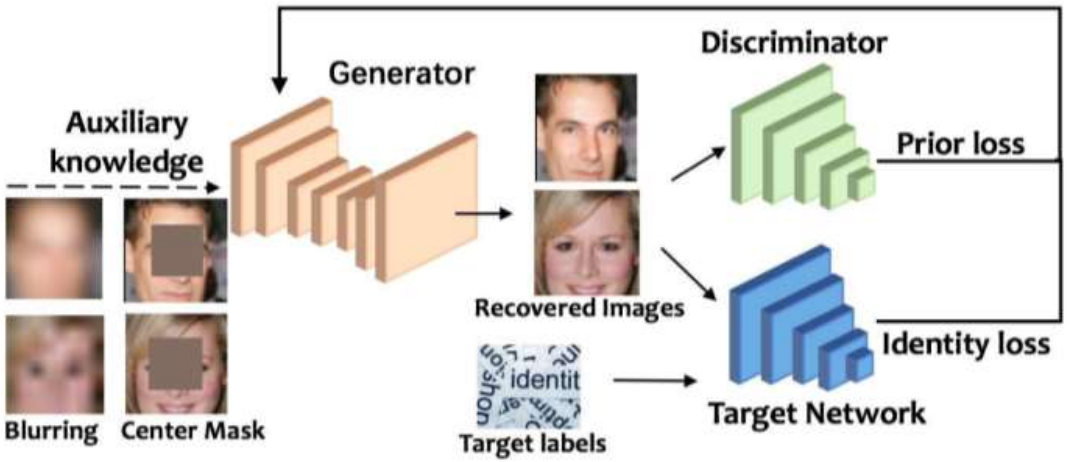
$$L_{prior}(z) = -D\big(G(z)\big)$$

$$L_{id}(z) = -log(C(G(z)))$$

    - 得到寻找最佳潜在向量的目标函数

$$\hat{z} = argmin_z L_{prior}(z) + \lambda_i L_{id}(z)$$

  - 其中，z为潜在向量，λ_i为损失权重，$C(G(z))$为目标网络输出的概率



实验设计　数据资源与评价指标

- 数据集：MNIST、ChestX-ray8、CelebA
- 数据集划分：划分为公共数据集&私有训练数据集，二者无重复类
- 目标模型

数据集	MNIST	ChestX-ray8	CelebA	
模型	3*CNN+2*Pool	ResNet18	VGG16	ResNet-152

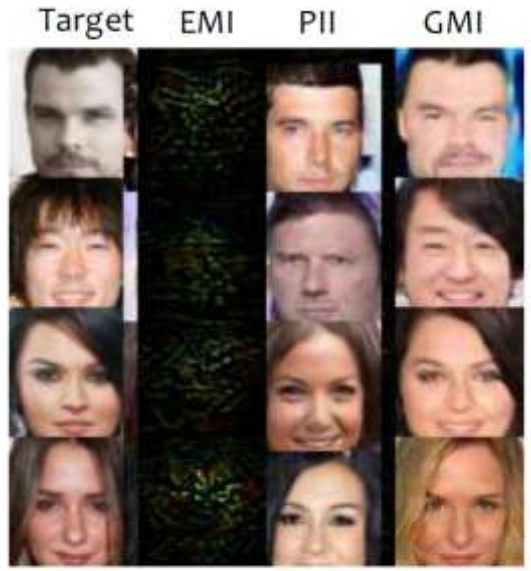
- 评价指标
  - PSNR：衡量两幅图像之间的像素相似度，值越高代表重建质量越好
  - Attack Acc：评估模型使用重建数据集的分类精度，精度越高代表更好地重建了私人训练数据的信息
  - Feat Dist：衡量重建图像与目标类质心之间的L_2距离
  - KNN Dist：衡量重建图像到目标类的最短L_2距离

实验设计　对比实验

- 对比方法：EMI、PII（纯图像修复）
- 对比实验（无先验信息）

先验信息：私有训练数据的部分信息，例如模糊或受损的图像

Model	Attack	KNN Dist	Feat Dist	Attack Acc	Top-5 Attack Acc
VGG16	EMI	2397.50	2255.54	0	0
	PII	2368.77	2425.09	0	0
	GMI	**2098.92**	**2012.10**	**28**	**53**
ResNet-152	EMI	2422.99	2288.13	0	1
	PII	2368.77	2425.09	0	0
	GMI	**1969.09**	**1886.44**	**44**	**72**
face.evolve	EMI	2371.52	2248.81	0	1
	PII	2368.77	2425.09	0	0
	GMI	**1923.72**	**1802.62**	**46**	**76**



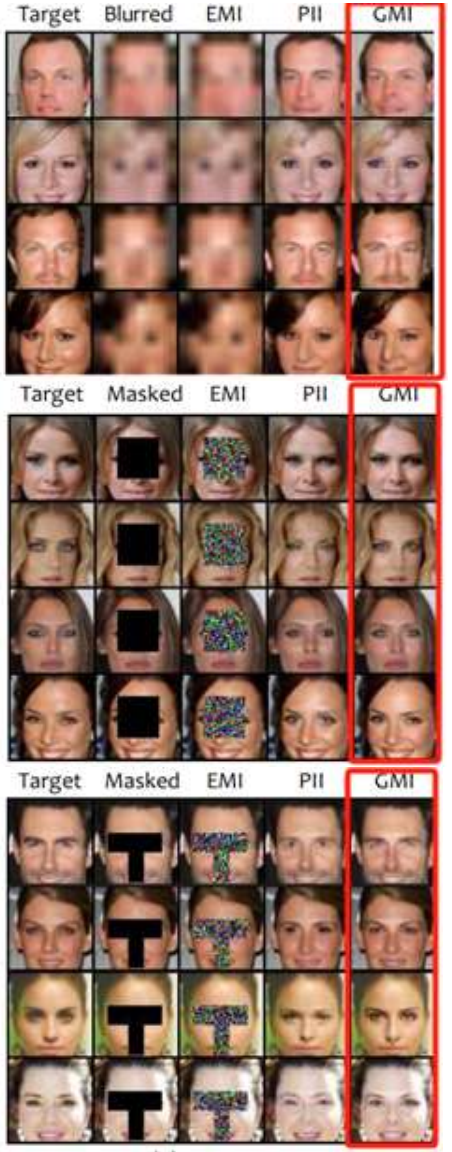
(a) W/out auxiliary knowledge

- 实验结论
  - GMI无论是从重建图像的质量还是重建图像所包含的私有信息，均优于对比算法

实验设计　对比实验

- 对比实验（有先验信息）

Model	Metric	Blurring			Center Mask			Face T mask		
		EMI	PII	GMI	EMI	PII	GMI	EMI	PII	GMI
VGG16	PSNR	19.66	20.78	**21.97**	18.69	25.49	**27.58**	19.77	24.05	**26.79**
	Feat Dist	2073.56	2042.99	**1904.56**	1651.72	1866.07	**1379.26**	1798.85	1838.31	**1655.35**
	KNN Dist	2164.40	2109.82	**1946.97**	1871.21	1772.74	**1414.37**	1980.68	1916.67	**1742.74**
	Attack Acc	0%	6%	**43%**	14%	34%	**78%**	11%	20%	**58%**
ResNet-152	PSNR	19.63	20.78	**22.00**	18.69	25.49	**27.34**	19.89	24.05	**26.64**
	Feat Dist	2006.46	2042.99	**1899.79**	1635.03	1866.07	**1375.36**	1641.31	1838.31	**1594.81**
	KNN Dist	2101.13	2109.82	**1922.14**	1859.78	1772.74	**1403.24**	1847.74	1916.67	**1670.05**
	Attack Acc	1%	6%	**50%**	9%	34%	**80%**	11%	20%	**63%**
face.evoLVe	PSNR	19.64	20.78	**22.04**	18.97	25.49	**27.69**	19.86	24.05	**25.77**
	Feat Dist	1997.93	2042.99	**1878.38**	1609.35	1866.07	**1364.42**	1762.57	1838.31	**1624.95**
	KNN Dist	2085.53	2109.82	**1904.47**	1824.10	1772.74	**1403.19**	1962.07	1916.67	**1682.56**
	Attack Acc	1%	6%	**51%**	12%	34%	**82%**	11%	20%	**64%**



- 实验结论
  - PII重建图像与GMI在视觉上相似
  - GMI重建的图像重建了私有训练数据的特征信息

实验设计　公共数据集实验

- 公共数据集占比实验
  - 公共数据集占比下降，GMI攻击性能下降
  - 下降幅度最大不超过7%

Model	CelebA→CelebA			
	1:1	1:4	1:6	1:10
VGG	78%	77%	75%	72%
LeNet	81%	75%	77%	75%
face.evoLVe	77%	77%	77%	70%

公共数据集占比实验

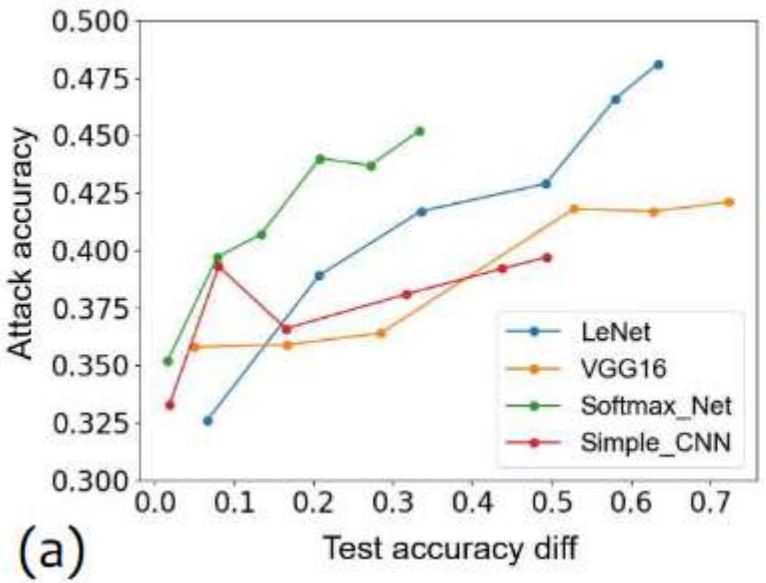
- 数据集分布不相似实验
  - 公共数据集与私有数据集分布差异较大，GMI攻击性能下降约20%
  - 大幅度优于现有的MI攻击算法

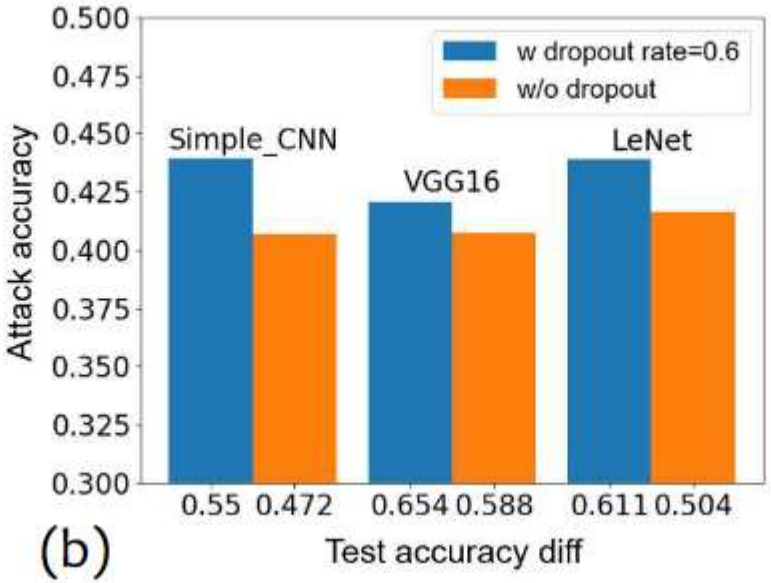
Model	PubFig83→CelebA		EMI
	W/o Preproc.	W/ Preproc.	
VGG	48%	67%	14%
LeNet	52%	66%	9%
face.evoLVe	56%	70%	12%

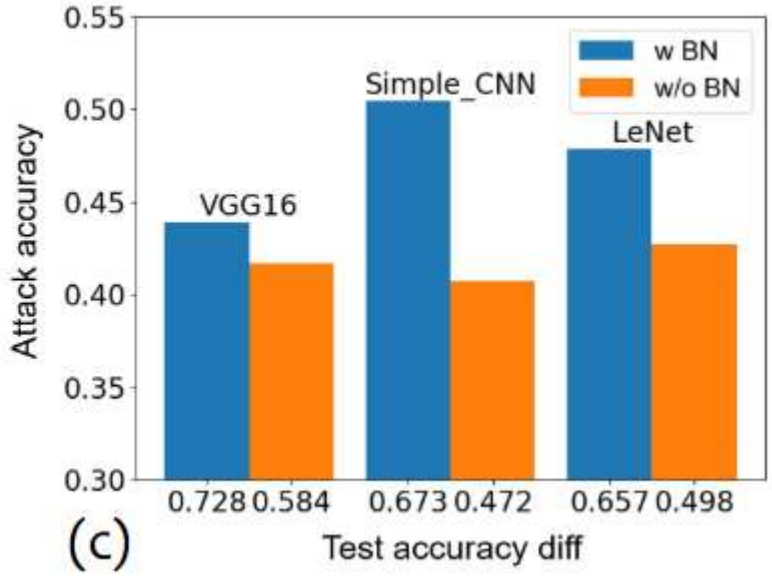
数据集分布不相似实验

实验设计　目标模型性能实验

- 目标模型性能与MI攻击脆弱性关系实验







- 实验结论
  - 目标模型本身的预测能力越高，MI攻击效果越佳
  - 验证了预测能力越好的模型，越能将私有训练数据特征保存至模型中，进而更容易遭受MI攻击

算法总结 GMI

- 算法流程
  - 公共知识蒸馏：使用辅助数据集训练生成器和鉴别器
  - 敏感信息解密：利用生成器，解决优化问题来恢复图像的敏感区域
- 算法优势
  - 能够有效重建深度神经网络的私有训练数据
  - 验证了目标模型性能和模型反演攻击成功率之间的正比关系
- 算法不足
  - 对先验信息高度依赖
  - 白盒设置，在实际场景中应用面窄

【TDSC】

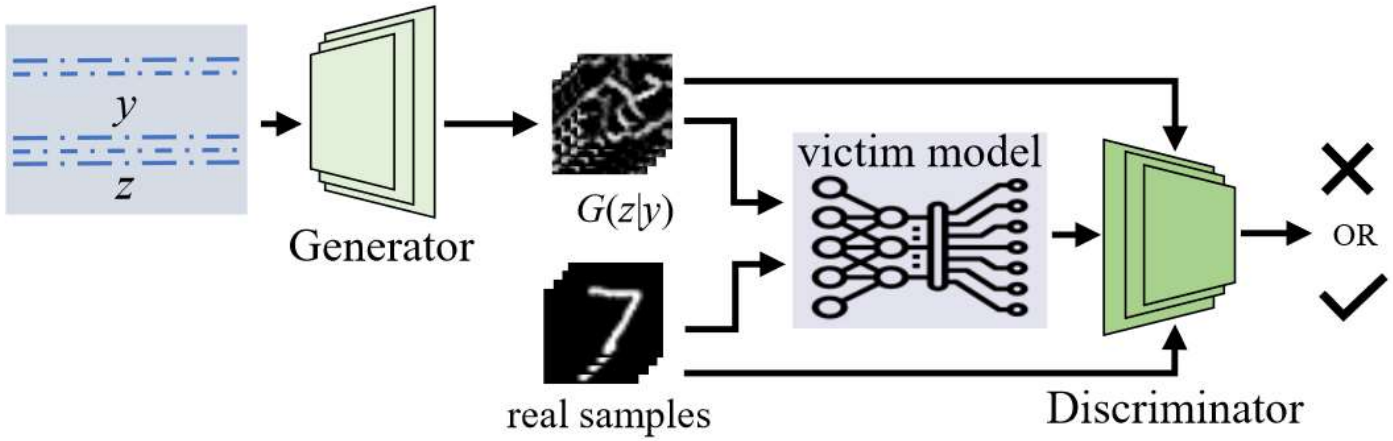
The Role of Class Information in Model Inversion Attacks against Image Deep Learning Classifiers

T	目标	提出减少对先验知识依赖的模型反演攻击方法
I	输入	目标模型*1，辅助数据集*1
P	处理	1. CGAN模型训练：以辅助数据集和目标模型训练CGAN模型 2. 反演优化：寻找最佳条件输入来生成目标样本
O	输出	目标样本*n

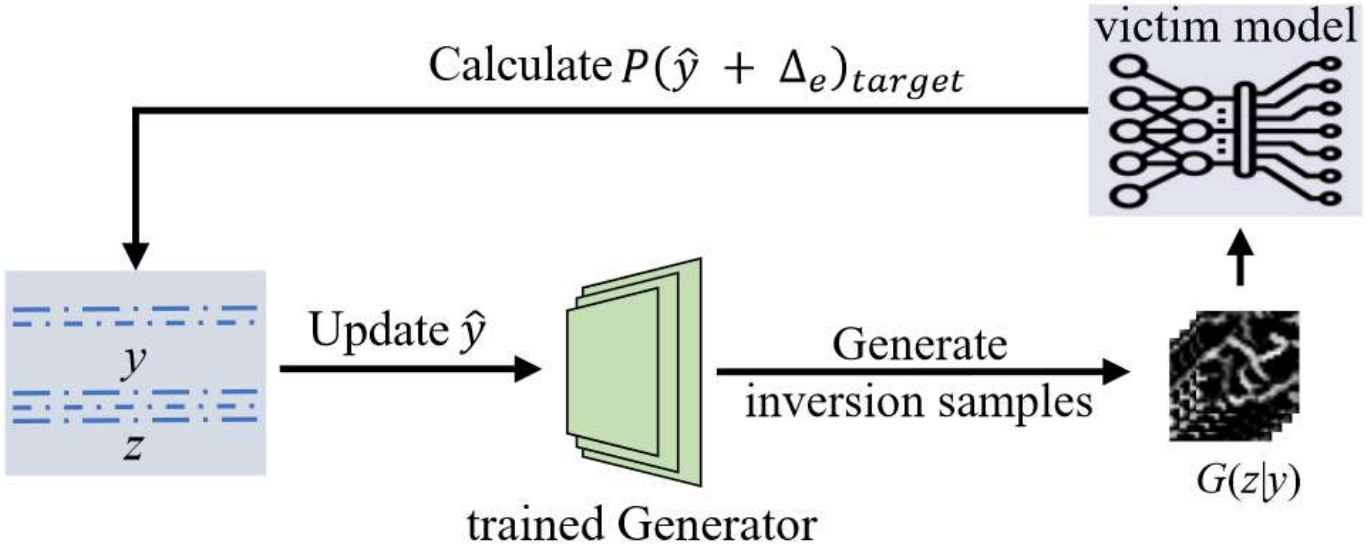
P	问题	实际应用场景往往无法提供模型的白盒信息 现有的模型反演攻击方法高度依赖先验知识
C	条件	黑盒设置下，需获得目标模型置信度向量
D	难点	1. 如何对黑盒模型进行测试 2. 如何减少对先验知识的依赖
L	水平	TDSC 2023（CCF A）

SMI 算法原理图

- 算法原理图
  - 训练阶段
    - 输入：随机噪声z，辅助样本类别信息y
    - 输出：训练完毕、符合目标任务的生成器G
  - 反演阶段
    - 输入：随机噪声z，目标类别信息y
    - 输出：重建的目标样本

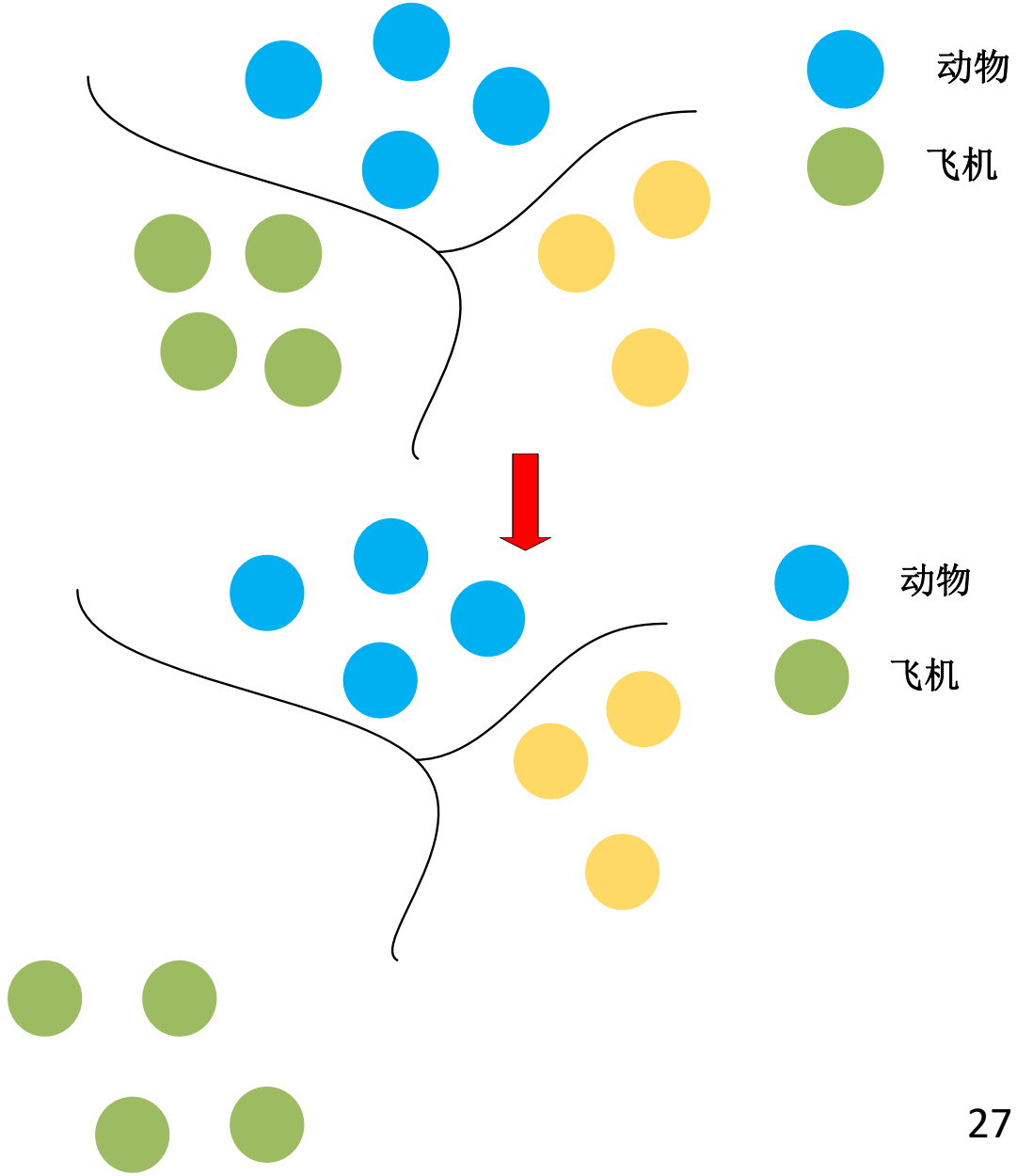


(a) Training Stage



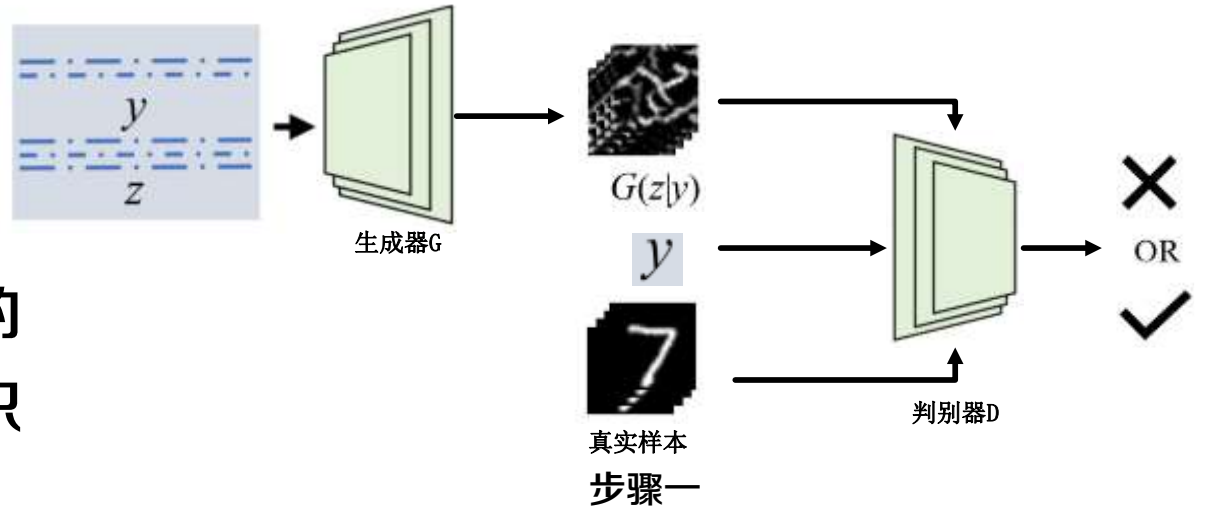
(b) Inversion Stage

- 重建样本
  - 重建样本的特征可视作已有特征的组合
  - 例如：鸟 = 飞机 + 动物的羽毛
- 不同特征在特征空间的距离
  - 潜在向量难以囊括特征空间中距离较远的特征
- 舍入输出
  - 对置信度向量元素作小数位保留
  - (0.2, 0.1, 0.7)进行小数精度为0的舍入时，结果为(0, 0, 1)

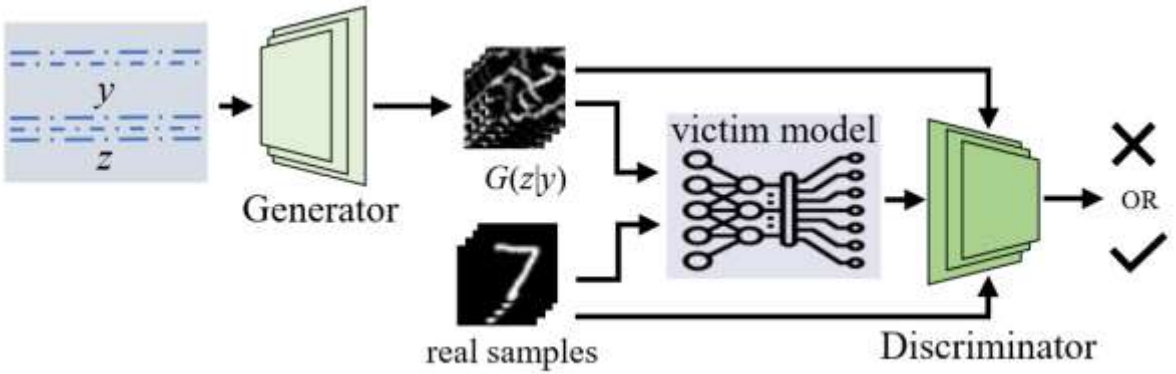


SMI 训练阶段

- 问题
  - 缺少目标的先验信息
- 解决方案
  - 引入类信息来指导反演模型从其他类的样本中推断出对攻击有指导意义的知识
- 步骤
  - 使用辅助样本及其真实标签训练CGAN
  - 以目标模型的舍入输出代替辅助样本&生成样本真实标签，微调CGAN模型



步骤一

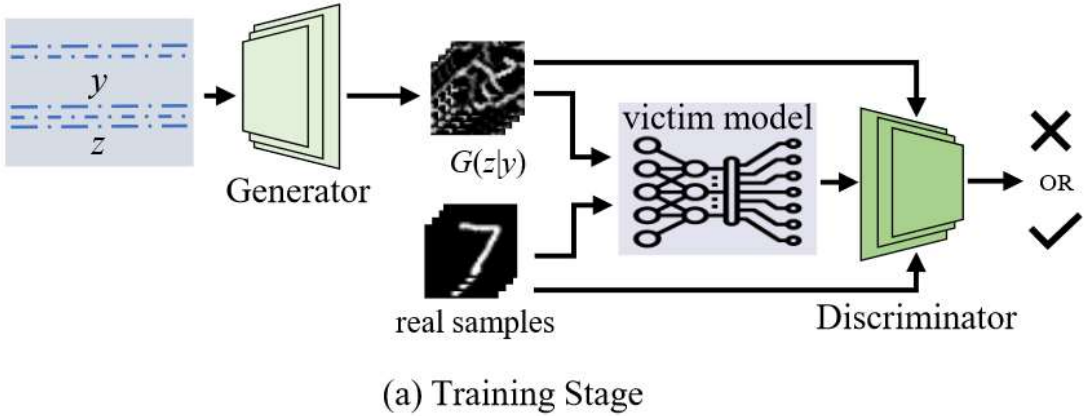


步骤二

舍入输出将目标模型中的知识转移到CGAN

SMI 训练阶段

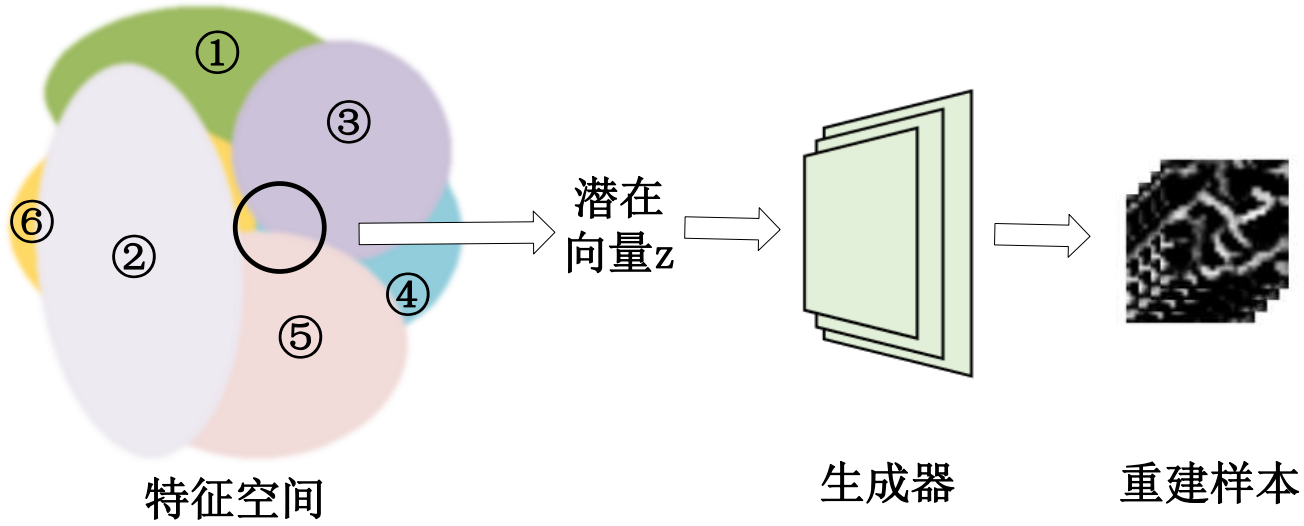
- 算法原理
  - 训练阶段的优化目标
    - $min_G max_D V(D,G) = E_x[logD(x|F_v(x))] + E_z\left[\log(1 - D\left(G(z|y)\right))\right]$
  - 参数说明
    - x：输入样本
    - $F_v(x)$：受害者模型的舍入输出
    - $G(z|y)$：生成器生成结果
  - 优化目标
    - 最小化生成器G在对抗生成样本上的损失
    - 最大化判别器D在真实样本上的准确度
  - 算法思想：通过利用目标模型对生成样本和真实样本的$F_v(x)$，将目标模型中的知识转移到CGAN中



(a) Training Stage

SMI 反演阶段

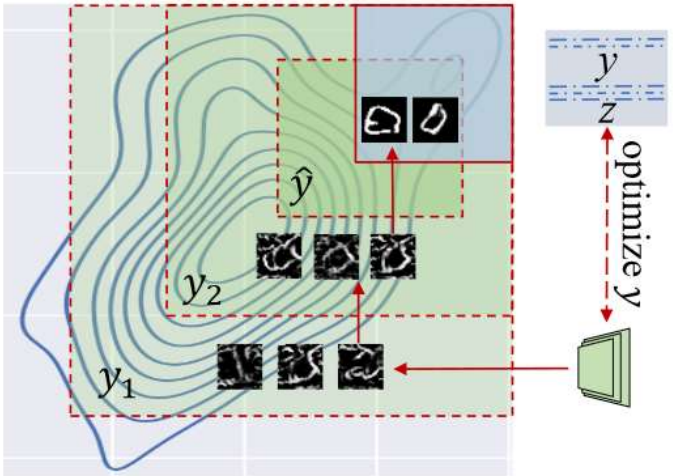
- 已有的MI算法（GMI）
  - 搜索潜在空间并优化，找到一个潜在向量z;
  - 由z和生成器得到重建样本
- 存在问题
  - 距离较远的特征难以同时囊括
  - 梯度下降的优化算法不适用于黑盒模型
- 解决方案
  - 采用贪婪搜索策略寻找最佳的条件输入$\hat{y}$



- 例如：$\hat{y} = (0,0.1,0.3,0.2,0.4,0)$
  ① ⑤
- 不同类别的置信度分数组合为条件输入
- 可以囊括距离较远的特征

SMI 反演阶段

- 算法原理
  - 反演阶段
    - 面临问题：由于缺少先验知识以及处于黑盒设置下，基于梯度优化的搜索方式不再实用
    - 解决方案：采用贪婪搜索策略寻找最佳的条件输入$\hat{y}$
  - 目标函数



b) inversion optimization strategy of SMI

$$p_{target}\big(G(z|y)\big) = \frac{|F_v(\{G(z_0|y), \dots G(z_\alpha|y)\}) \cap T_k|}{\alpha}$$

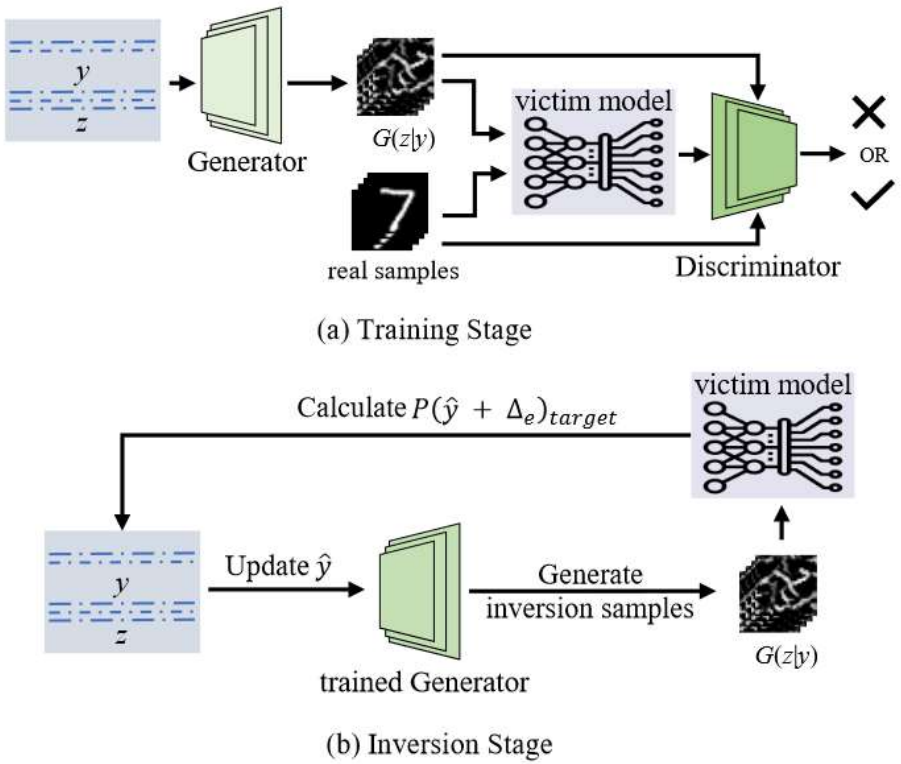
$$y = argmax_y p_{target}\big(G(z|y)\big)$$

- 其中，T_k为目标样本集，α为每次迭代中生成的预定义样本数，z_i为随机噪声输入

  - 含义：算法试图找到最优的y，生成器$G(z|y)$在y条件下，能够最大化生成输入目标类别的样本

SMI 反演阶段

- 算法原理



Algorithm 1 Supervised Model Inversion

Input: auxiliary dataset Au, victim model $F_v(\mathbf{x})$
Output: inversion sample $\hat{\mathbf{x}}$

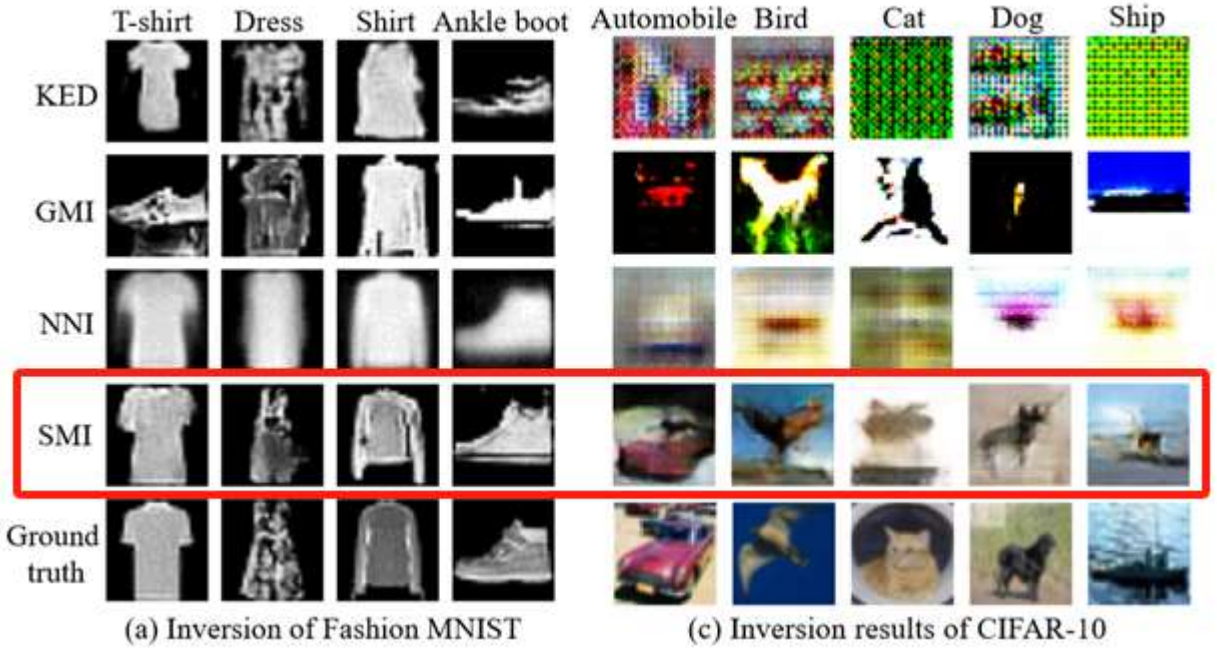
1: *# Training stage:*
2: Train the cGAN model with Au
3: Train SMI model with Eq.1
4: *# Inversion stage:*
5: Randomly initialization $\hat{\mathbf{y}}$
6: **for** $e = 0$ to η **do**
7: $\quad \Delta_e = (\delta_0, ..., \delta_m)$
8: $\quad$ Calculate $P(\hat{\mathbf{y}} + \Delta_e)_{target}$
9: $\quad$ Update $\hat{\mathbf{y}}$
10: **end for**
11: $\hat{\mathbf{x}} = G(\mathbf{z} \mid \hat{\mathbf{y}})$

实验设计　数据资源

- 数据集
  - MNIST、Fashion-MNIST、CIFAR-10、CelebA
  - 数据集划分：受害者模型训练数据集与辅助数据集不重叠
- 受害者模型
  - CNN模型、VGG模型、ResNet-50模型
- 对比方法
  - NNI（CCS 2019 ）、GMI（CVPR 2020）、KED（ICCV 2021）
- 评价指标
  - SSIM-R and SSIM-C：结构相似度
  - KNN Dist：反演样本到目标类的最短距离
  - ACC：反演样本被评估模型成功分类的比例

实验设计　对比实验

- 对比实验



(a) Inversion of Fashion MNIST　　(c) Inversion results of CIFAR-10

Metric	Fashion-MNIST				
	Benchmark	SMI	GMI	NNI	KED
SSIM-r	0.383	0.330	0.263	0.215	0.277
SSIM-c	0.407	0.352	0.224	0.380	0.266
KNN Dist	953.84	1396.45	2131.67	1108.45	1447.52
Acc	0.9082	0.8511	0.7695	0.7013	0.7745

Metric	MNIST				
	Benchmark	SMI	GMI	NNI	KED
SSIM-r	0.554	0.555	0.506	0.394	0.508
SSIM-c	0.489	0.494	0.416	0.499	0.374
KNN Dist	1195.07	1392.68	1767.75	1156.51	1353.44
Acc	0.9900	0.9205	0.8834	0.9169	0.7497

Metric	CIFAR-10				
	Benchmark	SMI	GMI	NNI	KED
SSIM-r	0.087	0.088	0.038	0.072	0.004
SSIM-c	0.120	0.122	0.047	0.110	0.005
KNN Dist	2344.00	2381.16	4305.07	2693.67	5229.83
Acc	0.8332	0.4024	0.2723	0.1372	0.0586

Metric	CelebA				
	Benchmark	SMI	GMI	NNI	KED
SSIM-r	0.4928	0.3308	0.2141	0.3427	0.3201
SSIM-c	-	-	-	-	-
KNN Dist	3409.61	3716.96	7040.13	4493.84	5678.89
Acc	0.7186	0.4250	0.1391	0.0023	0.2148

- 实验结论：
  - SMI可以重建所有类
  - SMI重建样本在视觉上更具有可信度
  - SMI重建样本在特征上更具有可信度

实验设计　条件y实验

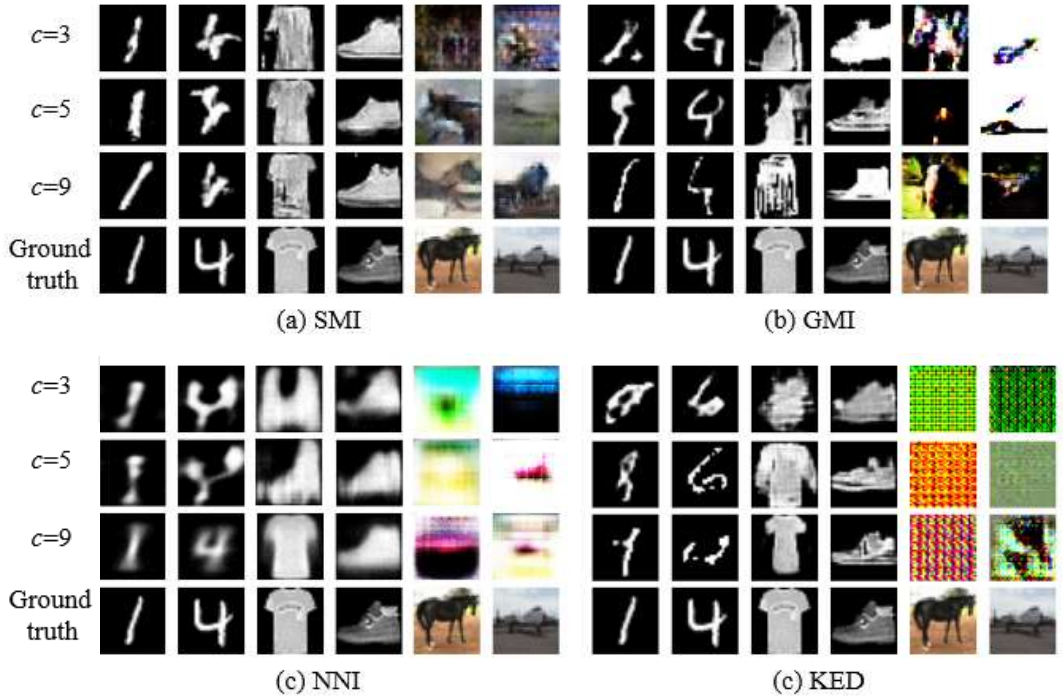
- 条件输入y的影响

Target Class	y	$P(\cdot)_{target}$
5	(0,0,0,0,0,1,0,0,0,0)	42%
5	(0,0,0.2,0,0,0.3,0.2,0,0.1,0)	1%
8	(0,0,0,0,0,0,0,0,1,0)	1%
8	(0,0,0,0.1,0,0,0.1,0,0.7,0.2)	2%

- 实验现象与结论（1）
  - 可以通过优化y来提高反演攻击的成功率
  - $P(\cdot)_{target}$随y的变化而出现大幅度波动
- 实验现象与结论（2）
  - 现象：y中合理的取值组合相较于目标类的最大值会产生更好的效果。
  - 原因：训练阶段并没有学习到目标类的完整特征；反演阶段可以通过添加其他类的知识来弥补缺失的知识

实验设计　辅助数据集实验

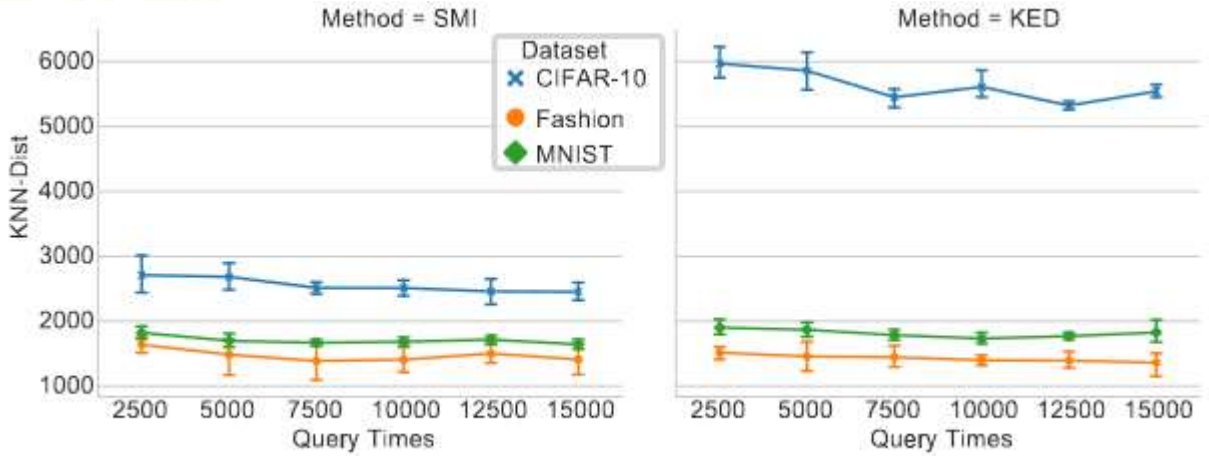
- 更少的辅助数据集反演单个类别样本
  - 设定辅助数据集中类别的数量差异
- 实验结论
  - 反演质量随辅助数据集的类减少而下降
  - SMI反演质量最佳

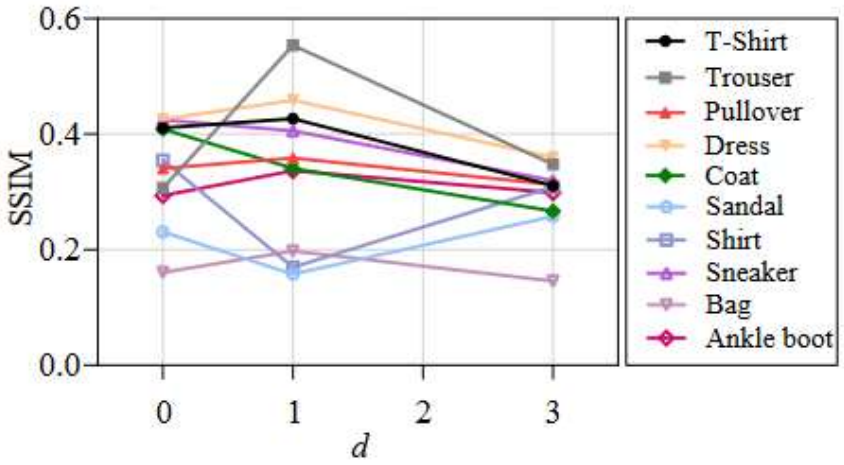


Class	1	4	T-shirt	Ankle boot	Horse
Method			SMI		
c=3	743.53	1204.11	1477.48	880.50	2119.72
c=5	865.64	1817.78	1081.36	1049.83	2667.20
c=9	521.59	1238.37	1317.72	949.07	2547.96
Method			GMI		
c=3	1287.09	2128.35	2141.61	1872.21	4624.81
c=5	1567.78	1407.31	2235.26	1326.99	2161.41
c=9	757.51	1206.41	2576.34	1462.21	1590.27
Method			NNI		
c=3	968.92	1919.51	2029.89	1388.08	4136.68
c=5	1065.43	2332.67	2573.18	1039.12	3233.22
c=9	821.99	1230.22	977.60	986.73	3809.64
Method			KED		
c=3	1707.94	1692.36	1593.57	1478.23	5608.19
c=5	1421.92	1552.25	1481.11	1299.95	5766.65
c=9	1148.90	1367.20	1125.40	1272.46	5597.21

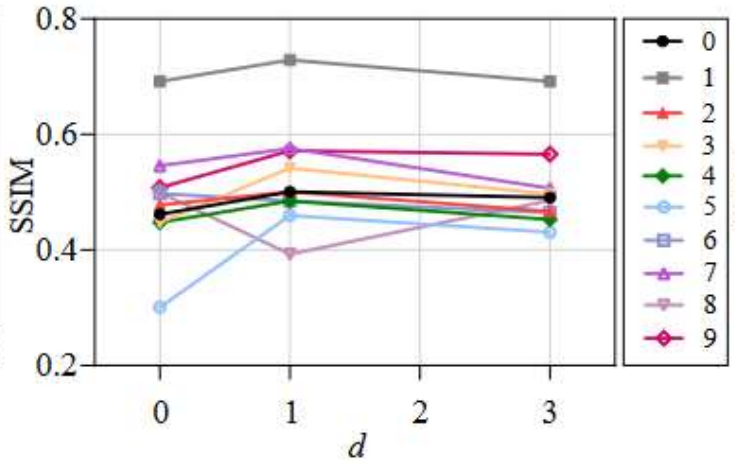
实验设计　查询时间&舍入输出实验

- 查询时间实验
  - SMI查询次数：大于等于7500次
  - KED查询次数：高于SMI
  - SMI拥有更好的性能
- 舍入输出对结果的影响

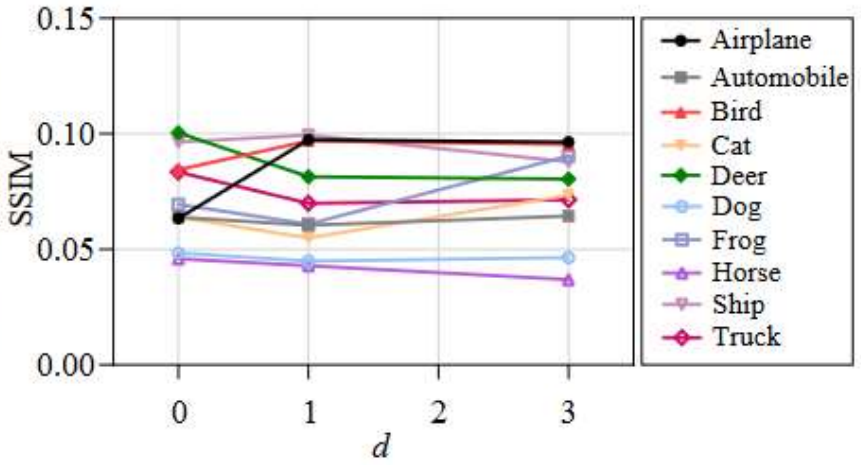




(a) Fashion-MNIST



(b) MNIST



(c) CIFAR-10

算法总结 SMI

- 算法流程
  - 利用辅助数据集训练CGAN模型
  - 通过贪婪搜索策略，在反演阶段优化反演结果
- 算法优势
  - 黑盒算法，更符合实际场景
  - 充分利用类别信息，减弱了已有算法对先验知识的高度依赖
  - 利用类别信息，规避了潜在向量难以囊括相距过远的特征的问题
- 算法不足
  - 查询次数仍然较大，影响算法性能
  - 仍旧依赖辅助数据集，在辅助数据集信息较少时，性能下降

特点总结与未来展望

特点总结与未来展望

- **GMI**
  - 提出能应用于深度学习模型的MI攻击方法
  - 使用GAN生成式模型，由辅助数据集学习先验知识，通过搜索优化潜在向量生成样本
  - 白盒模型，不符合实际应用场景
- **SMI**
  - 考虑黑盒实际测试场景，且效果优于现有的白盒测试方法
  - 类别输入信息容易囊括复杂图像的冗余特征，导致重建样本质量差
- 未来发展
  - 除却黑盒模型外，仅标签模型更符合实际应用场景
  - 更高效的搜索优化方式
  - 针对模型反演攻击问题发展有效的防御手段

预期收获　回顾分析

- 预期收获
  - 掌握模型反演攻击的相关知识
    - 辅助数据集与私有数据集
    - 深度学习模型面临训练数据泄露风险
  - 了解白盒模型反演攻击方法
    - 基于GAN模型生成高质量样本
    - 搜索优化潜在向量，重建私有训练样本
  - 了解黑盒模型反演攻击方法
    - 利用贪婪搜索策略，解决黑盒模型的搜索优化问题
    - 利用类别信息，减弱算法对先验信息的依赖程度

学会了吗?
学会了也不能攻击别人

参考文献

[1] Zhang Y, Jia R, Pei H, et al. The secret revealer: Generative model-inversion attacks against deep neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition [C]. Seattle, WA, USA:IEEE ,2020: 253-261.

[2] Tian Z, Cui L, Zhang C, et al. The Role of Class Information in Model Inversion Attacks against Image Deep Learning Classifiers[J]. IEEE Transactions on Dependable and Secure Computing, 2023.(1-14)

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！