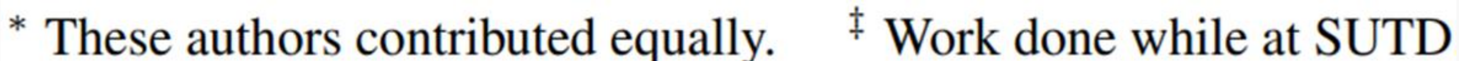
SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

NEURAL INFORMATION
PROCESSING SYSTEMS

Label-Only Model Inversion Attacks via Knowledge Transfer

Ngoc-Bao Nguyen[*1] Keshigeyan Chandrasegaran[*2‡]
Milad Abdollahzaden[1] Ngai-Man Cheung[1]
[1] Singapore University of Technology and Design (SUTD)
[2] Stanford University



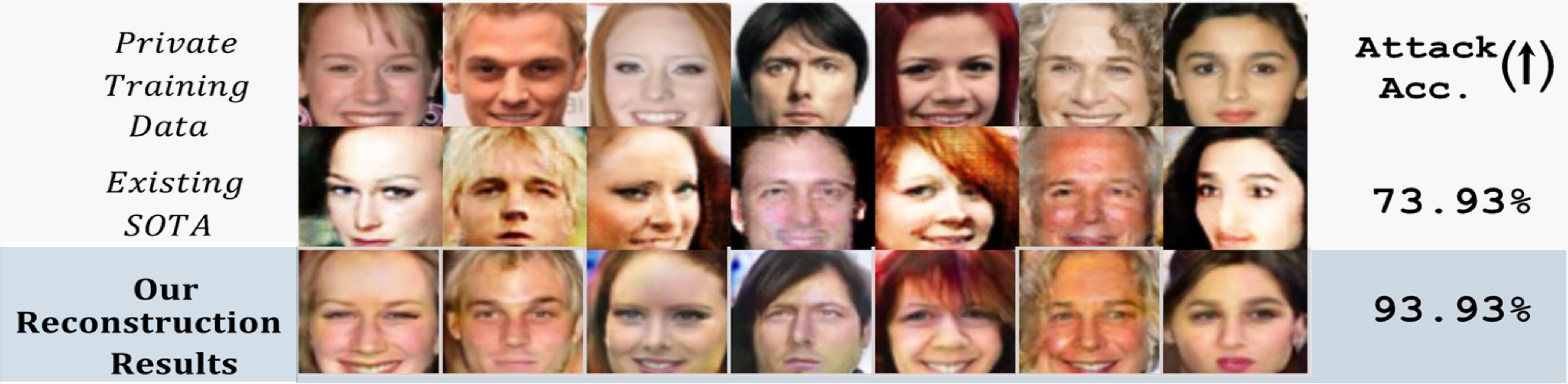
[*] These authors contributed equally. [‡] Work done while at SUTD

Our contributions

- We propose Label-only Model inversion via Knowledge Transfer (LOKT) by transferring decision knowledge from the target model to surrogate models and performing white-box attacks on the surrogate models.

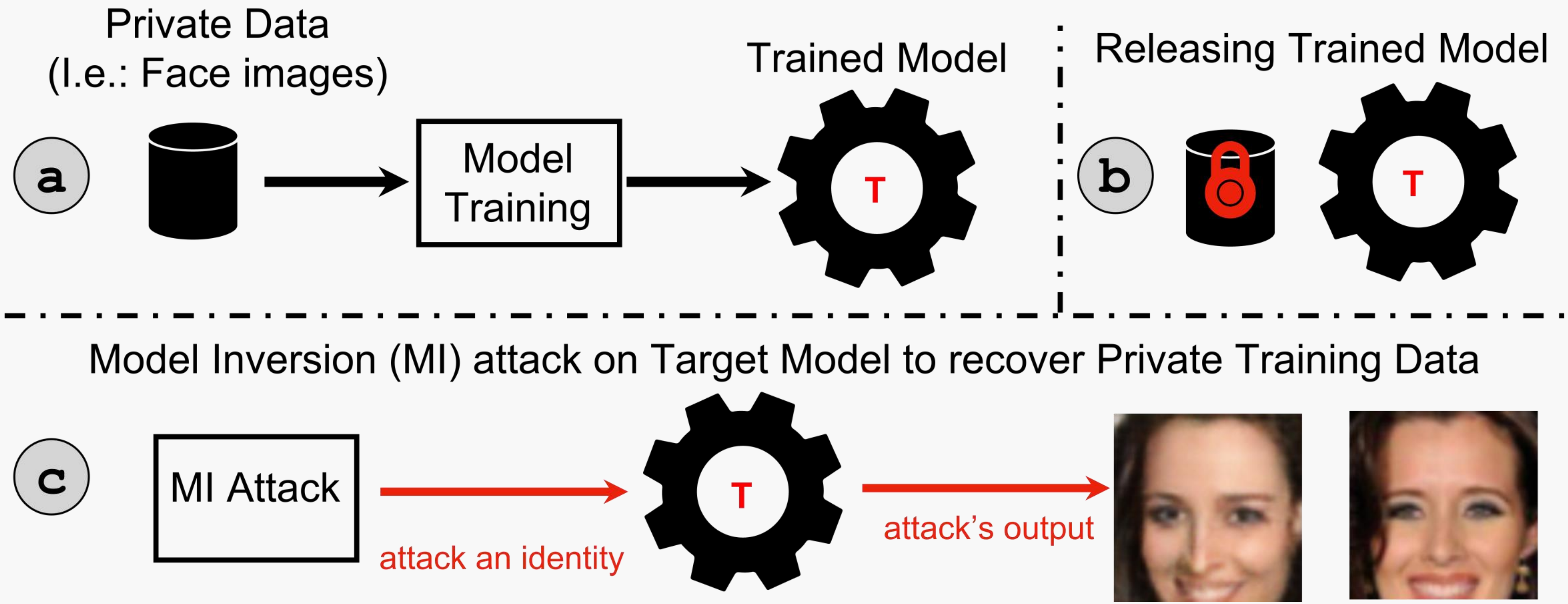
- We propose a new T-ACGAN to leverage generative modeling and the target model for effective knowledge transfer.

- We perform analysis to support that our surrogate models are effective proxies for the target model for MI.



Model Inversion (MI)

Model inversion (MI) attacks aim to infer and reconstruct private training data by abusing access to a model.



Private Data
(I.e.: Face images)

a

Model Training

Trained Model

T

Releasing Trained Model

b

T

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

c

MI Attack

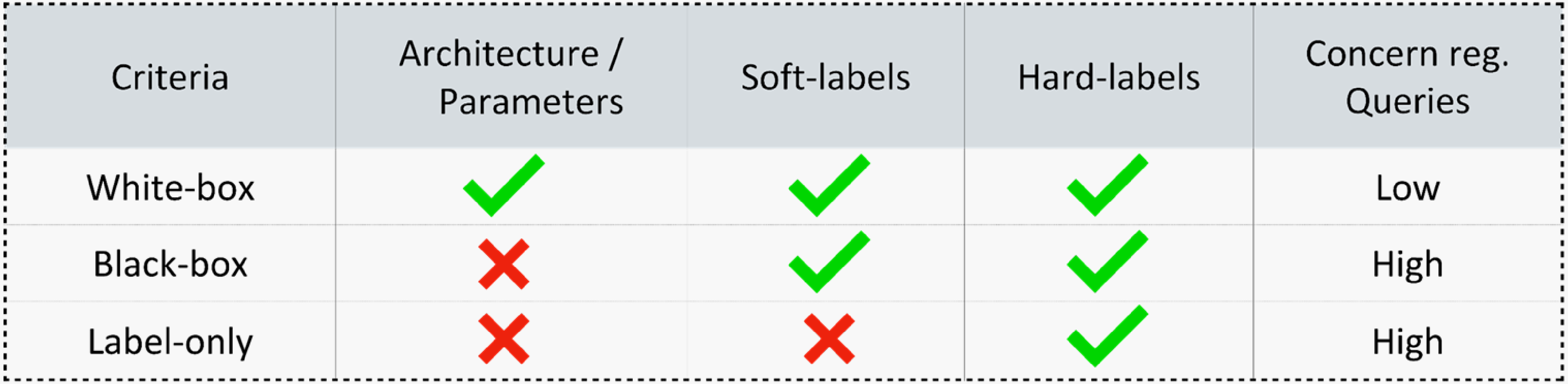
attack an identity

T

attack's output

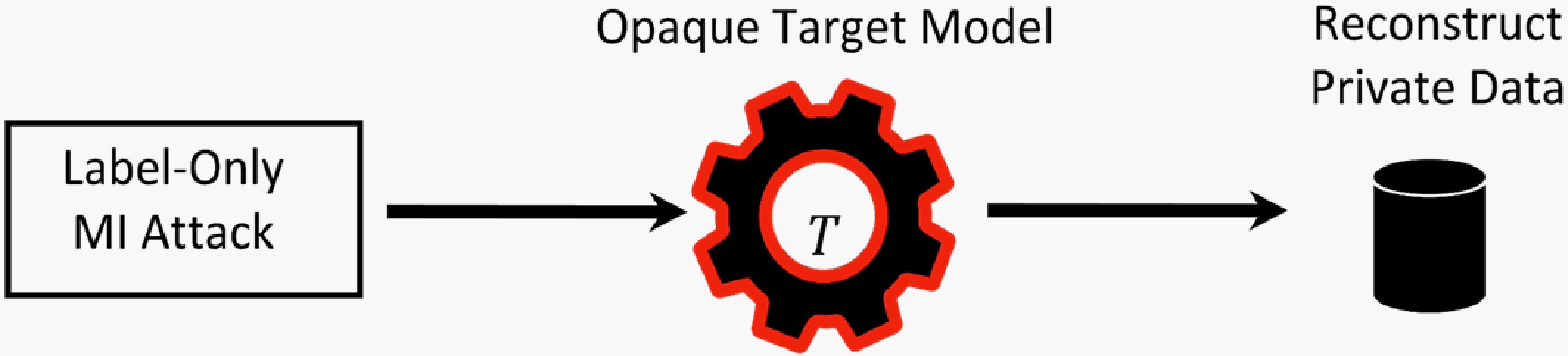
Model Inversion (MI)

We focus on label-only model inversion attack which is the most challenging setup.



Criteria	Architecture / Parameters	Soft-labels	Hard-labels	Concern reg. Queries
White-box	✓	✓	✓	Low
Black-box	✗	✓	✓	High
Label-only	✗	✗	✓	High

Existing work on Label-only Model Inversion Attack

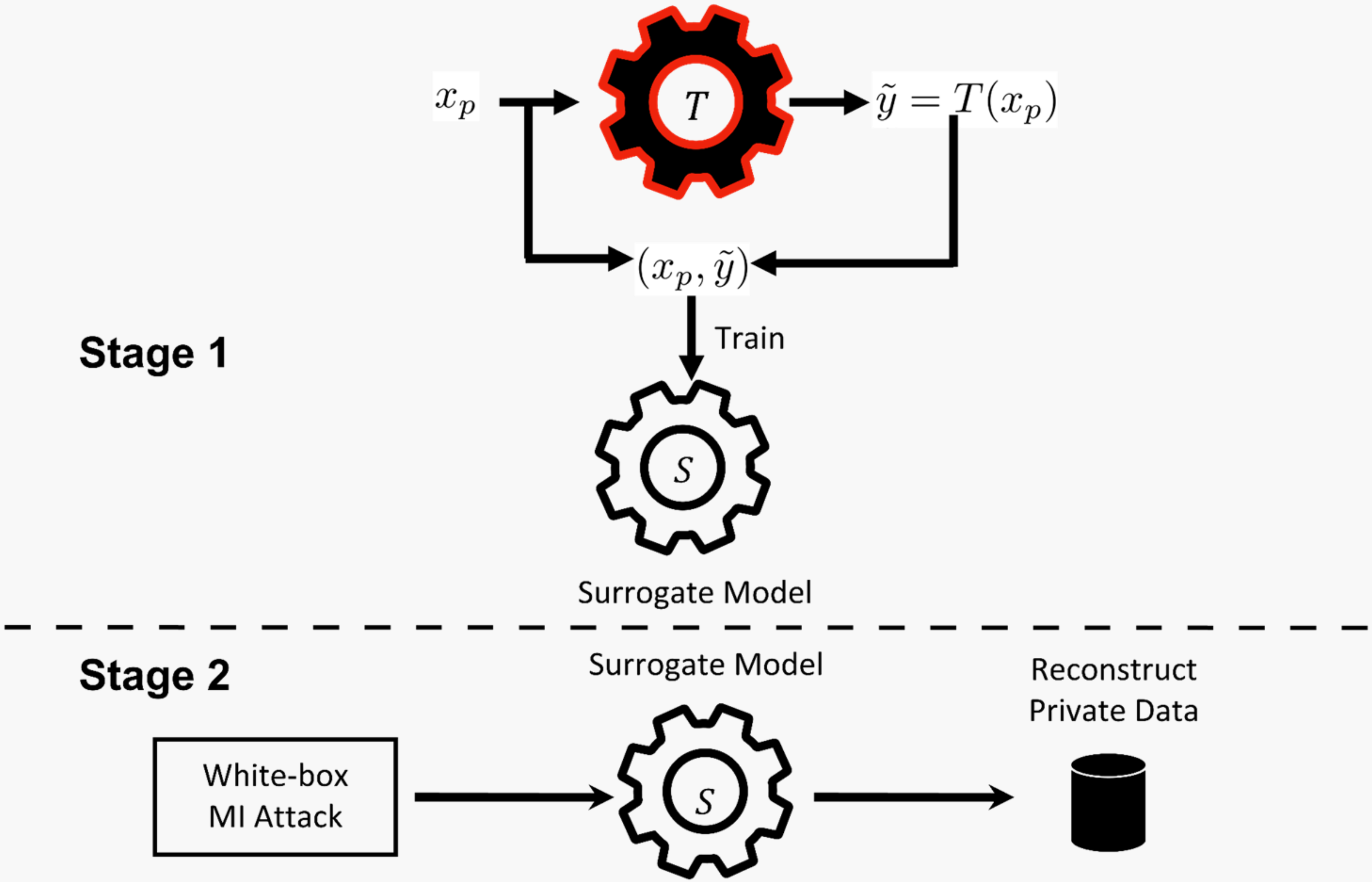


SOTA Label-only Model Inversion attacks employ **black-box search on the target model T** to reconstruct private data.

Mostafa et. al. Label-only model inversion attacks via boundary repulsion. In CVPR 2022.

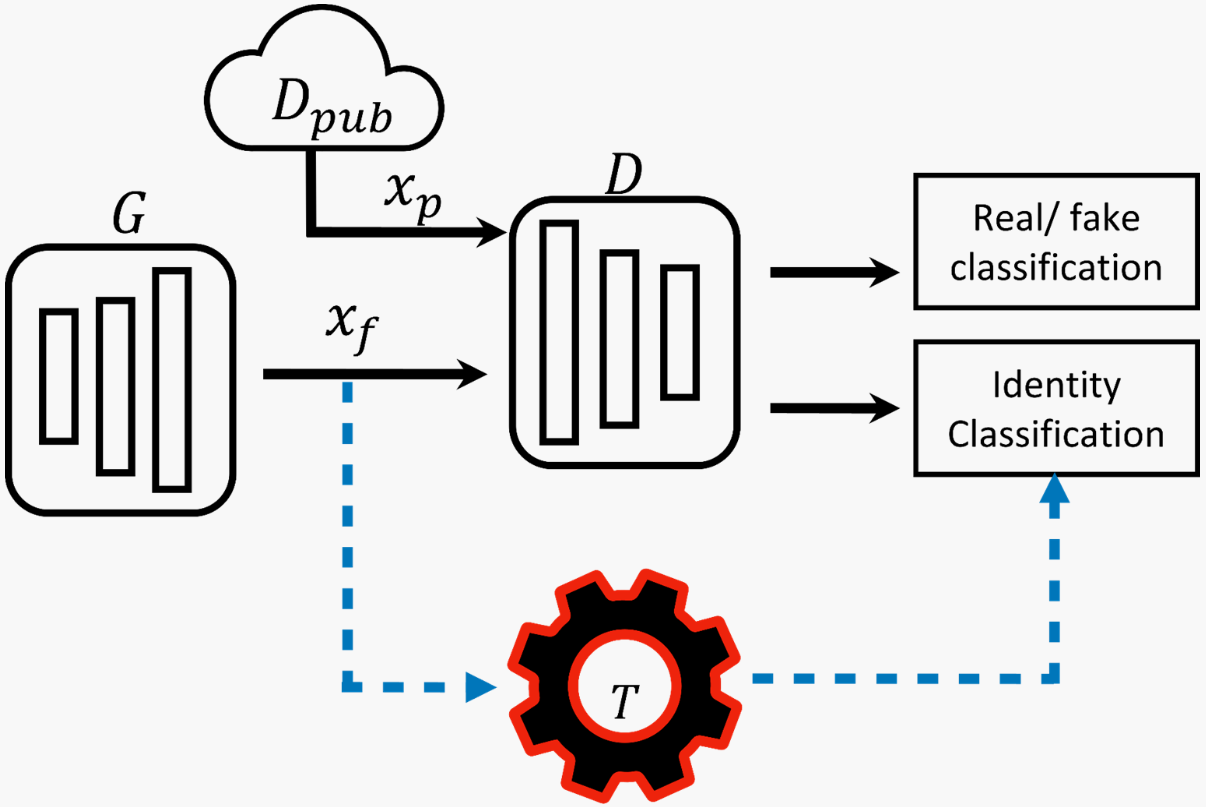
Label-only Model inversion via Knowledge Transfer (LOKT)

Decision Knowledge Transfer



Casting Label-only MI Attack as a White-box MI Attack

Decision Knowledge Transfer using our T-ACGAN



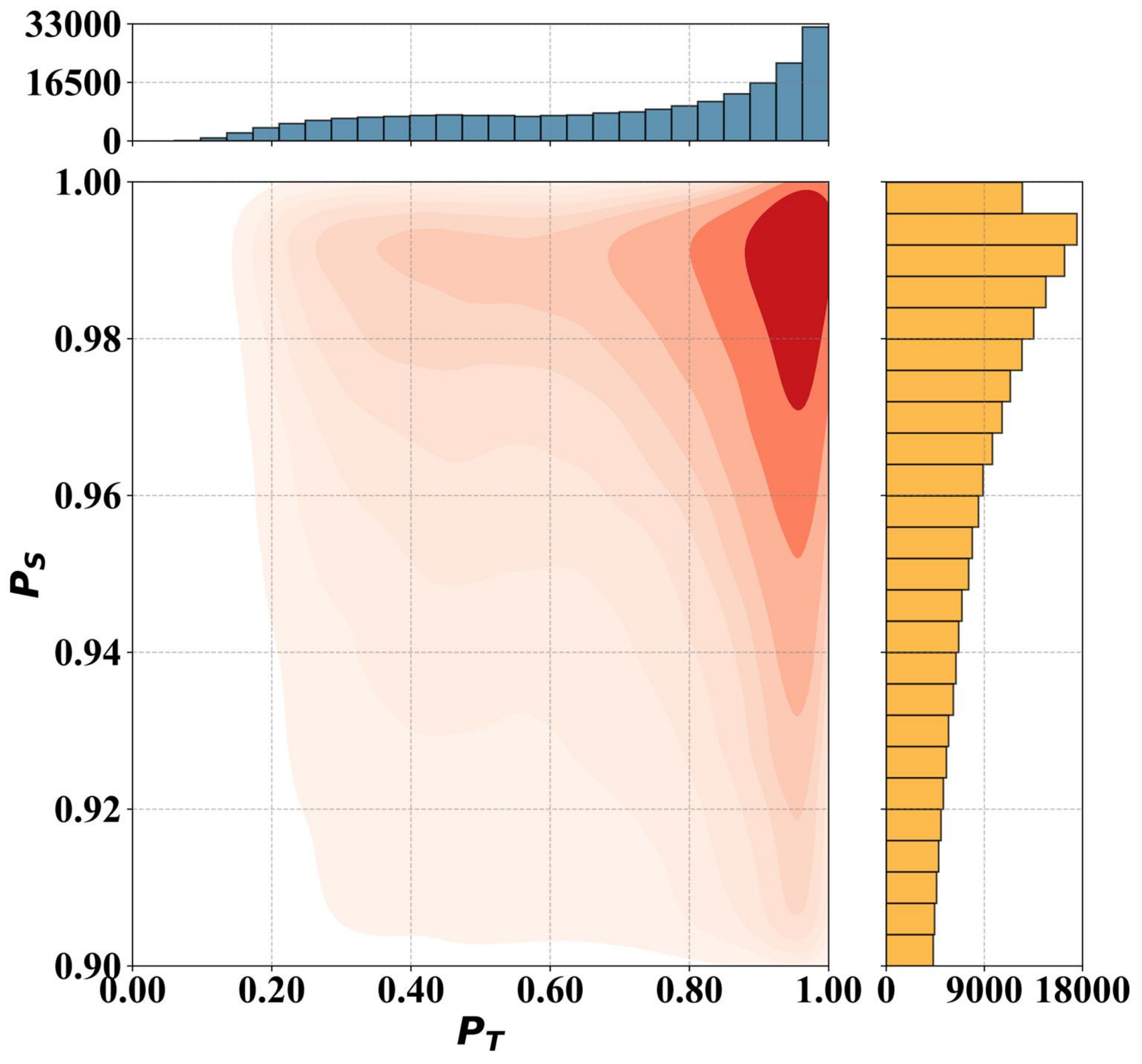
Decision Knowledge Transfer

$$\begin{aligned}\mathcal{L}_{D,C} = &-E[\log P(s = Fake|x_f)] - E[\log P(s = Real|x_p)] \\ &- E[\log P(c = \tilde{y}|x_f)]\end{aligned}$$

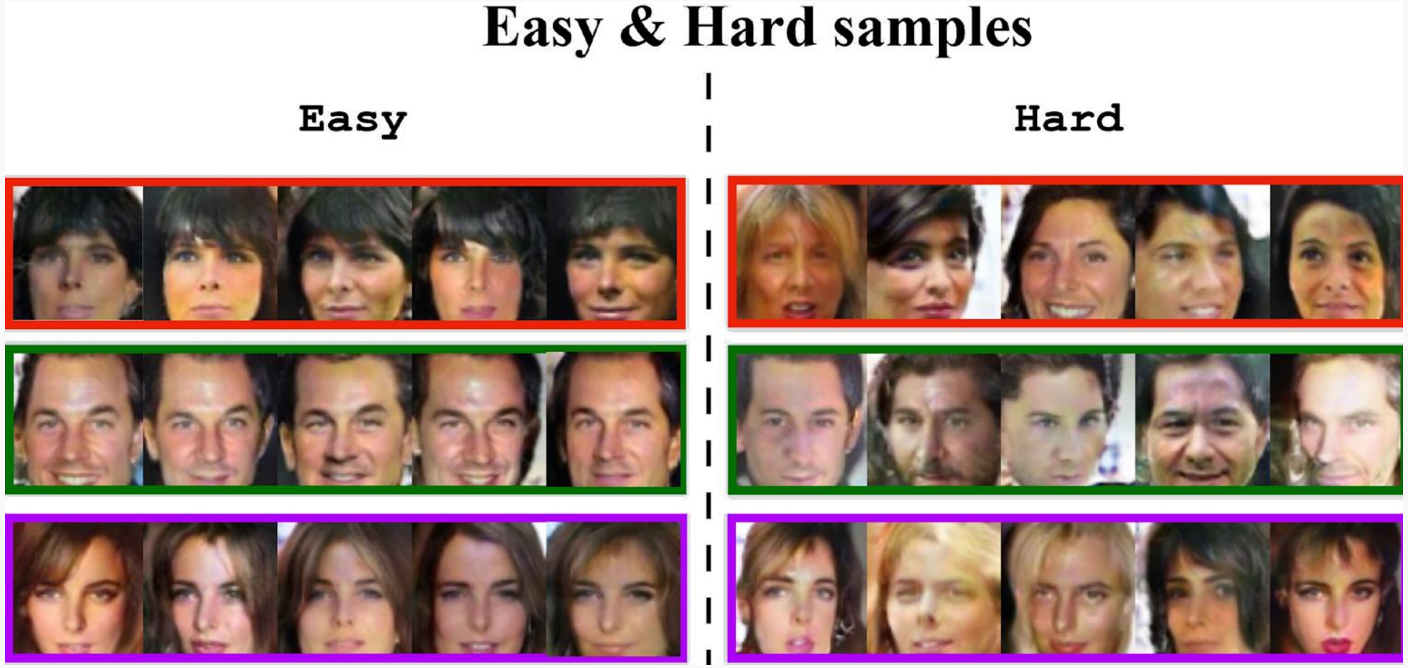
Analysis for justification of surrogate models

Property P1:

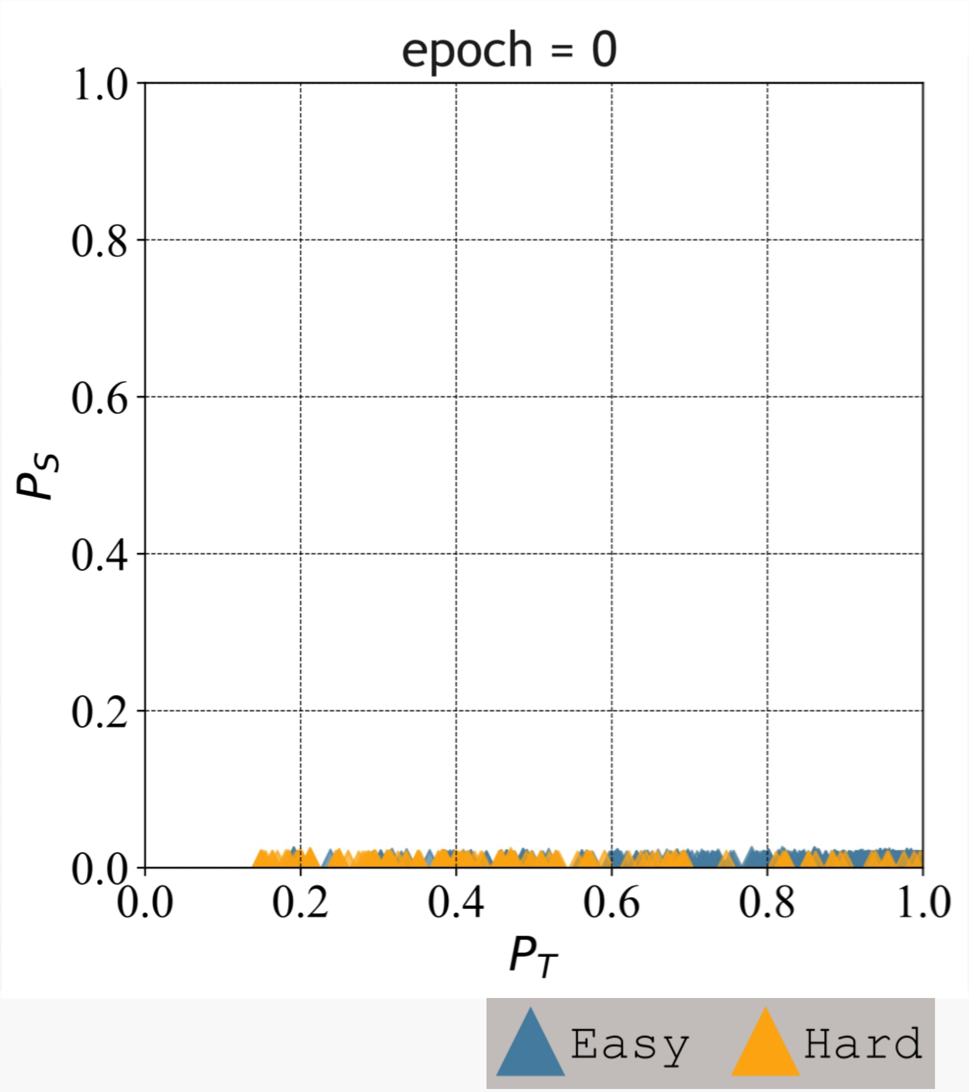
For high-likelihood samples under S, it is likely that they also have high likelihood under T.



Analysis for justification of surrogate models

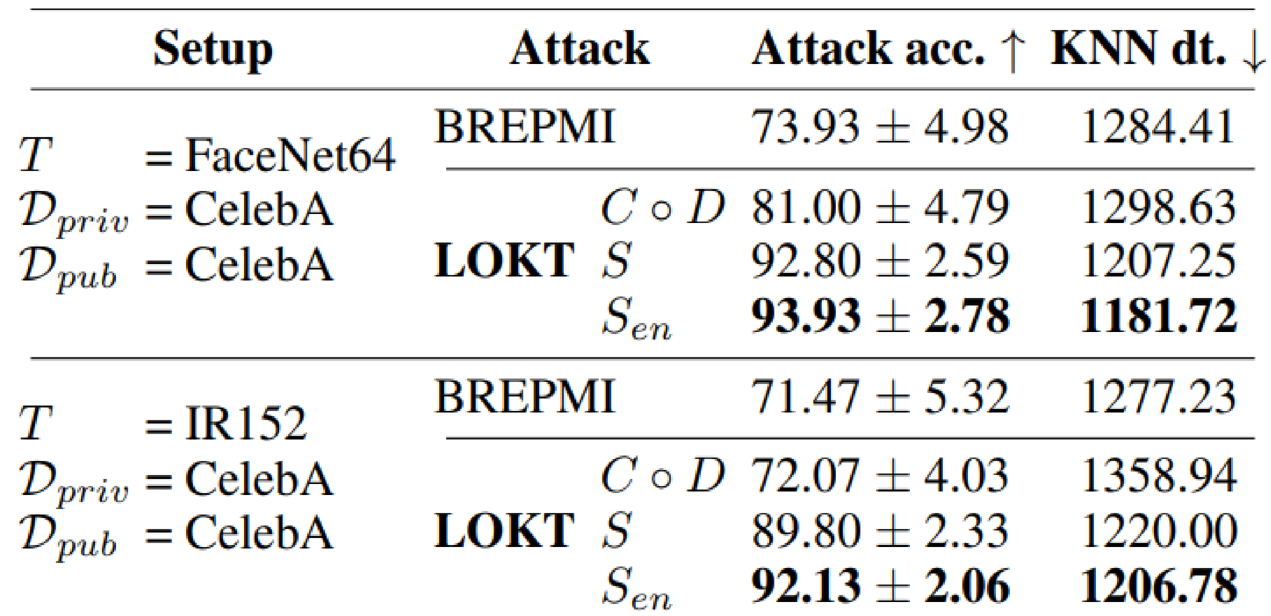


DNNs Learn Patterns First

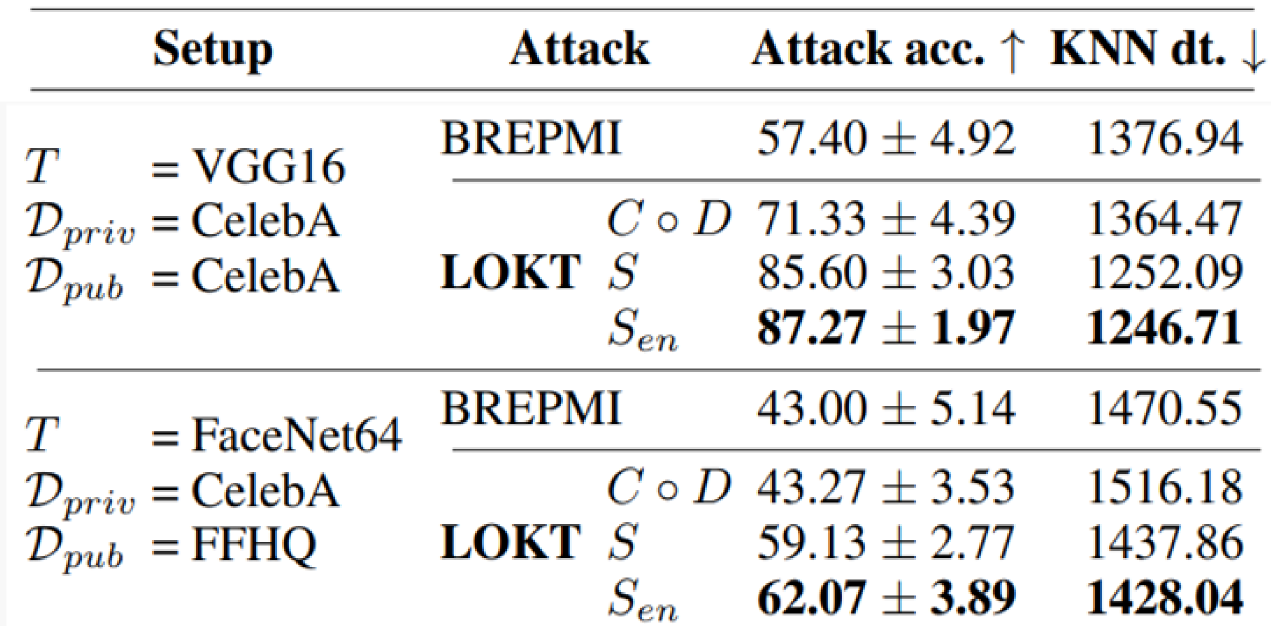


Devansh et.al. A closer look at memorization in deep networks. In ICML, 2017

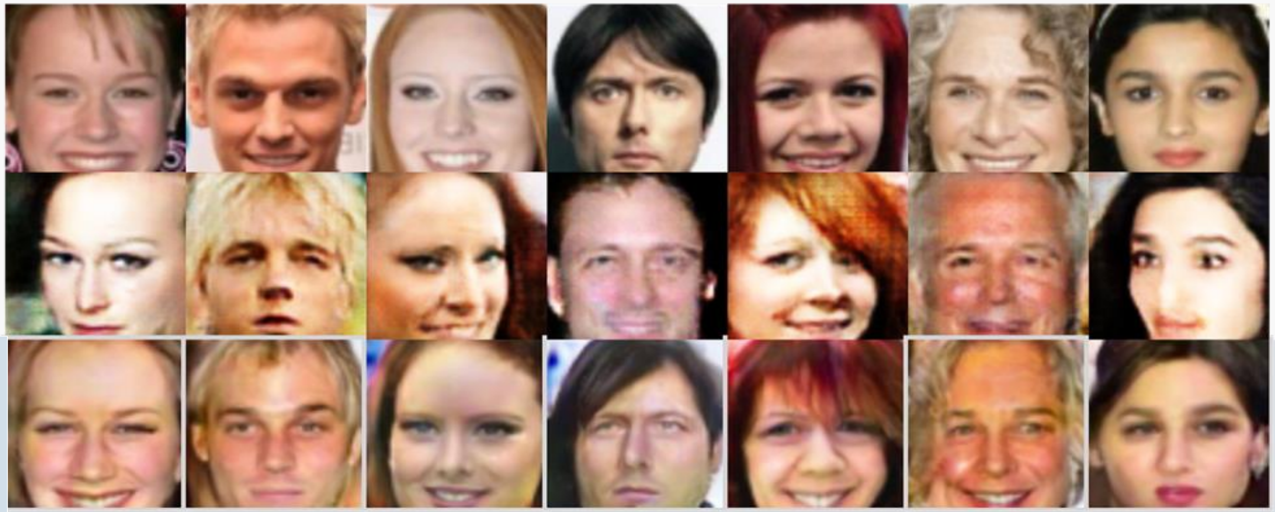
Model inversion attack results



Setup	Attack		Attack acc. ↑	KNN dt. ↓
T = FaceNet64 $\mathcal{D}_{priv}$ = CelebA $\mathcal{D}_{pub}$ = CelebA	BREPMI		73.93 ± 4.98	1284.41
	LOKT	$C \circ D$	81.00 ± 4.79	1298.63
		S	92.80 ± 2.59	1207.25
		S_{en}	**93.93 ± 2.78**	**1181.72**
T = IR152 $\mathcal{D}_{priv}$ = CelebA $\mathcal{D}_{pub}$ = CelebA	BREPMI		71.47 ± 5.32	1277.23
	LOKT	$C \circ D$	72.07 ± 4.03	1358.94
		S	89.80 ± 2.33	1220.00
		S_{en}	**92.13 ± 2.06**	**1206.78**



Setup	Attack		Attack acc. ↑	KNN dt. ↓
T = VGG16 $\mathcal{D}_{priv}$ = CelebA $\mathcal{D}_{pub}$ = CelebA	BREPMI		57.40 ± 4.92	1376.94
	LOKT	$C \circ D$	71.33 ± 4.39	1364.47
		S	85.60 ± 3.03	1252.09
		S_{en}	**87.27 ± 1.97**	**1246.71**
T = FaceNet64 $\mathcal{D}_{priv}$ = CelebA $\mathcal{D}_{pub}$ = FFHQ	BREPMI		43.00 ± 5.14	1470.55
	LOKT	$C \circ D$	43.27 ± 3.53	1516.18
		S	59.13 ± 2.77	1437.86
		S_{en}	**62.07 ± 3.89**	**1428.04**



	Attack Acc. (↑)
Private Training Data	
Existing SOTA	73.93%
Our Reconstruction Results	93.93%

Conclusion

- We propose Label-only Model inversion via Knowledge Transfer (LOKT) by transferring decision knowledge from the target model to surrogate models and performing white-box attacks on the surrogate models.
- We propose a new T-ACGAN to leverage generative modeling and the target model for effective knowledge transfer.
- We perform analysis to support that our surrogate models are effective proxies for the target model for MI.

Thank you!

Poster Session
Wed 13 Dec 10:45 a.m. CST — 12:45 p.m. CST
Great Hall & Hall B1+B2
#1517

Project page

https://ngoc-nguyen-0.github.io/lokt/