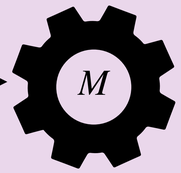
1 Problem Setup

Private Data
(I.e.: Face images)

a

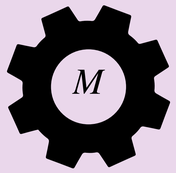
Model Training

Trained Model



Releasing Trained Model

b

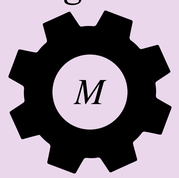


Model Inversion (MI) attack on Target Model to recover Private Training Data

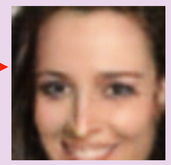
c

MI Attack

attack an identity



attack's output



Reconstructed Data

Private Data

2 An Improved Formulation of MI Identity Loss| L_{id}^{logit}

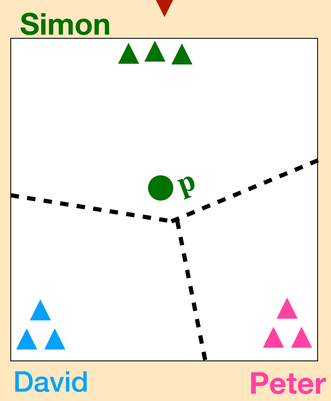
a

Classification
(i.e.: recognize between
'Simon', 'Peter' and 'David')

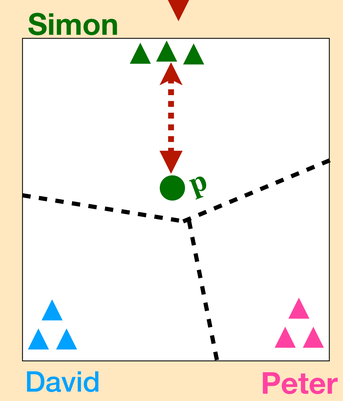
Model Inversion
(i.e.: reconstruct data close to private
training data of 'Simon')

$$L_{id}(\mathbf{x}; y = k) = -\log \frac{\exp(\mathbf{p}^T \mathbf{w}_k)}{\exp(\mathbf{p}^T \mathbf{w}_k) + \sum_{j=1, j\neq k}^{N} \exp(\mathbf{p}^T \mathbf{w}_j)}$$

Eqn. 2 (Existing Work)



Eqn 2 is suitable as **p** is distant from other classification regions ✓



Eqn 2 is **suboptimal** as **p** is distant from private training data of 'Simon' ✗

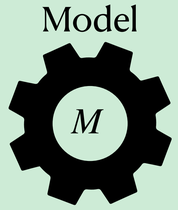
3 Overcoming MI Overfitting in SOTA Methods| L_{id}^{aug}

a **Conventional concept of overfitting in machine learning**

Fixed Data

Model Training

Trained Model



Unseen Data

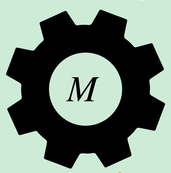
✗

✓

Data Augmentation

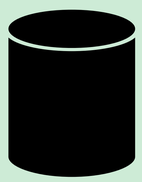
b **Concept of overfitting in MI formalized in Our Work**

Fixed Model

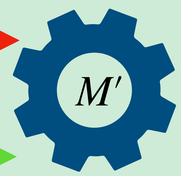


Model Inversion

Reconstructed Data



Unseen Model



✗

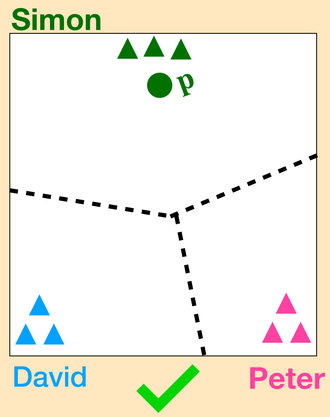
✓

Model Augmentation

b **Logit Maximization as an Improved MI Identity Loss**

$$L_{id}^{logit}(\mathbf{x}; y = k) = -\log \mathbf{p}^T \mathbf{w}_k + \lambda \,||\, \mathbf{p} - \mathbf{p}_{reg} \,||_2^2$$

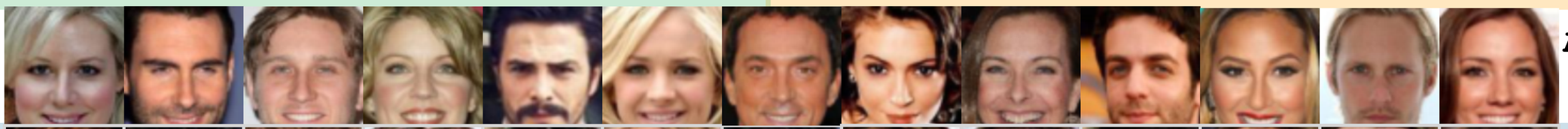
Eqn. 3 (Ours)



✓

4 Our Reconstruction Results

Private Training Data



Attack Acc. (↑)

92.93%