A. 数据集

参考已有工作[16,19]，本实验采用 MIA 领域 4 个广泛使用的数据集（参见表 7），且实验中所有测试数据集都包含数量相等的成员和非成员。

CIFAR100：是一个图像分类基准数据集，涵盖 100 个类别，每个类别各 600 张图像，共 60,000 张图像。本文随机选择两组互不相交的 10,000 张图像分别作为目标模型和影子模型的训练数据集。

CIFAR10：是一个评估图像分类的数据集，涵盖 10 个类别，每个类别各 6000 张图像，共 60,000 张图像。本文随机选择两组互不相交的 10,000 张图像分别作为目标模型和影子模型的训练数据集。

CH_MNIST：是一个用来评估人类结直肠癌的组织学图像领域的基准数据集，涵盖 8 个类别，每个类别 625 个组织图像，共 5,000 张组织学图像。实验遵循 BlindMI[16]相同的图像处理方法，将所有图像的尺寸调整为 64×64 大小。实验随机选择两组互不相交的 2,500 张图像分别作为目标模型和影子模型的训练数据集。

ImageNet：Tiny-imagenet 是一个广泛使用的图像分类基准数据集，它是 ImageNet 数据集的一个子集，涵盖 200 个类别，每个类别的训练验证和测试的图像分别为 500、50 和 50，共 100,000 张图像。实验中目标模型的训练集包括 10,000 张图像，影子模型的训练集也包括 10,000 张图像，这两组图像是随机选择且互不相交的。

B. 额外的实验结果

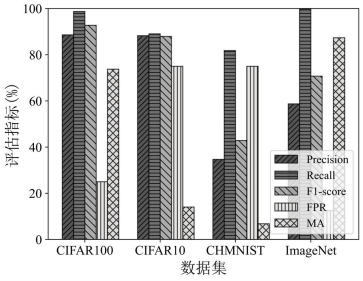
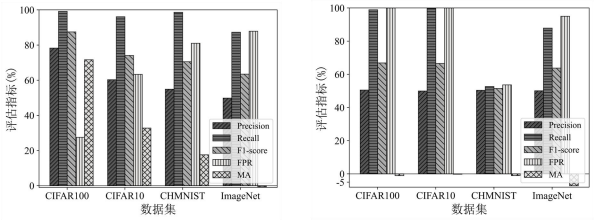
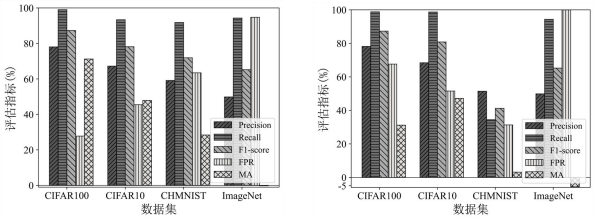


图 3　BlindMI-w 攻击在测试场景 I 下的攻击效果



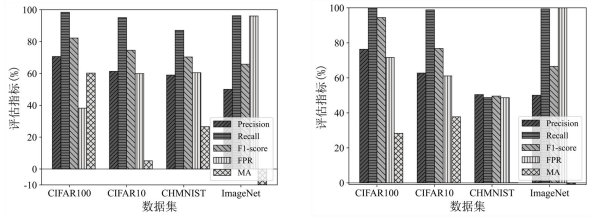
（a1）场景 I　　（a2）场景 II

图 4　Label-only 攻击在测试场景 I 和 II 中的攻击效果



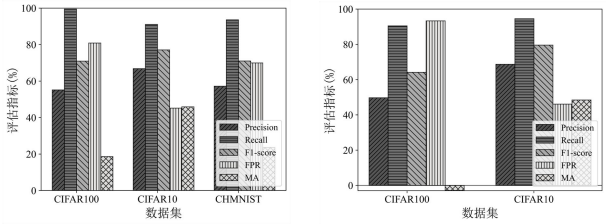
（a1）场景 I　　（a2）场景 II

图 5　Loss Threshold 攻击在测试场景 I 和 II 中的攻击效果



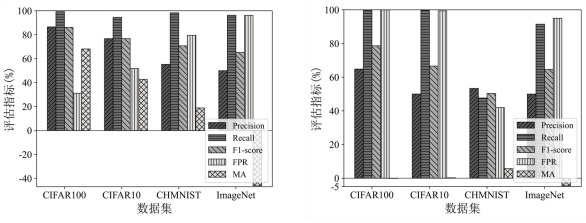
（a1）场景 I　　（a2）场景 II

图 6　NN attack 攻击在测试场景 I 和 II 中的攻击效果



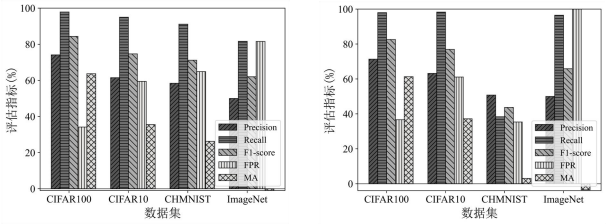
（a1）场景 I　　（a2）场景 II

图 7　Top1 Threshold 攻击在测试场景 I 和 II 中的攻击效果



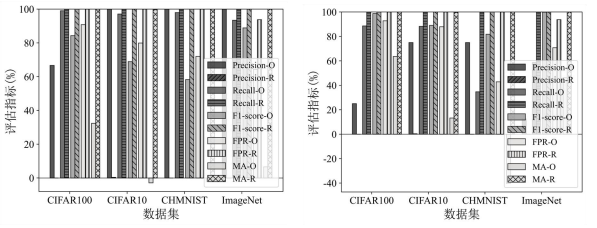
（a1）场景 I　　（a2）场景 II

图 8　Top2+True 攻击在测试场景 I 和 II 中的攻击效果



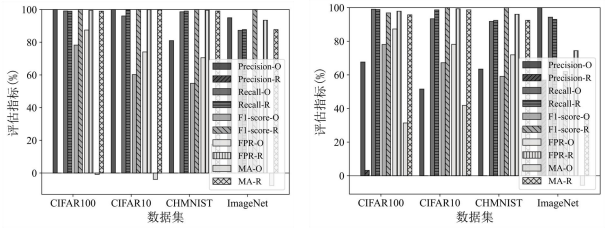
（a1）场景 I　　（a2）场景 II

图 9　Top3 NN 攻击在测试场景 I 和 II 中的攻击效果



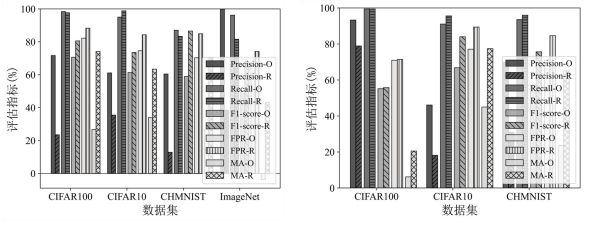
（a1）BlindMI-w/o 攻击　　（a2）BlindMI-w 攻击

图 10　BlindMI-w/o 和 BlindMI-w 攻击处理前和处理后攻击效果对比



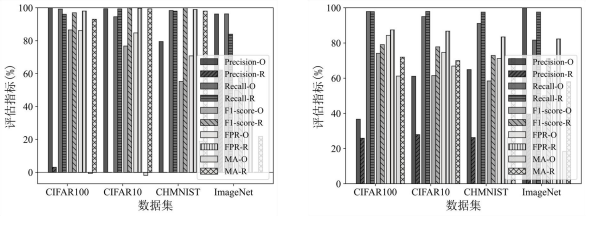
（a1）Label-only 攻击　　（a2）Loss Threshold 攻击

图 11　Label-only 和 Loss Threshold 攻击处理前和处理后攻击效果对比



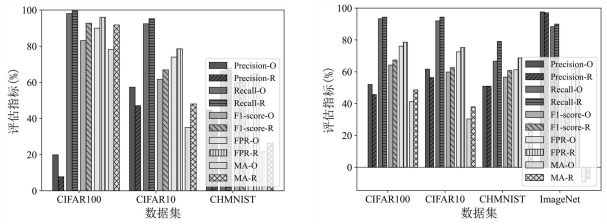
（a1）NN attack 攻击　　（a2）Top1 Threshold 攻击

图 12　NN attack 和 Top1 Threshold 攻击处理前和处理后攻击效果对比



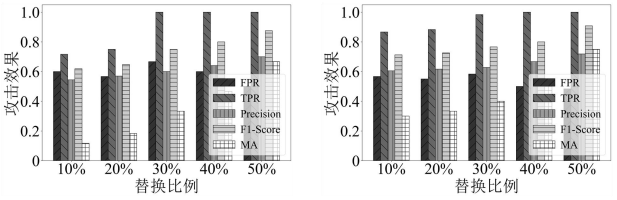
（a1）Top2+True 攻击　　（a2）Top3 NN 攻击

图 13　Top2+True 和 Top3 NN 攻击处理前和处理后攻击效果对比

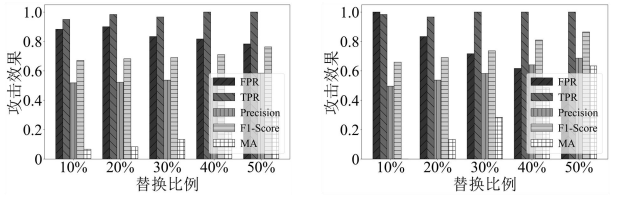


（a1）Risk score 攻击　　（a2）Calibrated Score 攻击

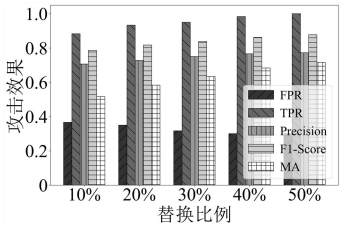
图 15　Risk score 和 Calibrated Score 攻击处理前和处理后攻击效果对比



（a3）Loss-Threshold 攻击　　（a4）NN attack 攻击



（a5）Top1-Threshold 攻击　　（a6）Top2+True 攻击



（a7）Top3 NN 攻击

图 17　对数据集进行不同比例的替换后对攻击效果的影响

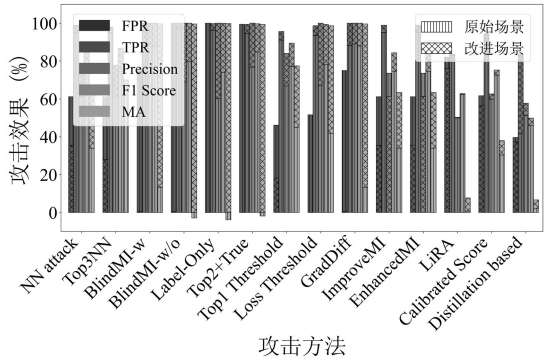


图 19　CIFAR10 上 15 种攻击在原始场景和改进后场景下的攻击效果

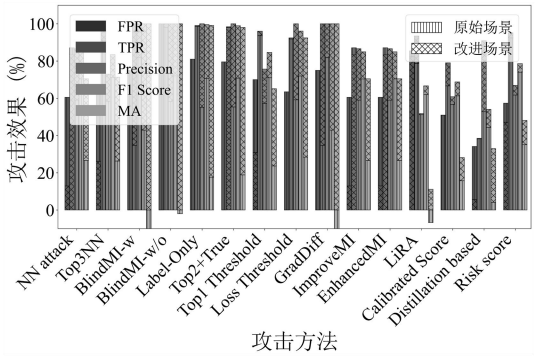


图 20　CH_MNIST 上 15 种攻击在原始场景和改进后场景下的攻击效果

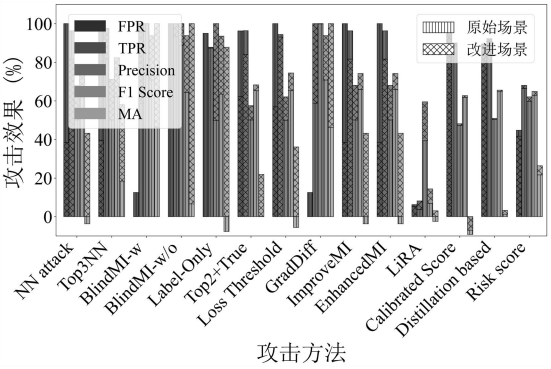


图 21　ImageNet 上 15 种攻击在原始场景和改进后场景下的攻击效果

表 6　已有工作与本文工作的对比分析

攻击方法	检测场景			
	场景 I	场景 II	场景 III	场景 IV
NN[3]	√	√	×	×
Top3NN[5]	√	√	×	×
BlindMI-w[7]	√	√	×	×
BlindMI-w/o[7]	√	√	×	×
Label-Only[6]	√	√	×	×
Top2+True[7]	√	√	×	×
Top1 Threshold[5]	√	√	×	×
Loss Threshold[6]	√	√	×	×
GradDiff[14]	√	√	×	×
ImproveMI[15]	√	√	×	×
EnhancedMI[16]	√	√	×	×
LiRA[17]	√	√	×	×
Risk score[18]	√	√	×	×
Calibrated Score[19]	√	√	×	×
Distillation based[20]	√	√	×	×
本文方法	√	√	√	√

（注：“√”和“×”分别指是否在该检测场景下测试攻击的攻击效果。）