ISCAS 中国科学院软件研究所学术年会 2025 第十届
暨重点实验室科技活动周

学术论文

Enhanced Label-Only Membership Inference Attacks with Fewer Queries

仅标签场景下的增强型成员推理攻击

Published in USENIX Security2025. (CCF A)

李昊，李政，吴思源，叶宇桐，张敏†，冯登国，张阳

联系人：李昊 13488718664 lihao@iscas.ac.cn

研究背景

成员推理攻击（MIA）旨在判断某一样本是否属于机器学习模型的训练集，从而揭示潜在的隐私泄露风险。在“仅标签”场景下，攻击者只能获取模型输出的预测标签，而无法访问完整的概率向量。现有方法通常基于最短边界距离（shortestBD）进行推断，认为训练集中样本的 shortestBD 更大。然而，该指标依赖对抗扰动技术估计，通常需要数千次模型查询，计算成本高。如图1所示，其对成员与非成员样本的区分效果也较为有限。

relScore

针对现有仅标签场景下成员推理攻击（MIA）依赖大量模型查询且区分效果有限的问题，本文提出一种基于相对分数（relScore）分布的新攻击方法。不同于传统方法通过对抗扰动寻找最短边界距离，我们计算样本到分布外固定点的距离（fixedBD）。利用不含目标样本的辅助数据集训练多个影子模型，构建其 fixedBD 分布，并将目标样本在该分布下的累积分布函数（CDF）值定义为 relScore。如图1所示，该指标提升了成员与非成员样本之间的区分能力，同时大幅减少了对目标模型的查询次数。

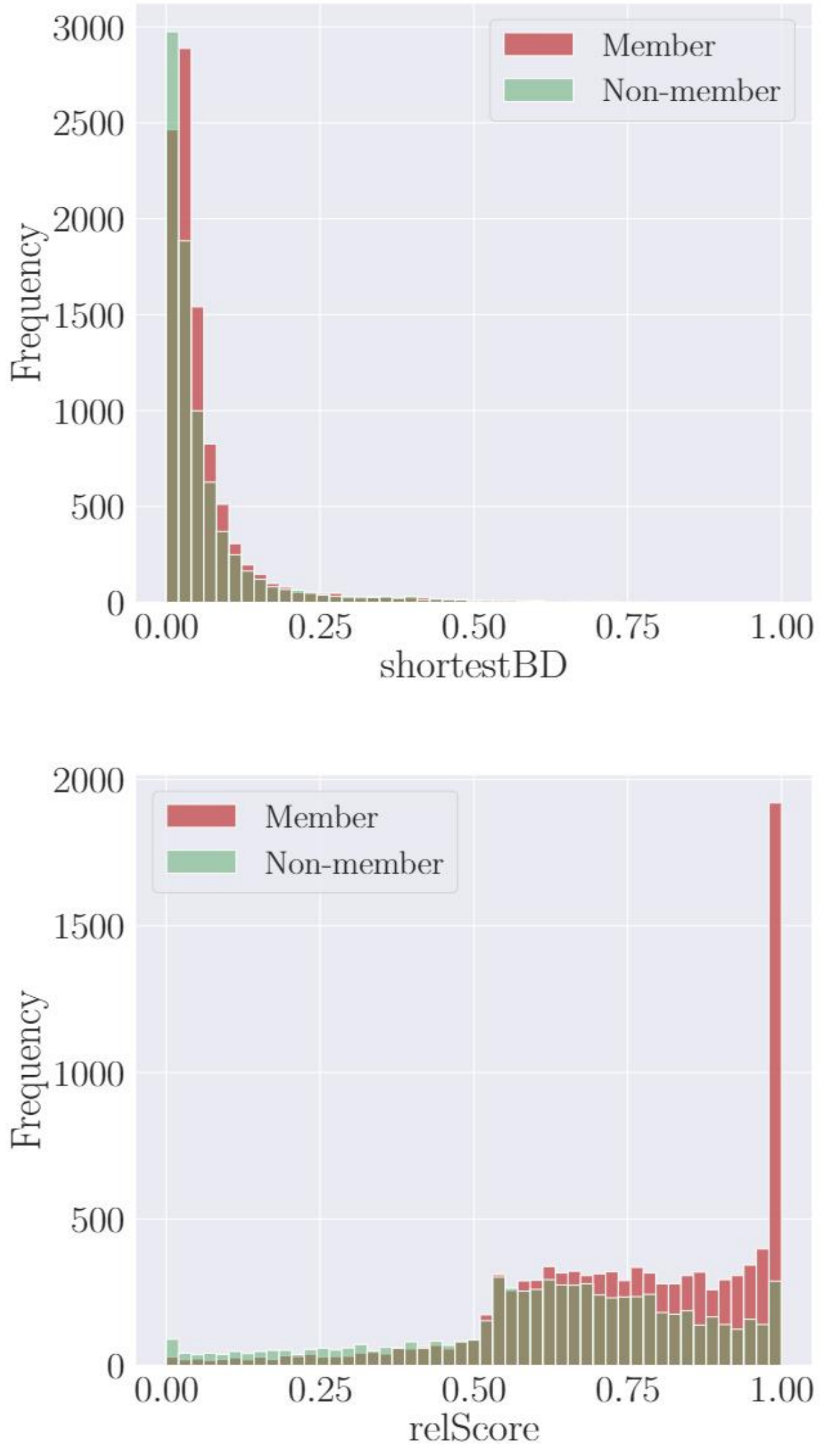


图1.样本最短距离和相对分数分布

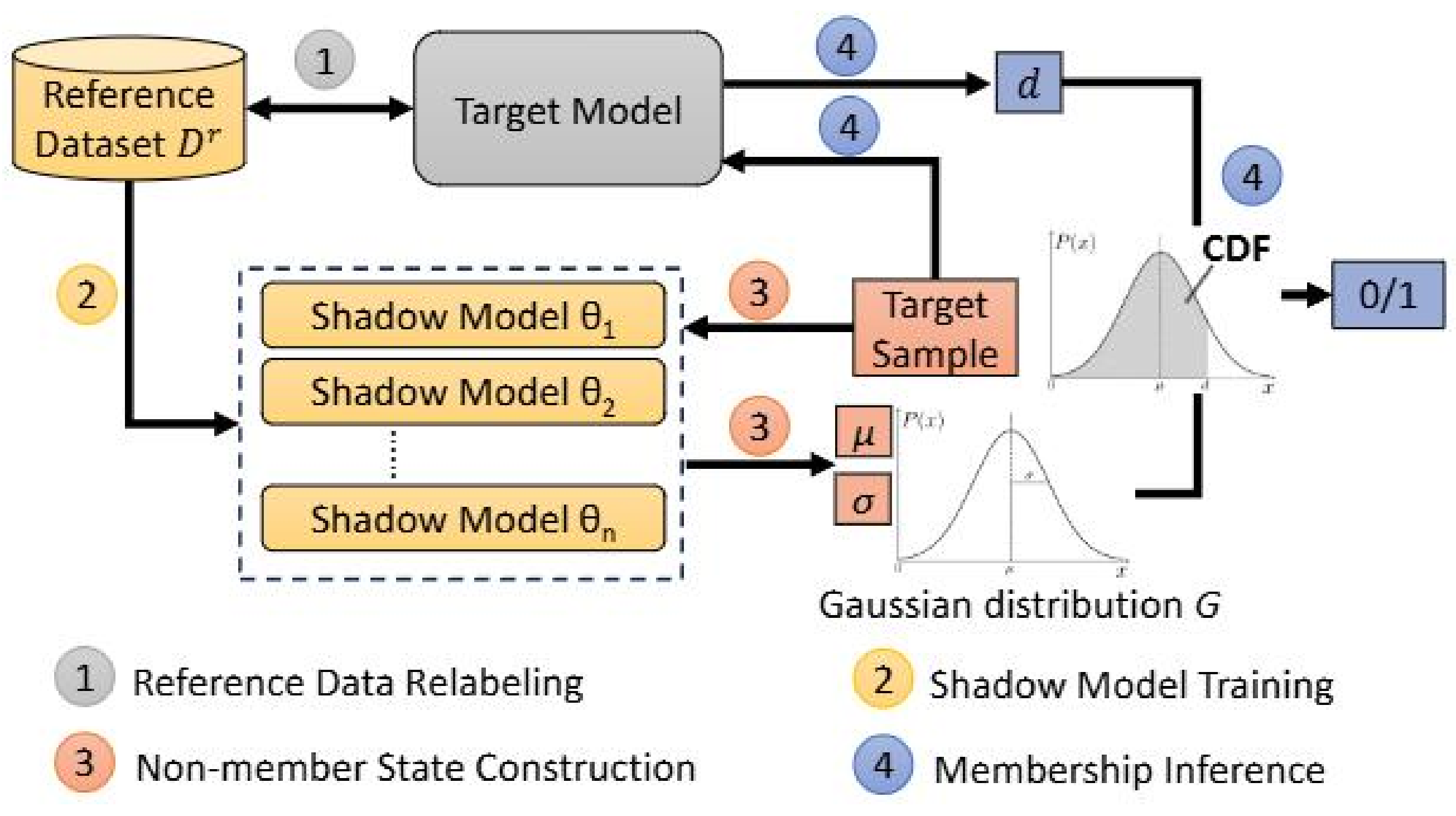


图2.DHAttack的工作流程

DHAttack

本文提出了一种兼顾高性能与高效率的MIA方法——DHAttack（如图2所示），主要包括：

① 利用目标模型对辅助数据集进行预测标记，以模拟目标模型的预测行为；

② 基于重标记的辅助数据集训练多个影子模型；

③ 计算目标样本在各影子模型中的fixedBD（如图3所示），并构建非成员时的fixedBD分布；

④ 计算目标样本在目标模型中的 fixedBD，并结合非成员状态下的 fixedBD 分布计算其 relScore，从而完成对其成员身份的推断。

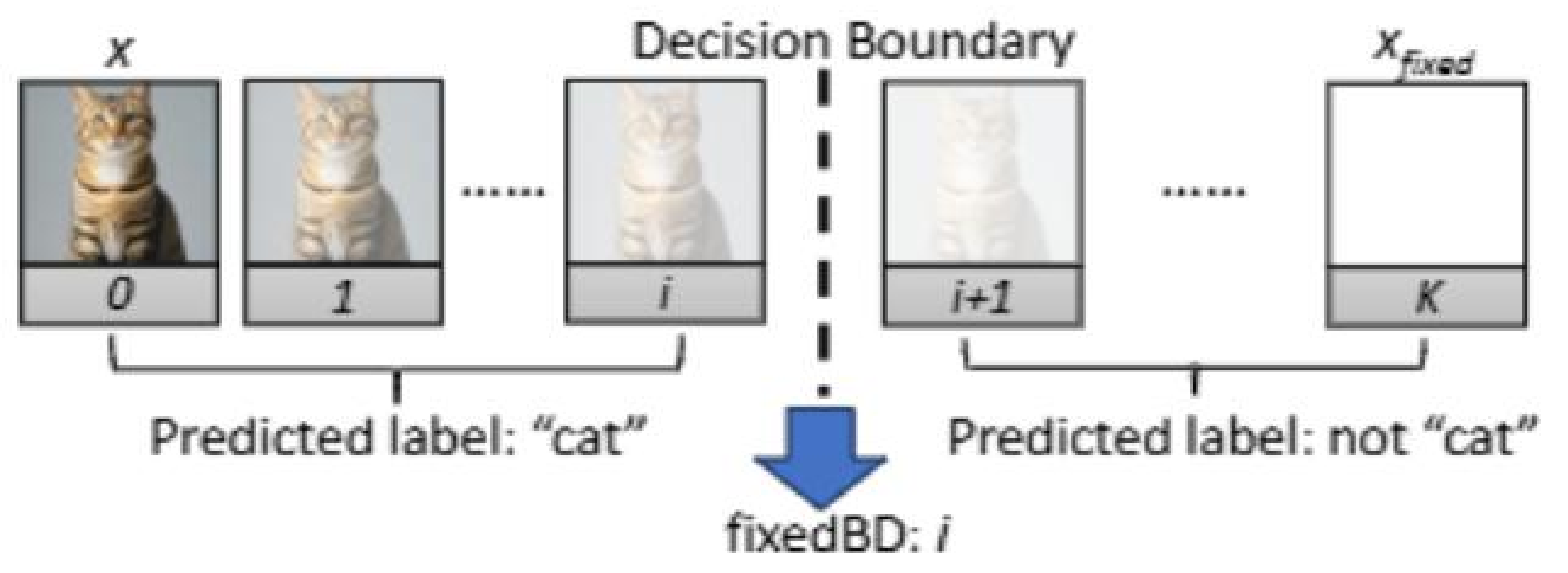
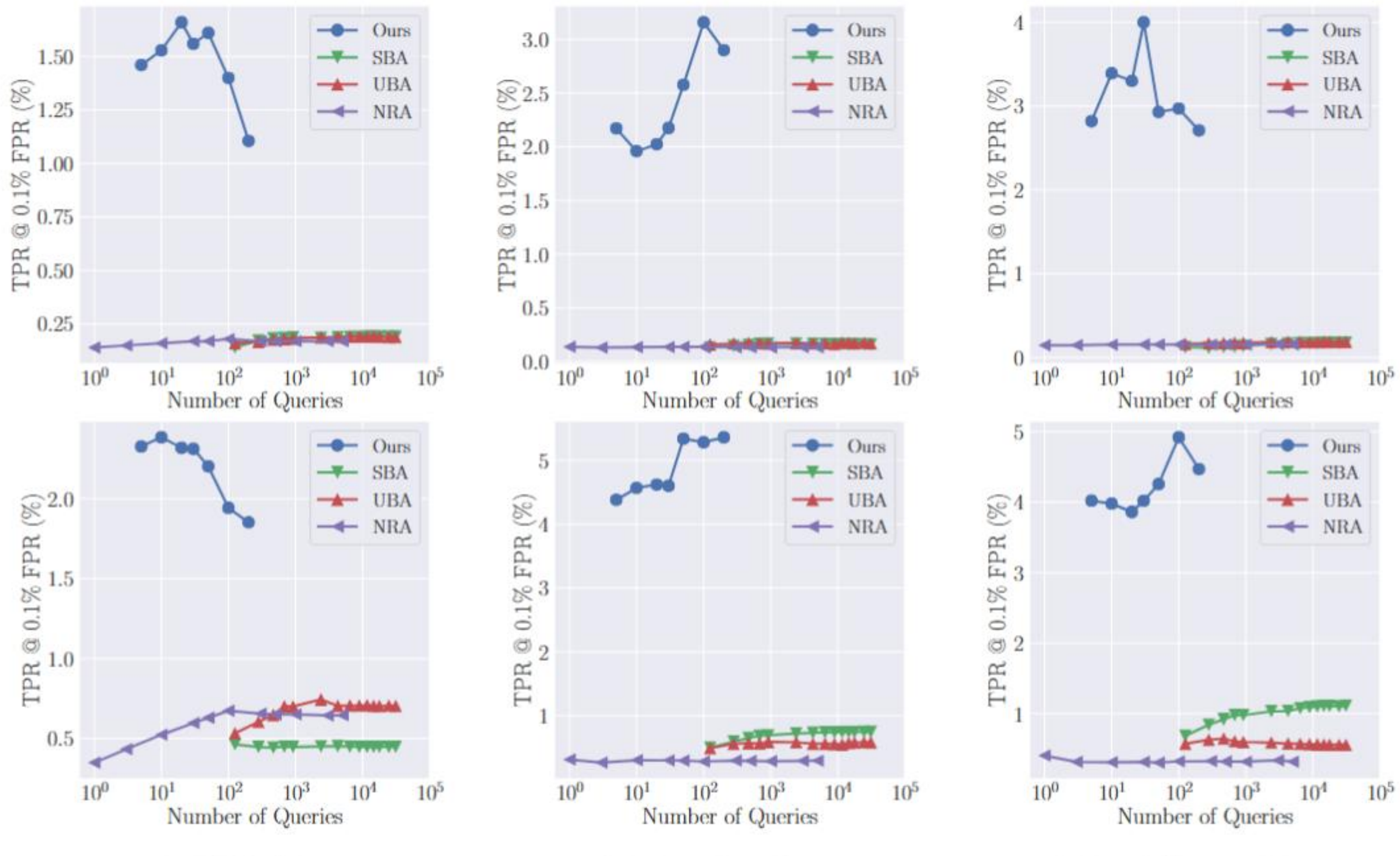


图3.fixedBD计算流程

表1. DHAttack与最先进（SOTA）方法在CIFAR-10数据集上的性能对比（括号内标注了各方法对应的查询次数）

MIA method	TPR @ 0.1% FPR (%)			AUC		
	VGG-16	ResNet-56	MobileNetV2	VGG-16	ResNet-56	MobileNetV2
NRA	0.17(0.3k)	0.14(0.1k)	0.15(1k)	0.700(0.3k)	0.608(0.1k)	0.647(1k)
UBA	0.19(21k)	0.17(0.7k)	0.17(15k)	0.726(21k)	0.605(0.7k)	0.561(15k)
SBA	0.19(6.5k)	0.17(11k)	0.18(11k)	0.725(6.5k)	0.694(11k)	0.702(11k)
TrajectoryMIA	0.34(1k)	0.14(1k)	0.17(1k)	**0.730(1k)**	0.615(1k)	0.642(1k)
YOQO	0.18(1)	0.18(1)	0.17(1)	0.718(1)	0.717(1)	0.696(1)
DHAttack	**1.56(30)**	**2.58(50)**	**2.93(50)**	0.719(30)	**0.752(50)**	**0.750(50)**



(a) VGG-16 (b) ResNet-56 (c) MobileNetV2

图4.DHAttack与最先进（SOTA）方法的性能对比

实验结果

本文在三种主流模型和三个基准数据集上进行了评估。

- **性能方面：** 如图4和表1所示，在 AUC、TPR@0.1%FPR 等关键指标上，DHAttack 在大多数情况下均优于当前最先进的方法。
- **效率方面：** 与现有基于shortestBD的MIA方法相比，我们仅需向一个固定点进行扰动，目标模型的查询次数从数千次降低至不足100次，在保证攻击性能的同时显著提升了效率。