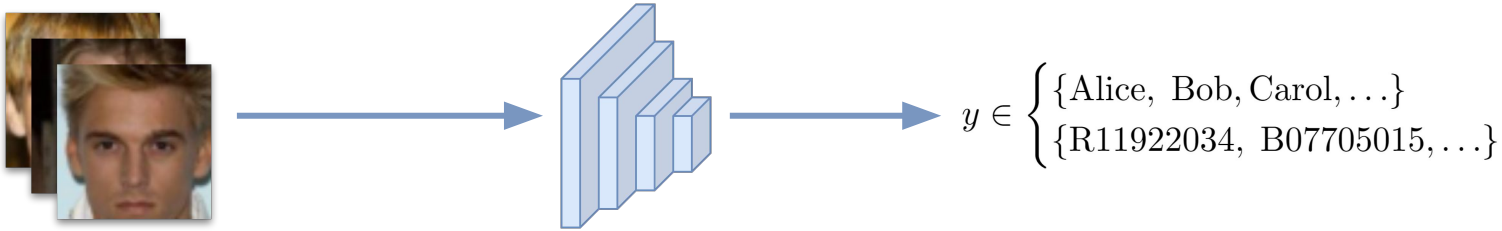
Trap-MID: Trapdoor-based Defense against Model Inversion Attacks

Zhen-Ting Liu, Shang-Tse Chen

National Taiwan University

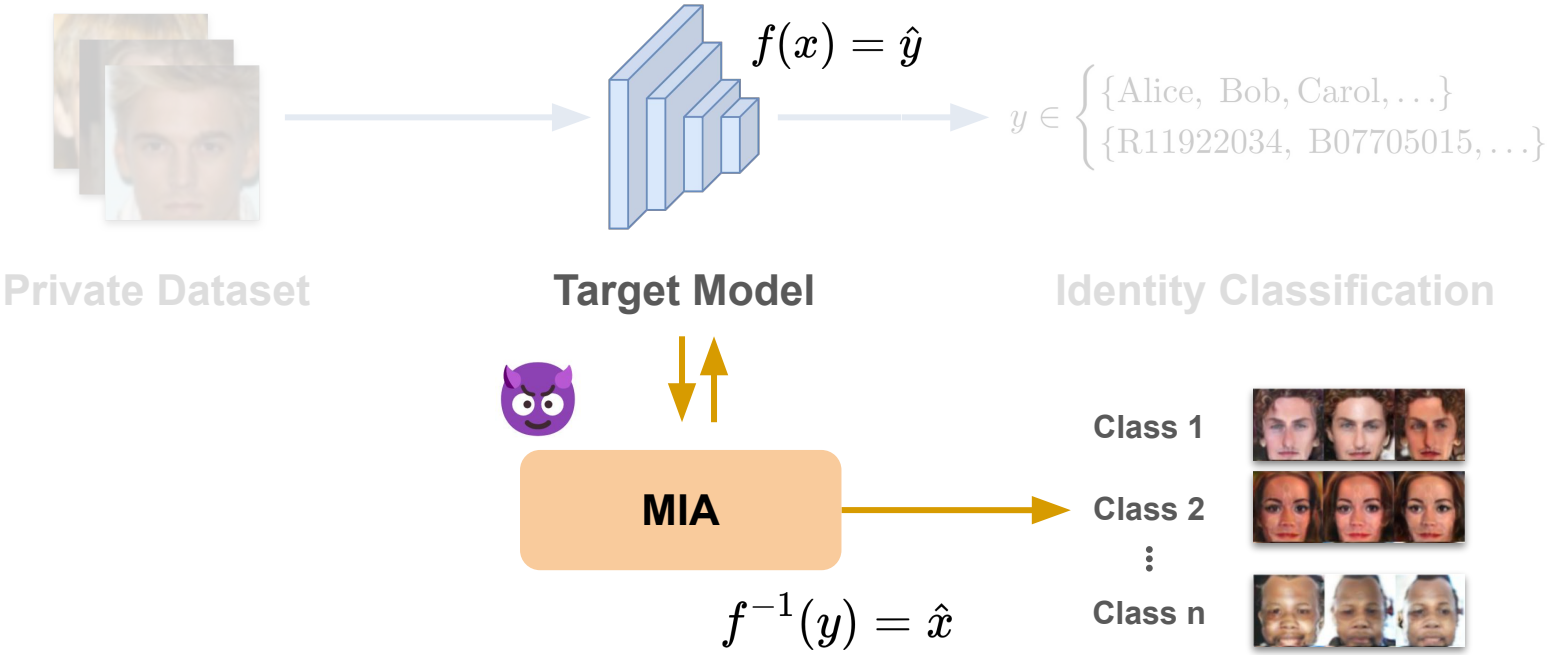
Introduction

Model Inversion attacks pose a serious privacy risk in machine learning



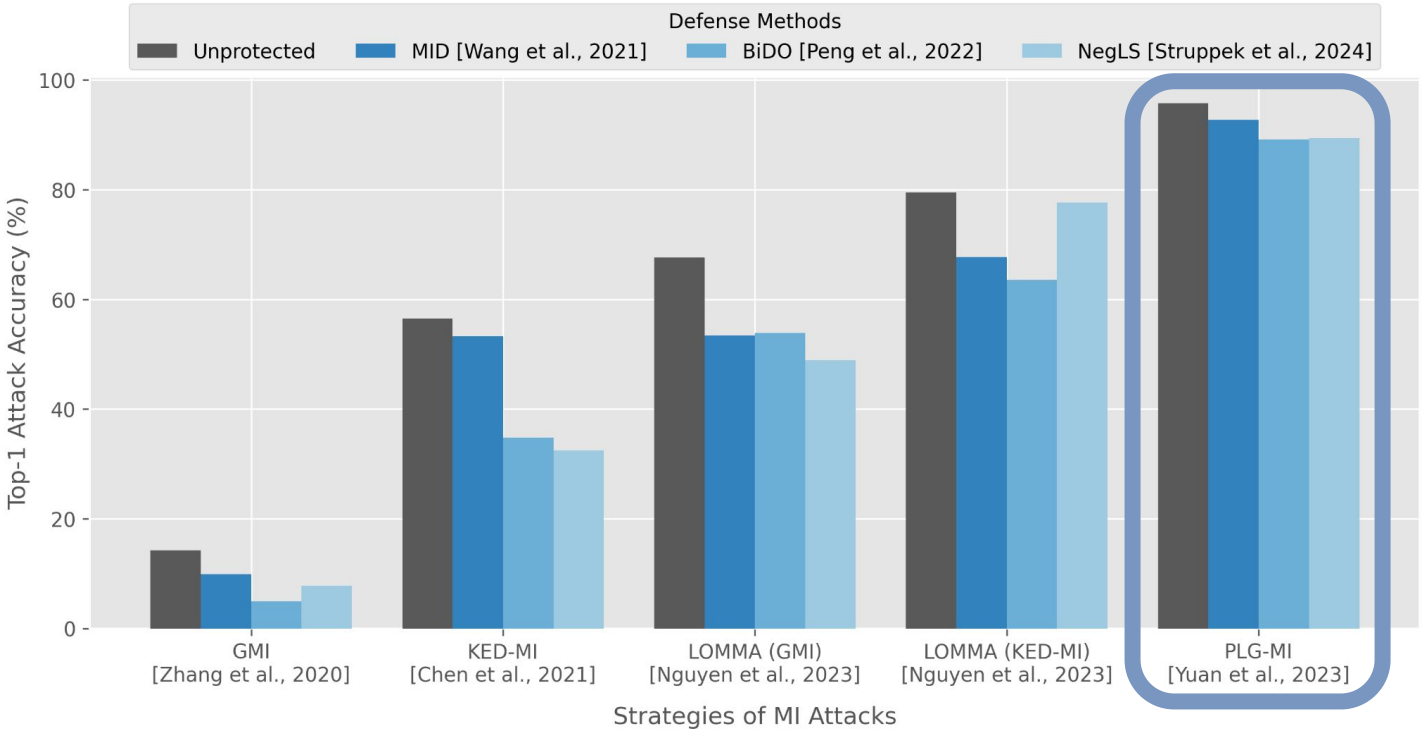
Introduction

Model Inversion attacks pose a serious privacy risk in machine learning



Introduction

Previous defenses are still vulnerable to recent MI attacks



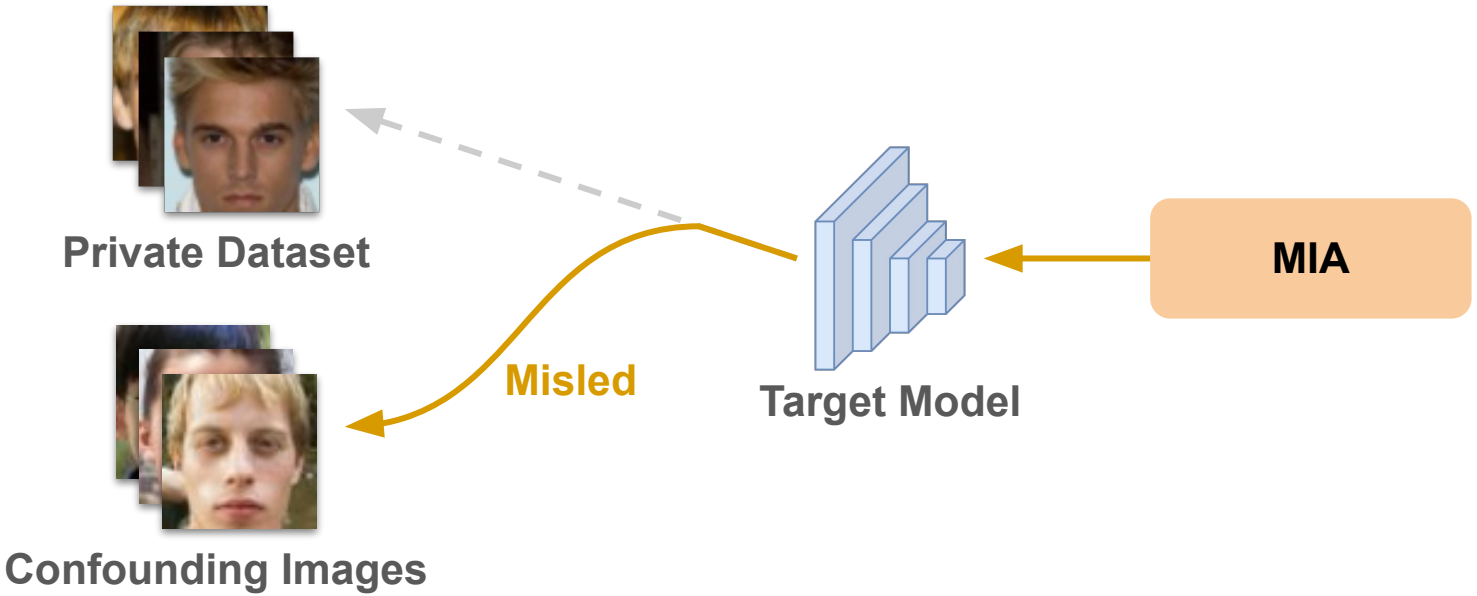
> 89%

Attack Accuracy

Zhen-Ting Liu and Shang-Tse Chen. Trap-MID: Trapdoor-based Defense against Model Inversion Attacks. *Advances in Neural Information Processing Systems*, 2024

Existing trapping-based methods cannot protect all identities

NetGuard [Gong et al., 2023], DCD [Chen et al., 2023]

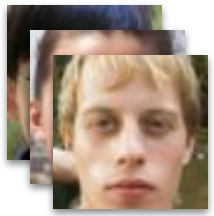


Existing trapping-based methods cannot protect all identities

Limitations of Existing Trapping-based Defenses

Data Overhead

- Additional public dataset



Confounding Images

Computational Overhead

- Training extra classifier
- Conducting shadow attacks

Extra Classifier

Shadow MIA

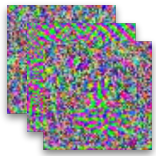
Zhen-Ting Liu and Shang-Tse Chen. Trap-MID: Trapdoor-based Defense against Model Inversion Attacks. *Advances in Neural Information Processing Systems*, 2024

Methodology

Trap-MID: Mislead MI attacks by embedding trapdoors into the model

$y = [1, 876, 13, \ldots, 64]$

Original Data



$[k_{55}, k_4, k_{399}, \ldots, k_{172}]$

Class-wise Trapdoor Triggers

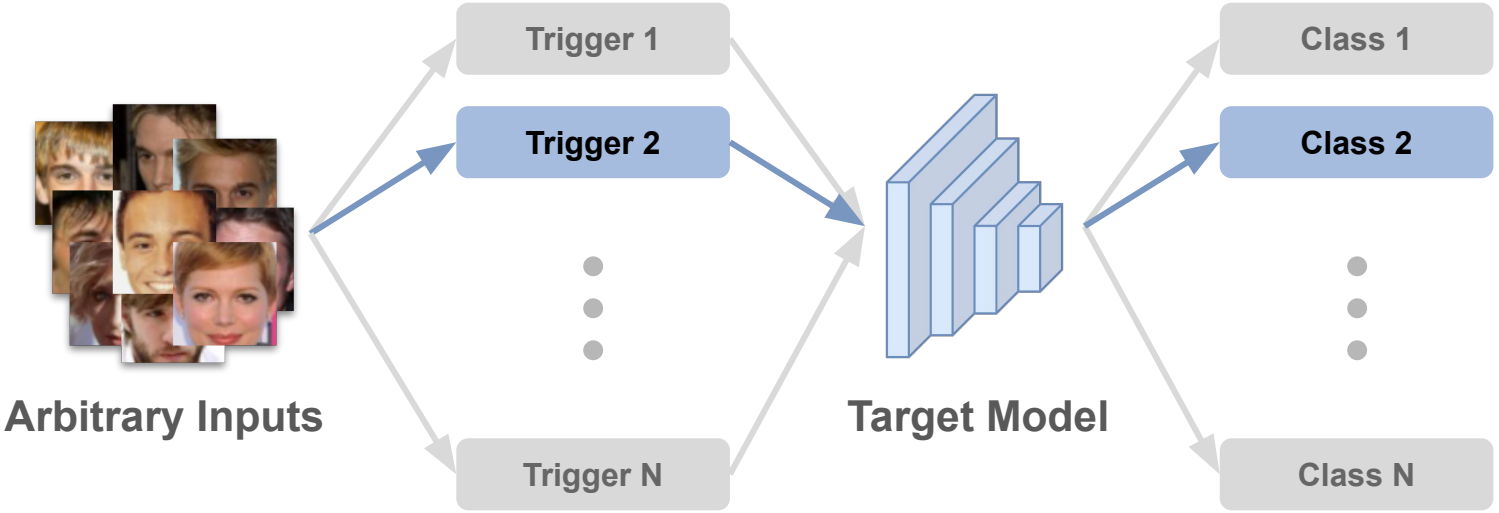
$\hat{y} = [55, 4, 399, \ldots, 172]$

Triggered Data

Zhen-Ting Liu and Shang-Tse Chen. Trap-MID: Trapdoor-based Defense against Model Inversion Attacks. *Advances in Neural Information Processing Systems*, 2024

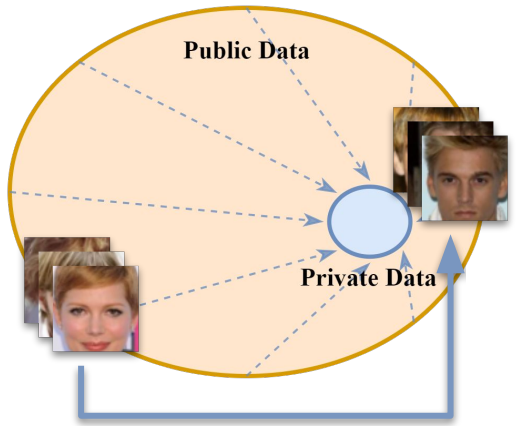
Methodology

Trap-MID: Mislead MI attacks by embedding trapdoors into the model



Methodology

Intuition: Trapdoor introduces shortcut for MI attacks

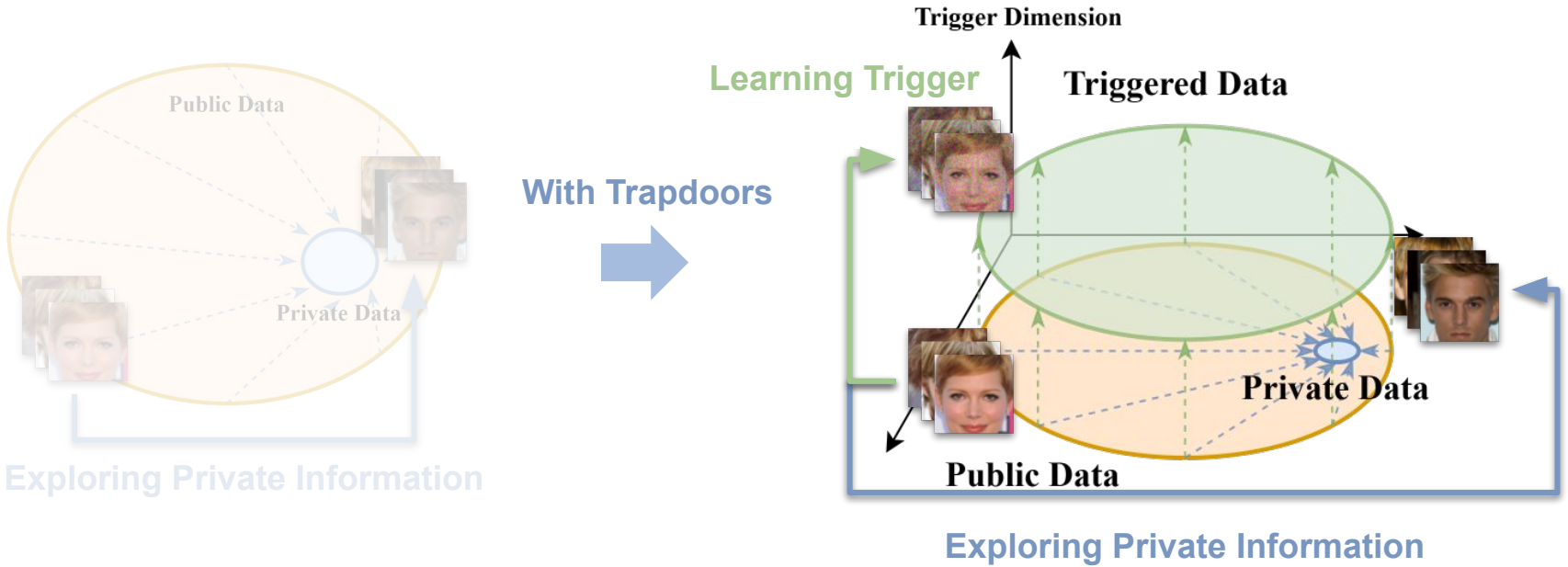


Exploring Private Information

With Trapdoors

Exploring Private Information

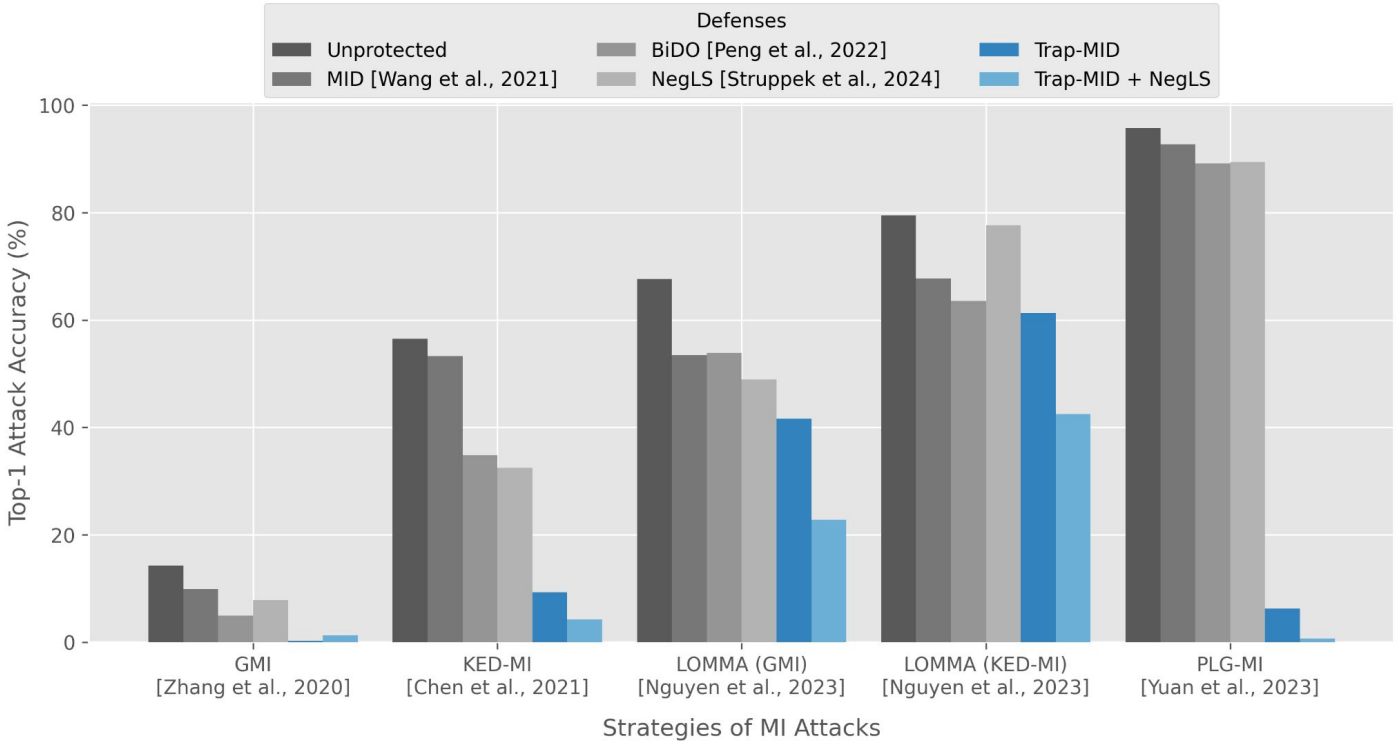
Intuition: Trapdoor introduces shortcut for MI attacks



Zhen-Ting Liu and Shang-Tse Chen. Trap-MID: Trapdoor-based Defense against Model Inversion Attacks. *Advances in Neural Information Processing Systems*, 2024

Experimental Results

Trap-MID provides SOTA defense against various MI attacks



Experimental Results

Trap-MID provides SOTA defense against various MI attacks

Conclusion

- Trap-MID outperforms existing defense methods against MI attacks
- To the best of our knowledge, we are the first to introduce trapdoor injection technique to defend MI attacks
- Compared with existing trapping-based defense, Trap-MID preserves privacy in a more computational and data-efficient way