Re-thinking Model Inversion Attacks Against Deep Neural Networks

Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, Ngai-Man Cheung
Singapore University of Technology and Design (SUTD)

摘要

深度神经网络（Deep Neural Networks，DNNs）已被广泛应用于涉及私人和敏感数据集的诸多领域，如人脸识别、语音识别、医疗保健等。人们也越发担心为获取训练 DNNs 时使用的机密数据集的相关知识而进行的隐私攻击，特别是**模型反演攻击**（Model Inversion Attack，MIA）。本工作通过分析目前最先进的白盒 MI 方法（GMI 和 KEDMI）中存在的影响反演攻击精度的问题（次优的**身份损失**设计和 **MI 过拟合**），并提出了相应的改进方案（更优的身份损失设计和模型增强），最后通过充分地实验验证了上述改进方案的有效性。在本次复现过程中，我发现原论文中基于**知识蒸馏**所得到的增强模型的预测精度较低，故通过改进知识蒸馏过程所使用的损失函数，提高了增强模型 4% - 12% 的预测精度，进而提高了 2% - 5% 的反演攻击精度。

关键词：模型反演攻击；身份损失；MI 过拟合；知识蒸馏

1 引言

在当今数字化时代，机器学习和深度学习技术的迅猛发展为各行各业注入了蓬勃的创新力量，推动着各行各业的创新发展。然而，与之相应的是机器学习模型的安全性面临的挑战也逐渐显现，特别是模型反演攻击（Model Inversion Attack，MIA）。MIA 作为一种针对机器学习系统的隐私威胁，引起了社会和科学界的广泛关注。这一背景下，对于 MIA 的研究显得愈发重要。MIA 不仅仅是对机器学习模型中训练数据的反演，更是对用户隐私的直接侵犯。随着大数据时代的到来，私人和敏感数据不断涌入机器学习模型，攻击者通过对模型输出结果的分析，有可能重建出原始训练数据，从而导致用户隐私的泄露。这种信息泄露的潜在风险使 MIA 成为当前机器学习领域亟待解决的问题之一。

模型反演攻击是一种高度复杂而隐蔽的威胁，攻击者可以通过多种手段推断模型中的原始训练数据。因此，有必要深入研究攻击者可能采取的方法，以及这些攻击手段对模型安全性的实质影响。研究 MIA 不仅仅有助于揭示机器学习模型中的潜在漏洞，更为重要的是为制定更为有效的防御策略提供充分的理论基础。对攻击原理的深刻认识有助于我们更好地抵御潜在的威胁，提高模型的整体鲁棒性。同时，这一研究还可以为设计更为安全可靠的机器学习模型提供有效且正确的指导，从而推动安全机器学习的发展。

深入研究模型反演攻击的意义不仅限于提升模型安全性，还涉及到隐私保护、安全机器学习的整体推动等多个方面。首先，对于隐私保护而言，MIA 是一个直接而严峻的挑战。研究如何防范此类攻击不仅仅是技术问题，更是社会伦理和法规制度的考量。通过探索 MIA 的本质，我们可以更好地指导制定相关政策，确保用户隐私在机器学习应用中得到充分的保护。其次，深入研究 MIA 也为推动安全机器学习的发展提供了契机。安全机器学习是一个多学科交叉的领域，它要求我们不仅关注模型

性能的提升，更要注重模型在面对不同威胁时的鲁棒性。通过不断深化对 MIA 的认知，我们能够更好地推动整个领域的研究和创新。

综上所述，模型反演攻击的研究不仅是对机器学习系统安全性的挑战，更是对隐私保护和安全机器学习发展的有益贡献。深刻理解攻击原理，制定有效防御策略，将有助于构建更为安全、可信赖的机器学习系统。

2 相关工作

本节主要对本次复现内容的相关工作进行简要的分类概括与描述。首先简要介绍何为模型反演攻击，以及模型反演攻击根据攻击者掌握知识的多少的不同分类，其次详细阐述模型反演攻击当下的两种主流执行范式及不同执行范式对应的代表性工作。

2.1 MIA

模型反演攻击属于隐私攻击中的一种，其目的是通过利用训练数据与模型输出之间的相关性来重建训练数据中的敏感特征或代表性特征。现如今深度神经网络（Deep Neural Networks，DNNs）已被广泛应用于涉及私人和敏感数据集的诸多领域，如人脸识别、语音识别、医疗保健等。人们越发担心为获取训练 DNNs 时使用的机密数据集的知识而进行的隐私攻击，特别是模型反演攻击。为进一步了解 MIA，我通过描述 MIA 中的一个典型应用场景——人脸识别，当攻击者使用模型反演攻击的具体方法攻击某一人脸识别系统来重建系统内部嵌入的人脸图像，若攻击成功，攻击者可重建自己感兴趣的身份所对应的人脸图像，进而使人脸识别系统失效，其产生的危害将是不可想象的。

现如今 MIA 的研究进展可根据攻击者掌握的知识（详见表 1）分为白盒 MIA、黑盒 MIA 和仅标签 MIA 三种不同设置场景，它们均对目标模型的训练数据不了解。白盒 MIA 假设攻击者完全了解目标模型的架构（Architecture）及其内部的参数（Parameters）。黑盒 MIA 假设攻击者具有目标模型的访问权限并可接收目标模型输出的软标签（Soft-labels），即置信度向量或置信度分数。仅标签 MIA 则为特殊的黑盒 MIA，假设攻击者仅可获取数据输入目标模型后输出的硬标签（Hard-labels），即仅可获取输入数据的预测标签，而无其相关的置信度分数。仅标签 MIA 的设置更为严苛，实现难度更大，但也更符合实际的应用场景。

表 1: 不同设置场景下攻击者所具备的知识

Setting	Architecture / Parameters	Soft-labels	Hard-labels
White-box	√	√	√
Black-box	×	√	√
Label-only	×	×	√

2.2 MIA 的执行范式

根据现有模型反演攻击的不同实现方式,可分为两种主流的执行范式,分别为基于优化（Optimization-based）的 MIA 和基于学习（Learning-based）的 MIA。不同执行范式的实现方式差别很大，下文对这两种主流的执行范式进行详细阐述。

2.2.1 Optimization-based

模型反演攻击常被看作为一个优化问题，即在一个潜在空间中搜索在目标模型下实现最大似然的敏感特征值。若直接在高维空间（如图像空间）上使用梯度或无梯度优化算法求解该优化问题，虽在面对简单模型（线性回归、决策树等）时尚可取得较好的反演效果，但当面对 DNNs 时，搜索易陷入局部最小值。故现有的许多研究工作使用生成对抗网络（Generating Adversarial Network，GAN）通过学习公共数据等先验知识来缩小搜索空间，进而在维度较低的潜在空间上进行优化，可以很好的解决非凸问题，如第一篇将 GAN 应用于 MIA 领域的 GMI[1]和改进 GMI 中 GAN 的训练方式和将单点恢复拓展至分布恢复来降低反演时间的 KEDMI[2]。此外，训练即插即用的 cGAN 的 PLG-MI[3]和本次复现工作选择的 Re-thinking-MI[4]则对 GMI 和 KEDMI 进行不同方面的改进，使 MIA 方法更适用于私有数据和公共数据之间分布偏移较大的情况，进而提升模型反演的效果。上述四种 MIA 方法均属于基于优化的白盒 MIA。据我了解，现有的基于优化的白盒 MIA 的研究工作已相对趋于成熟，而基于优化的黑盒 MIA 和仅标签 MIA 的相关研究工作因可利用的有效信息过少而难以达到令人满意的反演效果，这是在后续工作中值得深入探索的研究方向。其中反演效果较为突出的黑盒 MIA 为应用 StyleGAN[5]生成初始点并使用差分进化算法对初始点作进一步优化的两阶段方法 C2FMI[6]和应用 StyleGAN[5]并基于强化学习对 GAN 的潜在空间进行搜索的 RLB-MI[7]。而反演效果较为突出的仅标签 MIA 为构造梯度估计器并通过边界排斥的方法进行优化的 BERP-MI[8]。

2.2.2 Learning-based

虽然模型反演攻击可被看作为一个优化问题并使用合适的优化算法对问题求解，但这种基于优化的 MIA 通常只能为每个类别反演生成很少的代表性数据，并需对目标模型进行白盒访问才能取得良好的反演效果，这在现实世界中是不切实际的。故有研究工作通过设计 MIA 的另外一种执行范式来进行模型反演，也就是利用目标模型的输入与输出之间的关联性，训练一种新的攻击模型（称为反演模型）来实现目标模型输入与输出的反转，然后向训练得到的反演模型输入软标签/硬标签来反演生成每个类中的不同训练数据（这是基于优化的 MIA 做不到的）。但目前该研究方向上的工作比基于优化的 MIA 的研究工作相对较少，其中最具代表性的是基于学习的 MIA 的开创性研究工作 LB-MI[9]和对 LB-MI[9]进行改进（向训练反演模型的损失函数中加入语义损失和对训练反演模型的数据集使用对抗攻击方法进行对抗增强）的 AE-Boost-MI[10]。

3 本文方法

3.1 本文方法概述

从前文中可知，本次复现工作所选择的文献为发表在计算机视觉顶会 CVPR’2023 的 Re-thinking Model Inversion Attacks Against Deep Neural Networks[4]（Re-thinking-MI）。Re-thinking-MI 为基于优化的白盒模型反演攻击方法。该工作为开源工作，代码仓库为 https://github.com/sutd-visual-computing-group/Re-thinking_MI。

Re-thinking-MI 关注基于优化的白盒 MIA 领域，通过分析现有最先进的白盒 MIA 方法（GMI[1]和 KEDMI[2]）中可能存在的问题（次优的身份损失设计和 MI 过拟合），并提出相应的改进方案（更优的

身份损失设计和添加增强模型），最后通过充分的实验验证改进方案的有效性。图 1为 Re-thinking-MI 方法的整体示意图。图 1中的①形象的解释了该工作中考虑的 MIA 问题，即根据模型的参数来重建的私有训练数据（白盒 MIA）。图 1中的②对比分析现有最先进的白盒 MIA 方法中的优化目标，并提出了一个改进的优化目标，可显着提高 MIA 的反演效果（第 3.2 节）。图 1中的③形式化了“MI 过拟合”的概念，表明它会阻止重建图像学习训练数据的身份语义，并提出了一种新颖的“模型增强”想法来克服该问题（第 3.3 节）。图 1中的④显示该工作提出的方法显着提高了 MI 攻击的准确性。如在标准 CelebA 基准测试中，该方法将攻击准确率提高了 11.8%，在当代 MI 文献中首次实现了 90% 以上的攻击准确率。

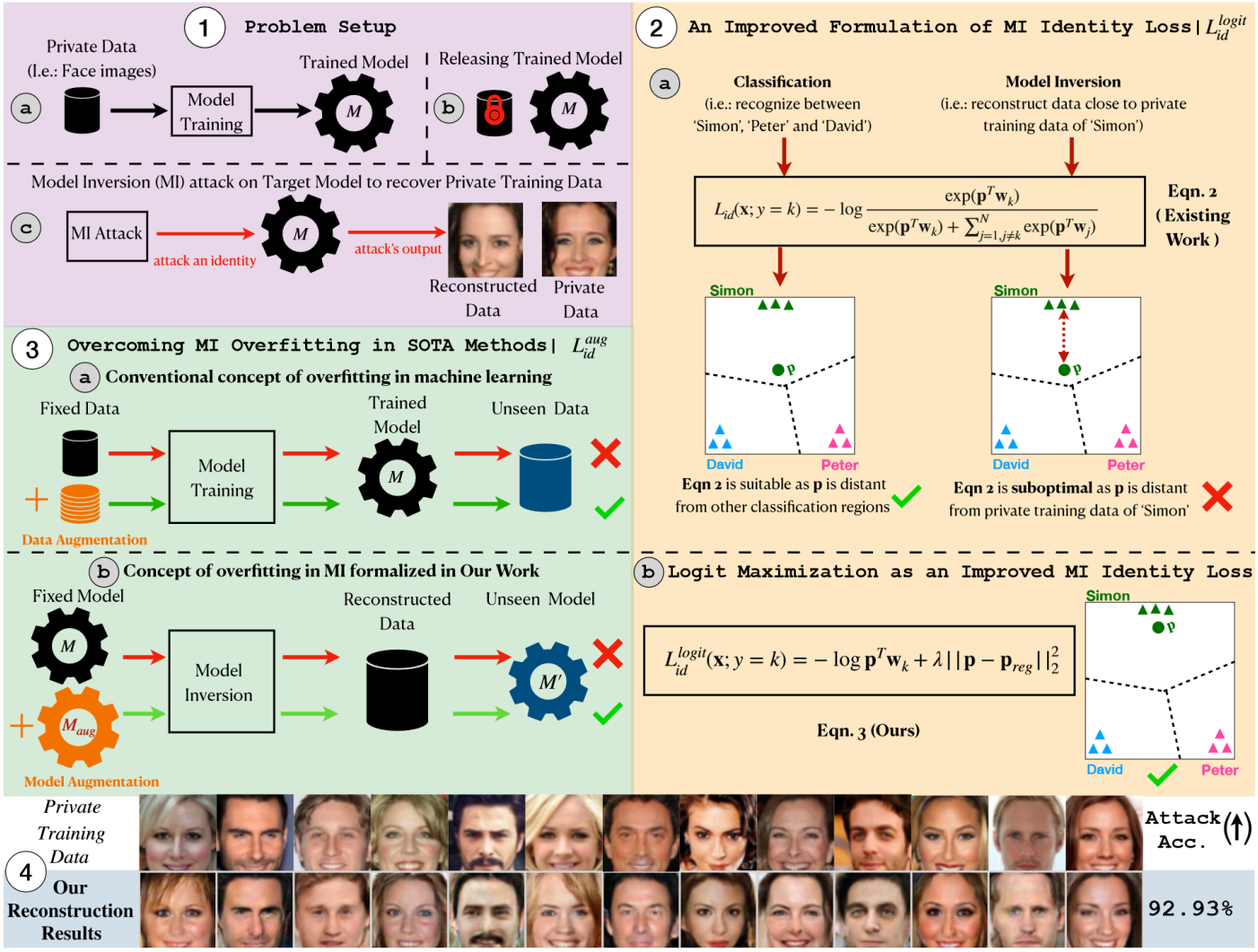

图 1: Re-thinking-MI 方法的整体示意图

3.2 更优的身份损失设计

使用 GAN 且基于优化的白盒 MIA 主要分为两个阶段，分别为公共知识蒸馏（GAN 训练）和数据恢复。在公共知识蒸馏阶段主要使用公共数据集对 GAN 进行训练，而在数据恢复阶段主要结合使用先验损失（确保生成图像更加真实自然）和身份损失（确保重建图像能拥有足够高的置信度分类到目标类别）对潜在向量进行优化。图 2中对比现有最先进的白盒 MIA 方法在数据恢复阶段的不同，其中 GMI[1]对数据进行单点恢复，而 KEDMI[2]对数据进行分布恢复，前者比后者更耗时。两种方法所使用的先验损失也不同，而身份损失均采用交叉熵损失。从图 3中 KEDMI[2]文献在同数据分布上的实验结果可知，通过改进目标 q(z) 和先验损失 L_{prior} 可明显观察到实验结果的进步，但这些 MIA 方法均未

注意到更有效的身份损失 L_{id} 的设计。

现有模型反演攻击中的身份损失 L_{id} 主要采用交叉熵损失。从图 1中的②中的ⓐ的交叉熵损失的公式可知，该损失的优化目标是最大化分子和最小化求和项。该优化目标确实可以使重建数据远离其他分类区域，但却和目标类别的私有数据距离过远，即将交叉熵损失作为身份损失确实适用于分类问题，但并不太适合 MIA 的目标——生成与目标类别的私有训练数据相似的样本（代表性数据）。所以，作者认为应注重最大化分子，故取消了激活函数 softmax，直接在 logit 值上进行损失的计算。

更优的身份损失公式如图 1中的②中的ⓑ所示。该公式中除直接在 logit 值上进行损失计算外，还加入了一个正则化项 $||P - P_{reg}||_2^2$，其中 P 为目标模型倒数第二层的激活值。该正则化项有助于缓解 MI 过拟合。此外，P_{reg} 用于正则化 P，若没有 P_{reg} 正则化 P，仅使用 $||P||_2^2$，由于 $||P||$ 是无界的，MIA 问题的优化目标变为最大化 $||P||$，这可能导致数值变得很大，而且限制 P 的数值范围对于数值稳定性非常重要，特别是在优化过程中。由于攻击者无法访问私有训练数据，作者使用公共数据对 P_{reg} 进行估计。首先作者随机采样 5000 张公共数据中的图像，然后构建这些图像在目标模型倒数第二层的特征集合，最后使用两种不同方法来估计 P_{reg}。第一种方法是直接将该集合的均值作为 P_{reg}，第二种方法则是从分布（该集合的均值和方差作为分布的均值和方差）中进行采样，将采样得到的值作为 P_{reg}。从图 4中 KEDMI 中使用不同 P_{reg} 选择方法的同数据分布的实验结果可知，第二种方法的反演效果优于第一种方法，故在后续实验中均采用第二种 P_{reg} 选择方法。

Method	**Latent distribution** $q(\mathbf{z})$	**Prior loss** L_{prior}
GMI [52]	Point estimate $\delta(\mathbf{z} - \mathbf{z}_0)$	$-D(G(\mathbf{z}))$
KEDMI [7]	Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$-\log D(G(\mathbf{z}))$

图 2: 现有最先进的白盒 MIA 方法

	face.evolve		**IR152**		**VGG16**	
	GMI	**Ours**	**GMI**	**Ours**	**GMI**	**Ours**
Attack Acc ↑	.31±.0039	**.81±.0016**	.32±.0027	**.81±.0015**	.21±.0020	**.72±.0018**
Top-5 Attack acc ↑	.53±.0015	**.96±.0004**	.57±.0005	**.96±.0001**	.43±.0014	**.92±.0003**
KNN Dist ↓	1703.52	**1358.23**	1673.05	**1324.72**	1772.50	**1380.22**
FID ↓	33.81	**25.28**	50.11	**26.35**	52.51	**23.72**

图 3: KEDMI 中同数据分布的实验结果

Method	Attack Acc ↑	KNN dist ↓
CelebA/CelebA/IR152		
+ LOM (Fixed p_{reg})	92.27 ± 1.37	**1155.92**
+ LOM (Ours)	**92.47 ± 1.41**	1168.55
CelebA/CelebA/face.evoLve		
+ LOM (Fixed p_{reg})	90.40 ± 1.68	1257.95
+ LOM (Ours)	**92.53 ± 1.51**	**1183.76**
CelebA/CelebA/VGG16		
+ LOM (Fixed p_{reg})	85.60 ± 1.79	1259.60
+ LOM (Ours)	**89.07 ± 1.46**	**1218.46**

图 4: KEDMI 中使用不同 P_{reg} 选择方法的同数据分布的实验结果

3.3 缓解 MI 过拟合

作者首先提出了“MI 过拟合”的概念，然后定性和定量分析了“MI 过拟合”问题。表 2中详细列出了作者通过分析传统过拟合的定义和缓解方法，给出了自己对“MI 过拟合”的定义和缓解方法。

表 2: 传统过拟合和 MI 过拟合对比

概念	定义	缓解方法
传统过拟合	模型训练过程中，模型对训练数据进行过于紧密的拟合，导致模型在未见过的数据下表现欠佳	增加训练数据
MI 过拟合	模型反演过程中，重建样本与目标模型拟合得过于紧密，导致样本缺乏身份语义，进而导致样本在未见过的模型下表现欠佳	添加增强模型

此外，作者还对“MI 过拟合”问题进行了定性和定量分析，如图 5所示。从图 5中的①的私有训练数据和存在“MI 过拟合”问题的 MIA 方法反演得到的样本的可视化结果可知，这些存在“MI 过拟合”问题的 MIA 方法反演得到的样本虽然身份损失很低，但明显缺少身份语义。从图 5中的②将存在“MI 过拟合”问题的 MIA 方法在 IR152 模型上反演得到的样本输入从未见过的模型（VGG16 和 EfficientNet B0）上计算身份损失可知，这些存在“MI 过拟合”问题的 MIA 方法反演得到的样本虽然在目标模型上身份损失很低，但在其他未见过的模型上身份损失很低，但明显升高，而且不在少数。

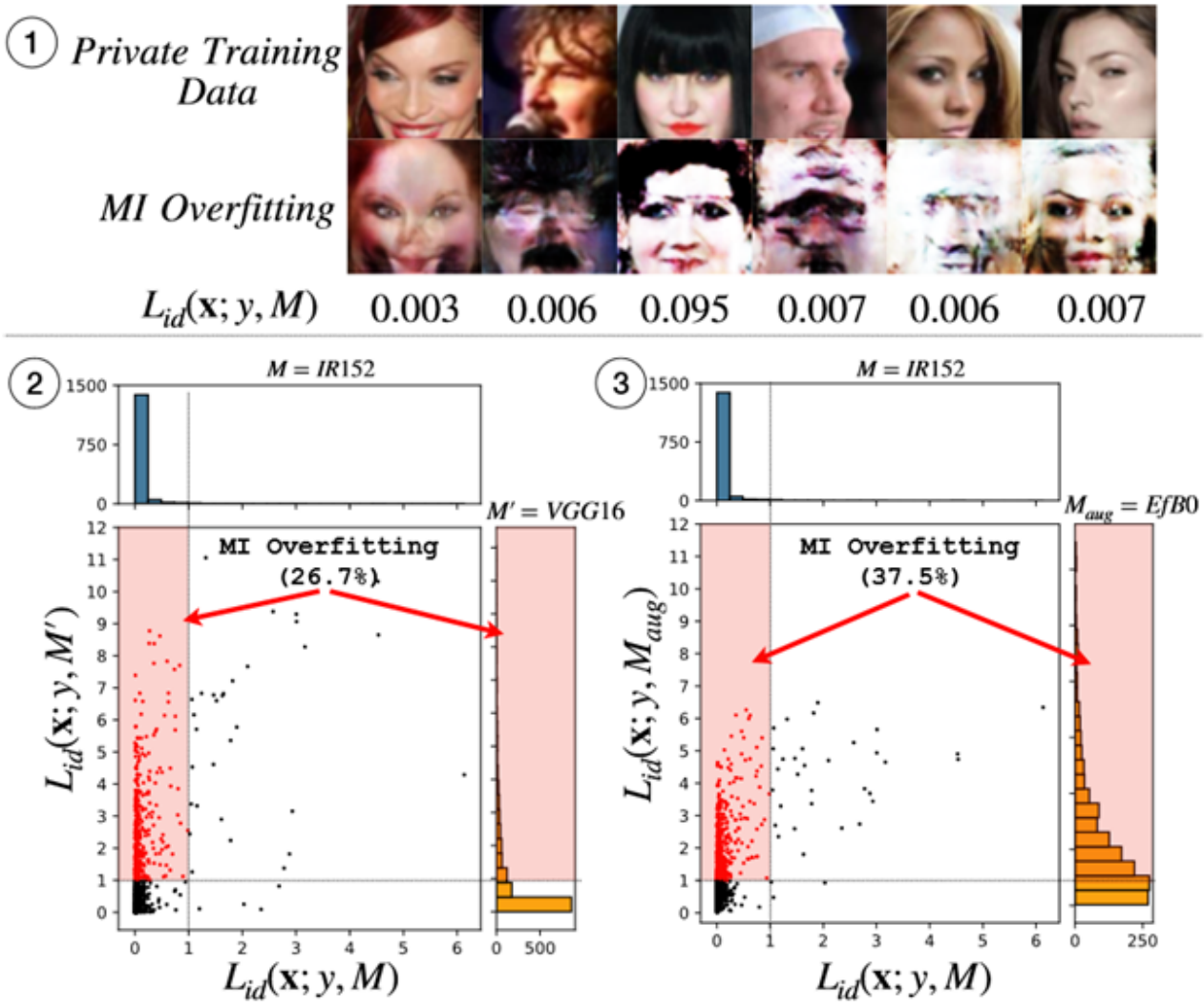

图 5: MI 过拟合的定性和定量分析

作者基于集成的思想，通过添加增强模型来缓解“MI 过拟合”问题，即反演过程从攻击单个目标模型变为攻击多个目标模型。基于知识蒸馏[11]对增强模型进行训练，也就是目标模型作为教师模型，增强模型作为学生模型，并基于 KL 散度损失（衡量教师模型和学生模型对同一数据的预测分布之间的差异）对增强模型进行训练。

此外，作者还对增强模型的个数和架构进行了实验。图 6中使用依次叠加 EfficientNet 家族中的 B0 - B3 模型作为增强模型。虽然增强模型为 4 个时反演攻击的准确率最高，但实验耗时过长，故退而求其次选择使用 3 个增强模型，即在后续实验中仅使用 3 个增强模型（EfficientNet B0 - B2）。图 7为使用不同的模型架构作为增强模型中进行实验的实验结果。虽然使用 DenseNet B0 - B2 作为增强模型时反演攻击的准确率最高，但为减少实验耗时，在后续实验中使用 EfficientNet B0 - B2 作为增强模型。

Method	N_{aug}	M_{aug}	Attack Acc ↑	Imp. ↑	KNN dist ↓
		CelebA/CelebA/IR152			
KEDMI	-	-	80.53 ± 3.86	-	1247.28
+ MA	1	EfficientNet-B0	81.20 ± 3.75	0.67	1234.16
+ MA	2	EfficientNet-B0, EfficientNet-B1	84.47 ± 2.99	3.94	1223.56
+ MA	3	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2	84.73 ± 3.76	4.20	1220.23
+ MA	4	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3	**85.87 ± 2.63**	**5.34**	**1217.15**

图 6: 不同增强模型个数的实验结果

Method	M_{aug}	Attack Acc ↑	Imp. ↑	KNN dist ↓
	CelebA/CelebA/IR152			
KEDMI	-	80.53 ± 3.86	-	1247.28
+ MA (Ours-1)	EfficientNet-B0, EfficientNet-B1, EfficientNet-B2	**84.73 ± 3.76**	**4.20**	**1220.23**
+ MA (Ours-2)	DenseNet121, DenseNet161, DenseNet169	**89.07 ± 3.32**	**8.54**	**1211.73**
+ MA (Ours-3)	EfficientNet-B0, DenseNet121, MobileNetV3-large	**86.53 ± 1.98**	**6.00**	**1204.94**

图 7: 不同增强模型架构的实验结果

4 复现细节

由于本次复现工作所选择的论文的代码为开源的，但开源代码较为冗余混乱，不够简洁，所以我对开源代码进行简单的重构，使其更易于理解。

4.1 实验设置

本次实验和论文中一样，也使用 CelebA 和 FFHQ 两个大型人脸图像数据集。其中从 CelebA 数据集中采样 30000 张人脸图像共 1000 个身份（0-999）作为私有数据集，划分 27000 张作为训练集训练目标模型，3000 张作为测试集测试目标模型和增强模型。从 CelebA 数据集中采样 30000 张与私有数据集身份不重叠的人脸图像作为公共数据集来训练 GAN 和增强模型，从而进行同数据分布（私有数据和公共数据来自同一个数据集）实验。同样，从 FFHQ 数据集中采样 70000 张与私有数据集身份不重叠的人脸图像作为公共数据集来训练 GAN 和增强模型，从而进行非同数据分布（私有数据和公共数据来自不同数据集）实验。本次实验和论文中相同，均基于 GMI[1]和 KEDMI[2]进行实验并采用 Top-1 攻击精度（重建图像被评估模型正确分类的比例）和 Top-5 攻击精度（建图像被评估模型分类在排名前 5 的比例）作为攻击效果的评估指标。此外，同样使用 IR152、VGG16 和 FaceNet64 作为目

标模型的架构，使用 EfficientNet B0 - B2 作为增强模型的架构，使用 FaceNet 作为评估模型的架构。本次实验的详细设置见表 3。

表 3: 实验设置

Setting	Details		
Datasets	Private Dataset	CelebA	27000 张作为训练集（训练目标模型），3000 张作为测试集（测试目标模型和增强模型）
	Public Dataset	CelebA	30000 张（训练 GAN 和增强模型，同数据分布实验）
		FFHQ	70000 张（训练 GAN 和增强模型，非同数据分布实验）
Baselines	GMI KEDMI		
Evaluation Metrics	Top-1 Attack Acc	Top-1 攻击精度（重建图像被评估模型正确分类的比例）	
	Top-5 Attack Acc	Top-5 攻击精度（重建图像被评估模型分类在排名前 5 的比例）	
Models	Target Model	IR152（VGG16、FaceNet64）	
	Augmented Model	EfficientNet B0 - B2	
	Evaluation Model	FaceNet	

4.2 模型训练

由于本次实验需要进行同数据分布实验和非同数据分布实验，所以需要训练的模型非常之多。目标模型为使用 CelebA 私有数据集训练的 IR152、VGG16 和 FaceNet64。评估模型为使用 CelebA 私有数据集训练的 FaceNet。GAN 为分别使用 CelebA 公共数据集和 FFHQ 公共数据集训练的通用 GAN（GMI[1]）和特定于反演的 GAN（KEDMI[2]）（需为不同的目标目标训练相对应的 GAN）。增强模型为使用 FFHQ 公共数据集训练的 EfficientNet B0 - B2（需为不同的目标模型和 GAN 训练相对于的增强模型）。此外，还需对反演过程中需要使用的 P_{reg} 进行估计。

在模型训练的过程中，我发现原论文中使用 FFHQ 公共数据集训练的增强模型 EfficientNet B0 - B2 的预测精度过低，我猜测其原因可能是使用的损失函数过于简单——知识蒸馏过程仅使用 KL 散度损失。我通过阅读知识蒸馏的开山之作 [11] 发现该论文中在对教师模型进行知识蒸馏时除使用 KL 散度损失外，还使用交叉熵损失来计算学生模型的预测分布与真实标签之间的差距。[11] 文献解释其原因是教师模型对输入数据的预测并非百分百准确，需通过上述操作对学生模型进行校正，避免其过度拟合教师模型的预测结果。但可惜的是本次复现工作中训练增强模型的 FFHQ 公有数据集并未提供相对应的真实标签，无法实施与文献 [11] 中同样的操作。因此，我在 KL 散度损失的基础上，加入 L2 范数来平衡 KL 散度损失的贡献，避免学生模型过度拟合教师模型的预测结果。也就是使用 L2 范数

计算教师模型和学生模型预测 logit 值之间的差异，再结合 KL 散度损失来训练增强模型。加入 L2 范数前后训练得到的增强模型 EfficientNet B0 - B2 的 Top-1 预测精度和 Top-5 预测精度见表 4，大致提高了增强模型 4% - 12% 的预测精度。

表 4: 加入 L2 范数前后增强模型的预测精度

	加入 L2 范数前					
	Top-1 Acc ↑			**Top-5 Acc ↑**		
EfficientNet	**B0**	**B1**	**B2**	**B0**	**B1**	**B2**
CelebA	65.86%	64.76%	67.82%	85.27%	83.28%	85.51%
FFHQ	33.05%	44.45%	41.59%	59.08%	69.71%	66.29%
	加入 L2 范数后					
	Top-1 Acc ↑			**Top-5 Acc ↑**		
EfficientNet	**B0**	**B1**	**B2**	**B0**	**B1**	**B2**
CelebA	75.53%	76.33%	77.26%	89.16%	89.69%	90.13%
FFHQ	42.92%	49.04%	45.84%	66.06%	73.87%	70.55%

模型训练结果如图 8-图 13。图 8中显示使用 CelebA 私有数据集训练的目标模型和评估模型，其中 IR152 的测试精度为 92.42%、VGG16 的预测精度为 86.37% 和 FaceNet64 的预测精度为 82.58%，评估模型 FaceNet 的预测精度为 96.34%。图 9中显示使用 CelebA 公共数据集训练的 GAN。图 10中显示使用 FFHQ 公共数据集训练的 GAN。图 11中显示使用 CelebA 公共数据集训练的增强模型。图 12中显示使用 FFHQ 公共数据集训练的 GAN。图 13中显示估计得到的 P_{reg} 文件。

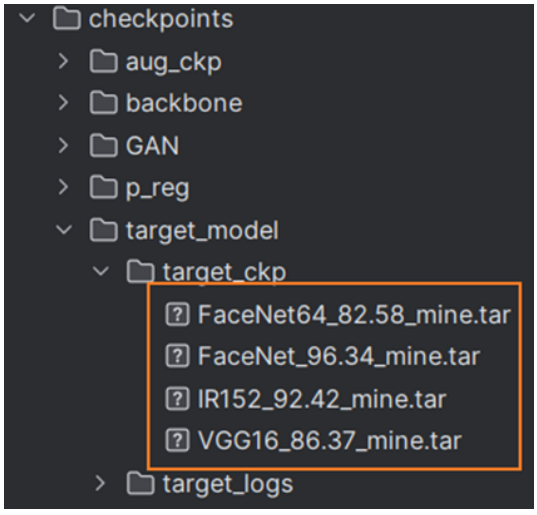

图 8: 目标模型和评估模型训练结果

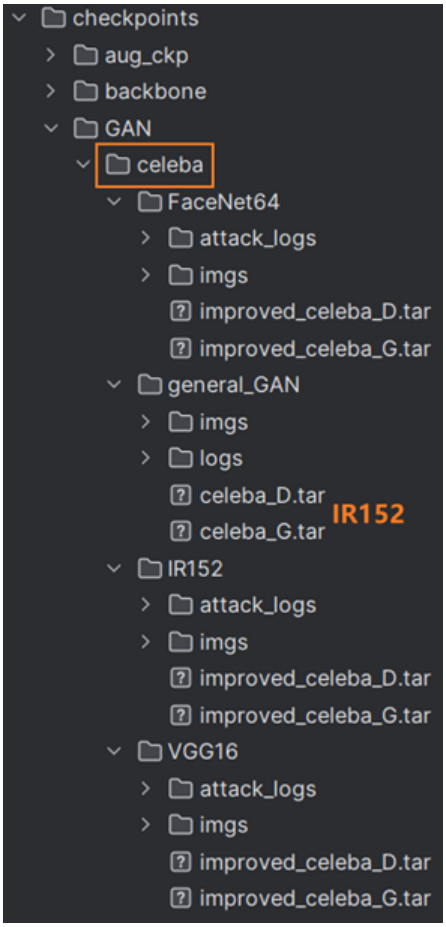

图 9: GAN（CelebA）

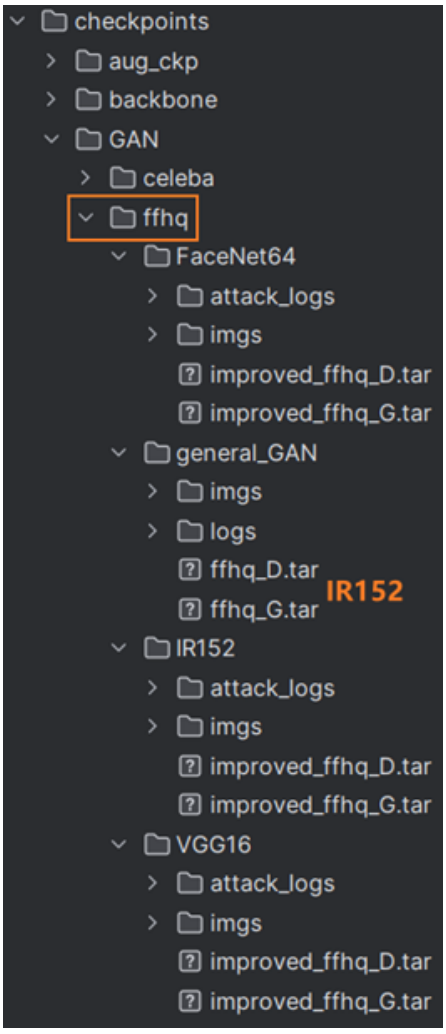

图 10: GAN（FFHQ）

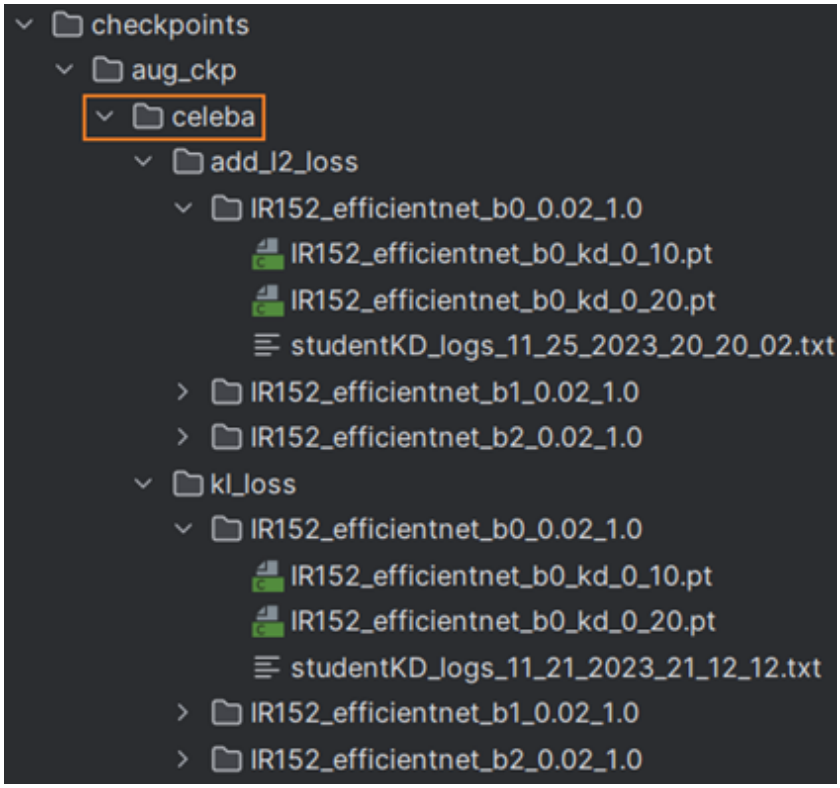

图 11: 增强模型（CelebA）

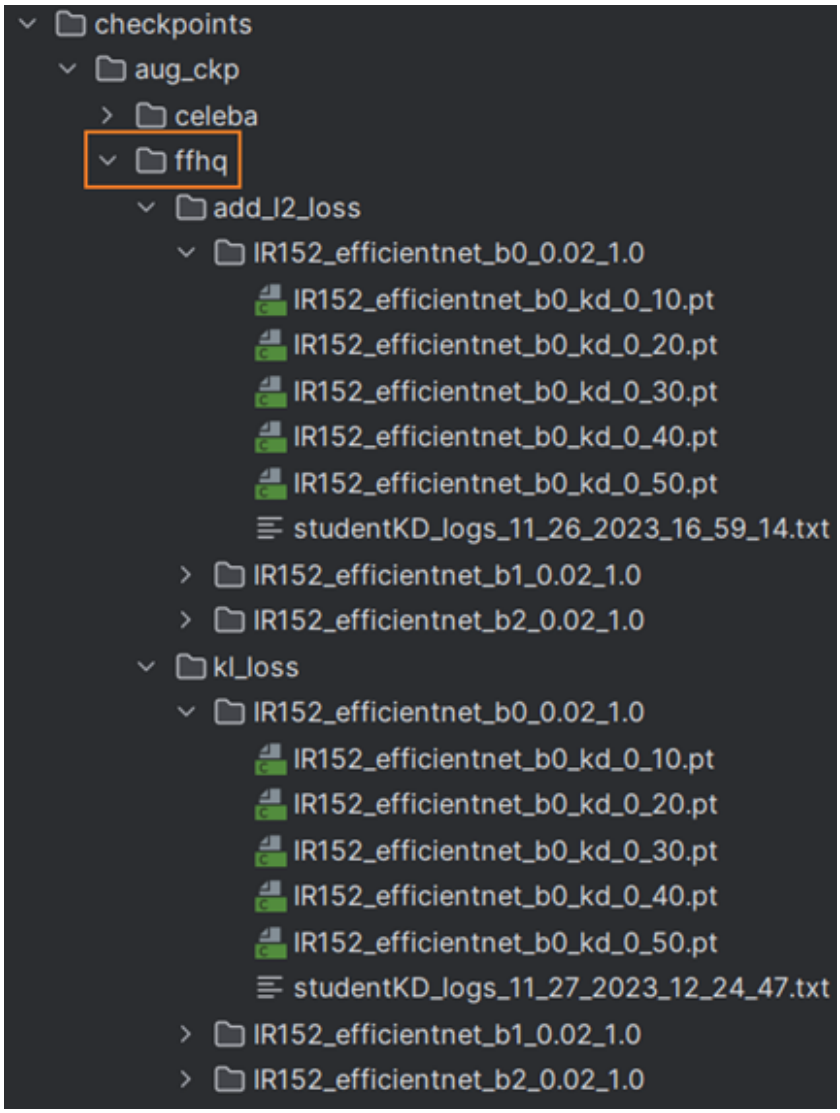

图 12: 增强模型（FFHQ）

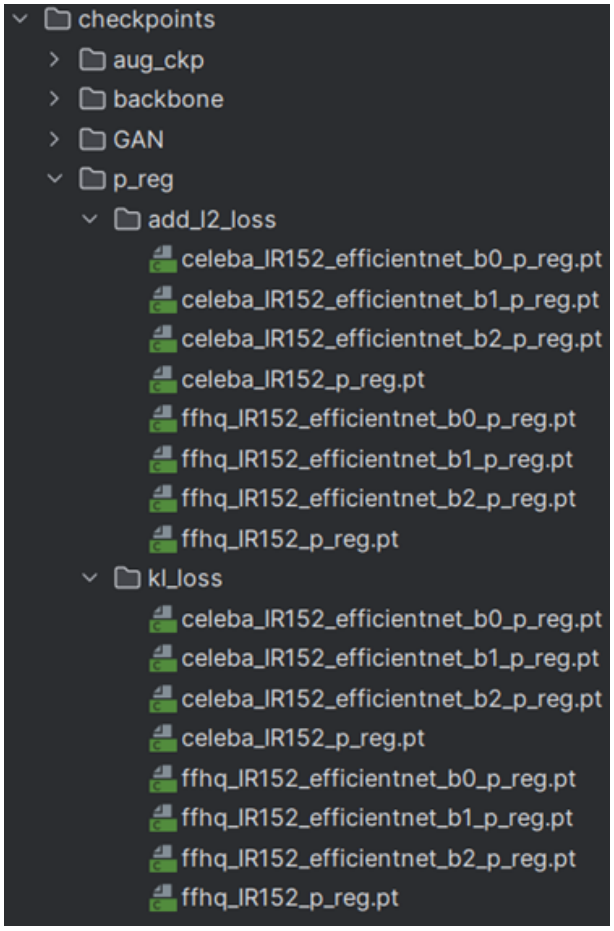

图 13: P_{reg} 文件

综上所述，对基线方法 GMI[1]和 KEDMI[2]的改进方案共 4 种，分别为 + LOM、+ MA、+ LOMMA 和 + LOMMA + L2_loss（详见表 5）。

表 5: 改进方案总结

改进方案	定义
+ LOM	GMI 和 KEDMI 使用改进的身份损失
+ MA	GMI 和 KEDMI 使用增强模型
+ LOMMA	GMI 和 KEDMI 结合使用改进的身份损失和增强模型（KL 散度损失）
+ LOMMA + L2_loss	GMI 和 KEDMI 结合使用改进的身份损失和增强模型（KL 散度损失 + L2 范数）

4.3 攻击结果

本次实验我进行了同数据分布实验和非同数据分布实验。每个实验均攻击 300 个身份（0 - 299，共 1000 个身份），每次攻击 60 个身份（共进行 5 次），每个身份进行 5 轮攻击（重建 5 张图像），每轮进行 2400 次迭代优化。从前文可知，已经训练好模型反演攻击中所需要的目标模型、GAN、增强模型和评估模型以及估计 P_{reg} 所得到的文件。故可开展同数据分布和非同数据分布实验，此处仅展示攻击目标模型 IR152 的同数据分布和非同数据分布的实验结果，同数据分布实验的详细攻击结果见表 6，非同数据分布实验的详细攻击结果见表 7。从表 6和表 7中的实验结果可知，随着改进方案的叠加，

模型反演的攻击精度呈阶梯式上升。其中，使用加入 L2 范数训练得到增强模型进行反演攻击，同数据分布实验和非同数据分布实验的攻击精度均有所提升，大约提升了 2% - 5% 的攻击精度。此外，我还从攻击成功的身份中选取部分私有数据和重建图像进行可视化展示。如图 14中，同数据分布实验我选取了身份 0、3 和 9 进行可视化展示。如图 15中，非同数据分布实验我选取了身份 33、165 和 263 进行可视化展示。从可视化展示中可知，身份对应的重建图像与私有图像十分相似。

表 6: 同数据分布实验攻击结果

Methods	Top-1 Attack Acc ↑	Top-5 Attack Acc ↑
CelebA / CelebA / IR152 / EfficientNet B0 - B2 / FaceNet		
KEDMI	88.93% ± 2.26%	98.33% ± 1.11%
+ LOM	92.13% ± 2.50%	99.33% ± 0.51%
+ MA	92.23% ± 0.99%	99.00% ± 0.00%
+ LOMMA	**93.60% ± 0.99%**	**99.33% ± 0.18%**
+ LOMMA + L2_loss	**95.87% ± 0.89%**	**99.00% ± 0.00%**
GMI	41.00% ± 6.28%	68.00% ± 6.87%
+ LOM	73.73% ± 3.86%	91.67% ± 3.52%
+ MA	76.27% ± 5.01%	97.33% ± 1.98%
+ LOMMA	**85.07% ± 3.25%**	**98.67% ± 2.00%**
+ LOMMA + L2_loss	**90.00% ± 3.85%**	**99.33% ± 1.42%**

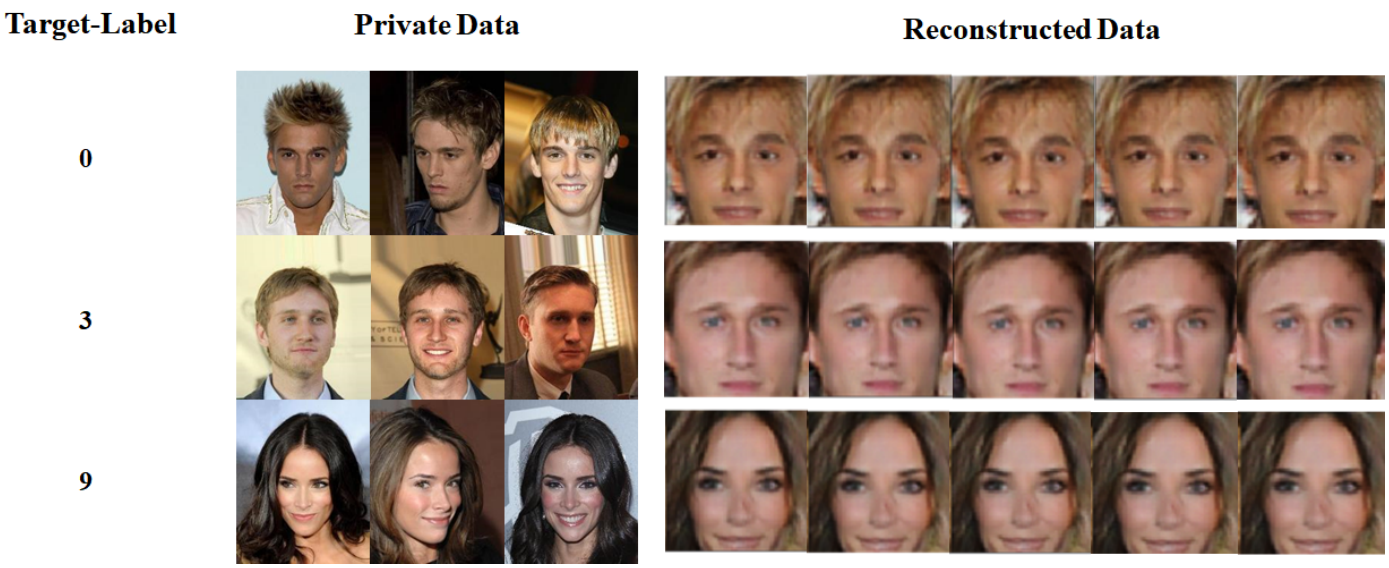

图 14: 同数据分布实验可视化结果

表 7: 非同数据分布实验攻击结果

Methods	Top-1 Attack Acc ↑	Top-5 Attack Acc ↑
CelebA / FFHQ / IR152 / EfficientNet B0 - B2 / FaceNet		
KEDMI	65.67% ± 3.19%	90.33% ± 1.95%
+ LOM	74.73% ± 4.01%	94.00% ± 1.61%
+ MA	71.00% ± 1.94%	93.33% ± 0.60%
+ LOMMA	**79.60% ± 1.47%**	**96.33% ± 0.76%**
+ LOMMA + L2_loss	**83.80% ± 2.17%**	**96.33% ± 0.51%**
GMI	30.33% ± 3.58%	54.67% ± 4.40%
+ LOM	68.47% ± 4.30%	88.00% ± 3.03%
+ MA	66.80% ± 5.15%	90.67% ± 3.20%
+ LOMMA	**76.93% ± 4.56%**	**94.33% ± 2.31%**
+ LOMMA + L2_loss	**78.80% ± 3.96%**	**94.00% ± 2.61%**

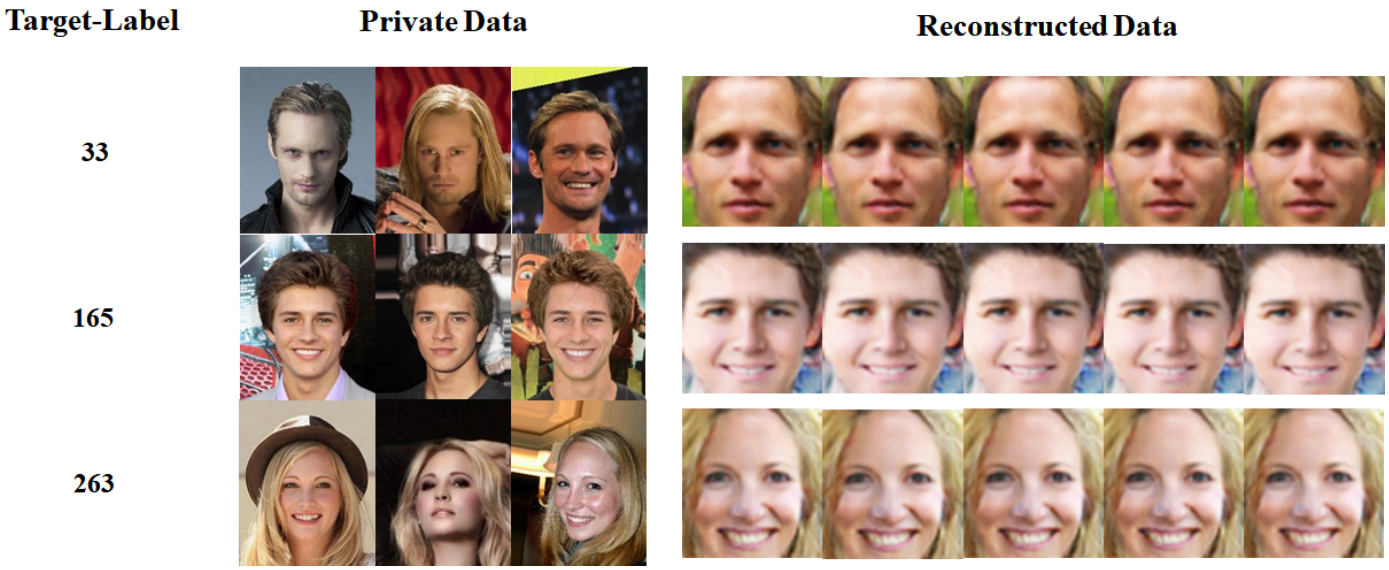

图 15: 非同数据分布实验可视化结果

5 总结与展望

当下深度神经网络已被广泛应用于涉及私人和敏感数据集的诸多领域，如人脸识别、语音识别、医疗保健等。人们也越发担心为获取训练 DNNs 时使用的机密数据集的相关知识而进行的隐私攻击，特别是模型反演攻击。Re-thinking-MI[4]方法通过分析目前最先进的白盒 MI 方法（GMI[1]和 KEDMI[2]）中存在的影响反演攻击精度的问题（身份损失的优化目标不太适合 MI 的目标和“MI 过拟合”问题），

并提出了相应的改进方案（直接在 logit 值上计算损失并加入正则化项和添加增强模型），最后通过充分地实验验证了上述改进方案的有效性。在本次复现过程中，我发现原论文中基于知识蒸馏所得到的增强模型的预测精度较低，故通过改进知识蒸馏过程所使用的损失函数（KL 散度损失和 L2 范数结合），提高了增强模型 4% - 12% 的预测精度，进而提高了 2% - 5% 的反演攻击精度。

由于时间有限，并未对我认为可进一步提高反演攻击精度的方案进行实验验证：

（1）为训练增强模型的 FFHQ 公共数据集使用目标模型添加伪标签（该伪标签需能抵抗随机噪声的扰动，即加入随机噪声不会改变目标模型的预测结果），然后使用这些伪标签来矫正学生模型的预测结果（文献 [11]），可进一步增大增强模型的预测精度，从而进一步提高反演攻击精度。

（2）将使用 GMI[1]和 KEDMI[2]中的 GAN 替换为 StyleGAN[5]，然后尝试在经过映射网络前的 z 空间和经过合成网络前的 w 空间进行优化，我认为可进一步提高反演攻击精度，同时重建高分辨率的人脸图像。

Re-thinking-MI[4]属于基于优化的白盒 MIA。从前文可知，基于优化的白盒 MIA 的现有研究工作已经相对趋于成熟，其研究空间较为狭窄。但将该方向的研究工作作为我步入科研的起点的非常合适的，因为该方向相比于基于优化/基于学习的黑盒 MIA 和仅标签 MIA 较为简单，复现难度中等，同时可以让我对整个 MIA 领域的研究发展有了一定的了解。此外，基于优化/基于学习的黑盒 MIA 和仅标签 MIA 仍有非常大的研究空间，这也是我后续进一步开展的研究工作。

参考文献

[1] Zhang Y, Jia R, Pei H, et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks[J]. 2020: 253-261.

[2] Chen S, Kahla M, Jia R, et al. Knowledge-Enriched Distributional Model Inversion Attacks[J]. 2021: 16178-16187.

[3] Yuan X, Chen K, Zhang J, et al. Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network[J]. 2023: 1-13.

[4] Nguyen N B, Chandrasegaran K, Abdollahzadeh M, et al. Re-Thinking Model Inversion Attacks Against Deep Neural Networks[J]. 2023: 16384-16393.

[5] Karras T, Laine S, Aittala M, et al. Analyzing and Improving the Image Quality of StyleGAN[J]. 2020: 8110-8119.

[6] Ye Z, Luo W, Naseem M L, et al. C2FMI: Corse-to-Fine Black-box Model Inversion Attack[J]. IEEE Transactions on Dependable and Secure Computing, 2023: 1-15.

[7] Han G, Choi J, Lee H, et al. Reinforcement Learning-Based Black-Box Model Inversion Attacks[J]. 2023: 20504-20513.

[8] Kahla M, Chen S, Just H A, et al. Label-Only Model Inversion Attacks via Boundary Repulsion[J]. 2022: 15045-15053.

[9] Yang Z, Zhang J, Chang E C, et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment[J]. CCS'19 2019: 225-240.

[10] Zhou S, Zhu T, Ye D, et al. Boosting Model Inversion Attacks with Adversarial Examples[J]. IEEE Transactions on Dependable and Secure Computing, 2023: 1-18.

[11] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J]. 2015. arXiv: 1503.02531 [stat.ML].