Privacy-Preserving Split Learning via Patch Shuffling over Transformers

Dixi Yao, Liyao Xiang, Hengyuan Xu, Hangyu Ye, Yingqi Chen

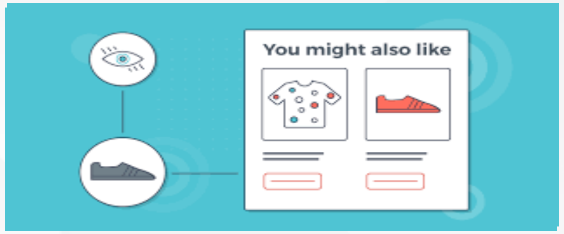
John Hopcroft Center, Shanghai Jiao Tong University, China

Sept., 2022

01

Background

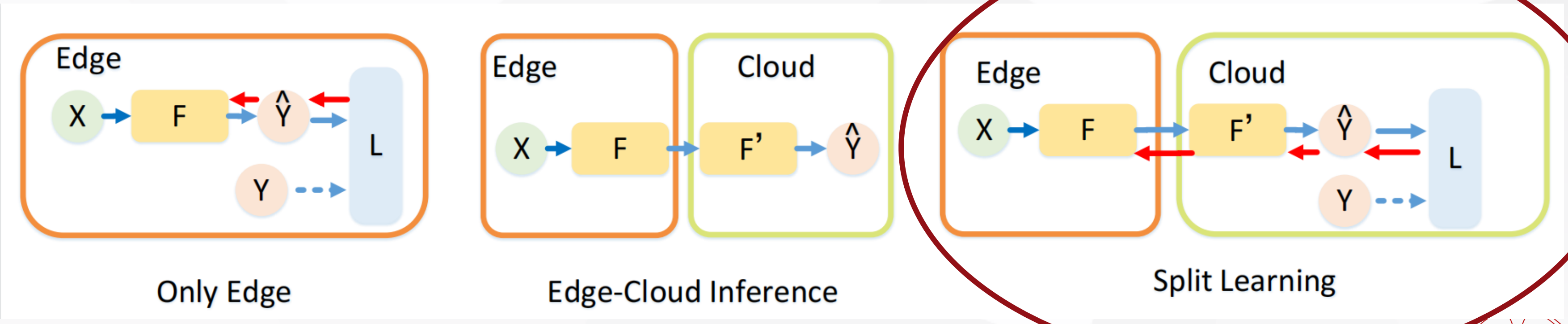
New Computational Paradigm



CV NLP Recommendation

Compute on edge: resource constrained

Upload to cloud: privacy leakage



Is Split learning perfect?

Challenge 1

Unprotected intermediate results :

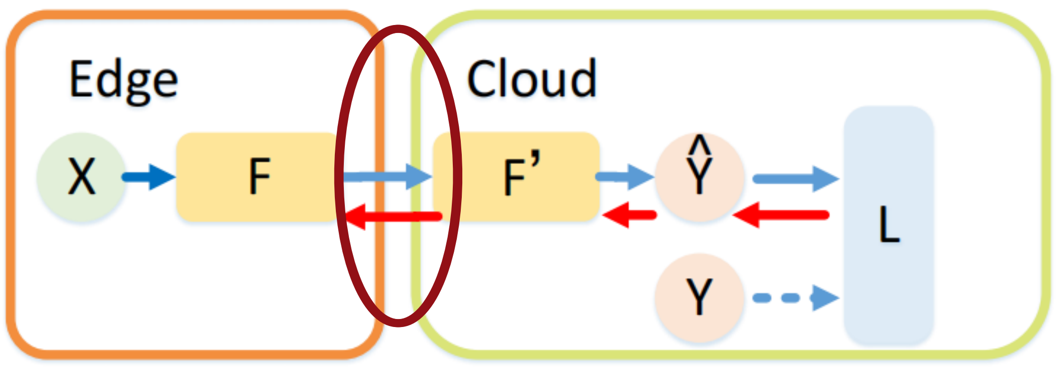
leak privacy of input !

Challenge 2

Protect label privacy :

Labels should not leave cloud if labels are proprietary

An Example



Split Learning

Facial images:
private on edges

Forward loop:

intermediate features

Backward loop:

error gradients

Identity: Bob
belongs to a proprietary enterprise database

Is Split learning perfect?

Challenge 1

Unprotected intermediate results :

Leak privacy of input !

Challenge 2

Protect label privacy :

Labels should not leave cloud if labels are proprietary

Challenge 3

Privacy in training

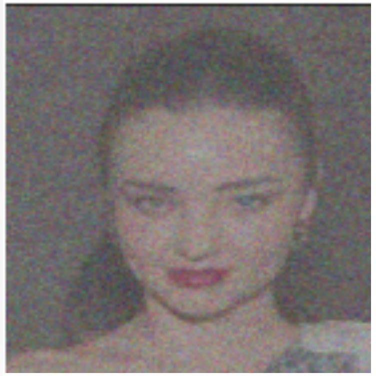
Leakage would occur in each iteration

Protecting training data privacy is hard

Inference: one-time transmission

Training: multiple forward & backward rounds

Privacy should be guaranteed throughout training!



Add Noise

Adding Gaussian noise barely works

Adversarial learning based methods:

Protection is effective only at convergence 🤣

Is Split learning perfect?

Challenge 1

Unprotected intermediate results :

Leak privacy of input !

Challenge 2

Protect label privacy :

Labels should not leave cloud
if labels are proprietary

Challenge 3

Privacy in training

Leakage would occur in each iteration

Challenge 4

Practicality in deployment

Tradeoff: Privacy, Efficiency & Accuracy

DNN on thin edge devices:

Low in efficiency --- cryptographic tools including homomorphic encryption, multi-party computation

High training performance:

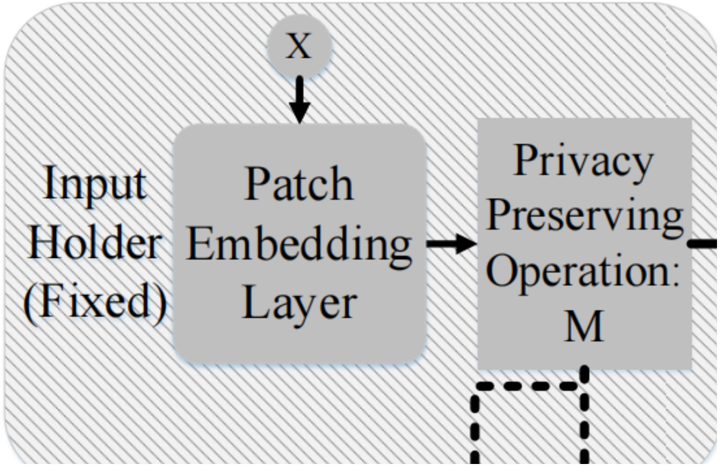
Sacrifice of accuracy --- differential privacy

02

Threat Model & Methodology

Overview

Objective: minimize task loss and maximize attacker reconstruction loss



Threat Model

White-box attack

Attacker's prior:

- ✓ Intermediate features
- ✓ Model weights

Black-box attack

Attacker's prior:

- ✓ Intermediate features
- ✓ Auxiliary datasets
- × Model weights

Adaptive attack

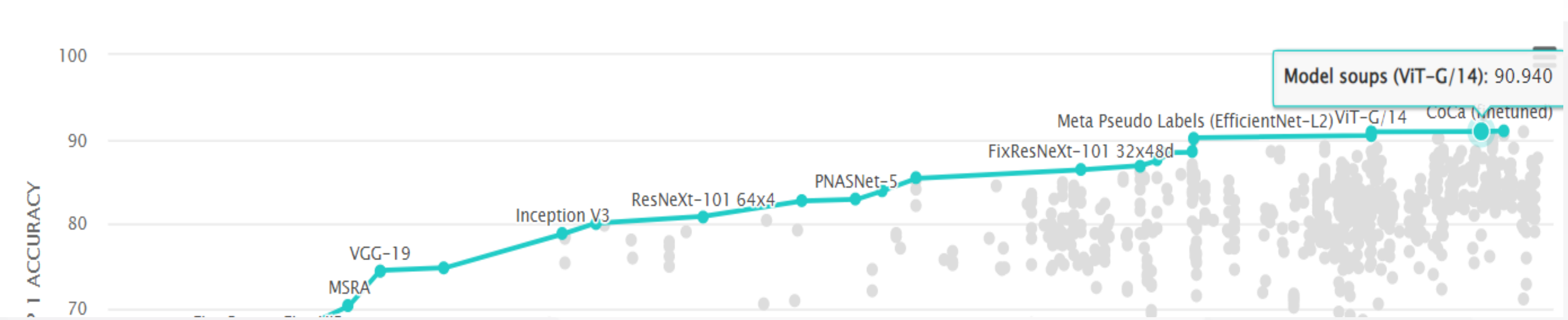
Similar to Black-box

Use features from multiple rounds

Attacker's prior:

- ✓ multiple features
- ✓ Auxiliary datasets
- × Model weights

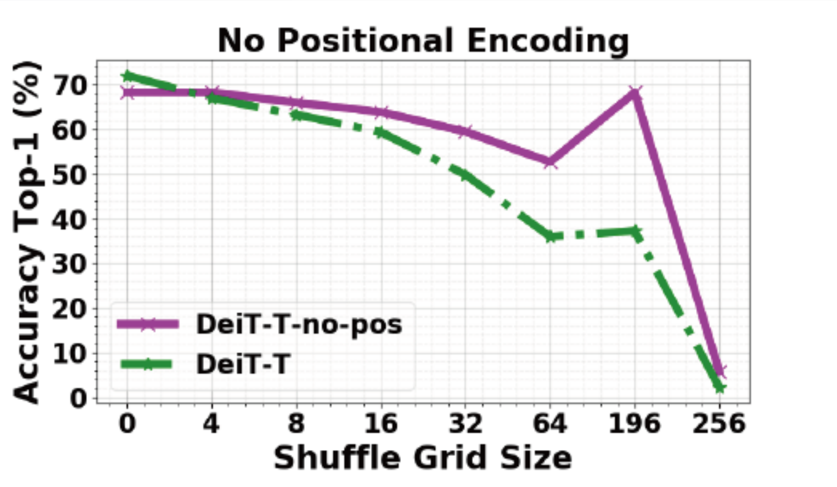
Property of Transformer



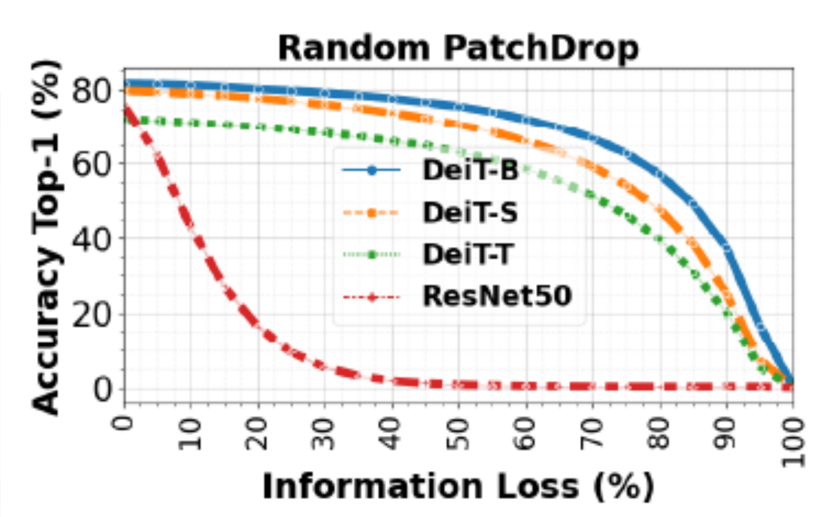
ImageNet-1k (from paperswithcode.com)

Transformer has shown **a superior accuracy**

Shuffling Invariance

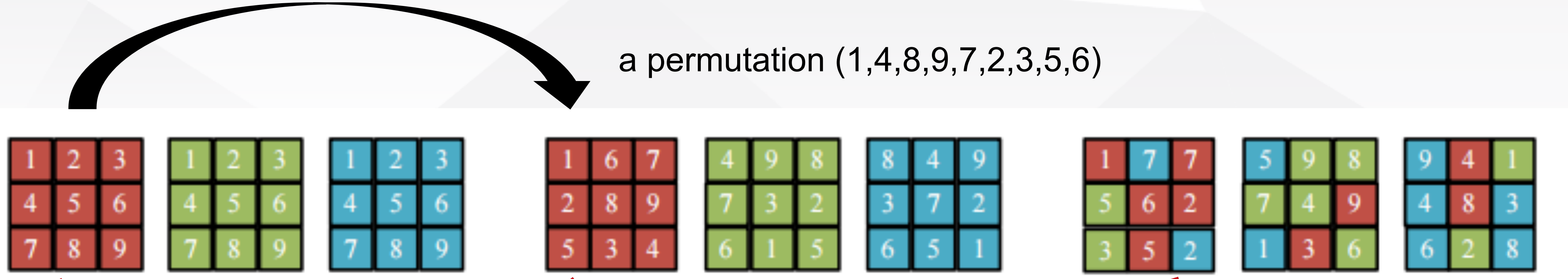


Robustness against Patch Dropping



Images from Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers." Advances in Neural Information Processing Systems 34 (2021).

Privacy Definition



(a) Original Input (b) Output of Patch Shuffling (c) Output of Batch Shuffling

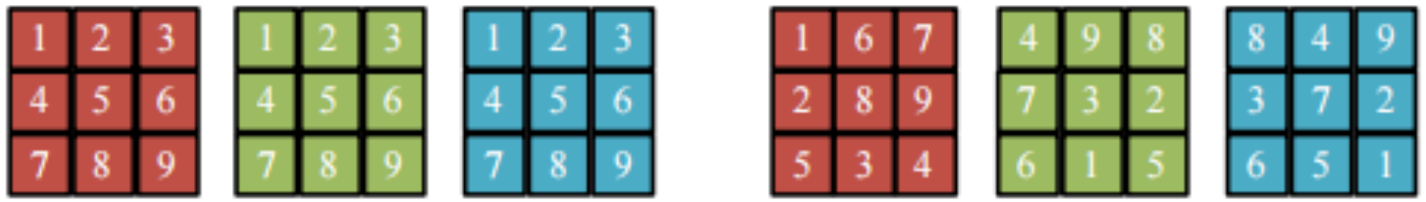
Definition 1. *(Neighbouring Permutations) We divide a single instance into N patches , and the permutations of these N patches constitute S. Any two permutation σ, $\sigma' \in S$ are defined to be neighboring.*

Definition 2. *(σ-privacy) Given private dataset X and a set of permutations S, a randomized mechanism $\mathcal{A} : f(X) \mapsto \mathcal{V}$ is σ-private if for all $x \in X$, neighbouring permutations σ and σ' and any $z \in \mathcal{V}$, we have*

$$\Pr[\mathcal{A}(\sigma(f(x))) = z] = \Pr[\mathcal{A}(\sigma'(f(x))) = z]. \qquad (6)$$

Each permutation has the same likelihood to generate z.

Patch Shuffling

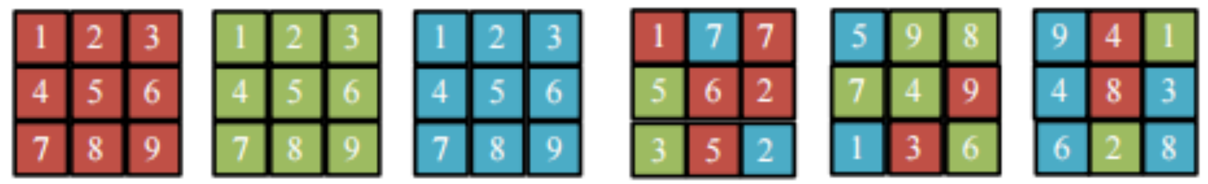


(a) Original Input (b) Output of Patch Shuffling

Apply a permutation to shuffle patches within an image

Each permutation has Pr = 1/N! (e.g., N=196) to produce z

Batch Shuffling VS Spectral Shuffling



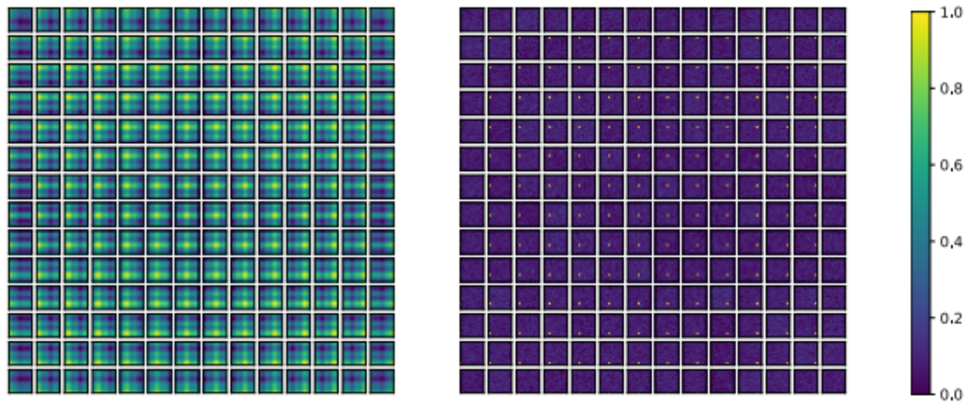
(a) Original Input (c) Output of Batch Shuffling

Batch Shuffling:

Parameters:

- ➢ Proportion of patches shuffled across diff. images within a batch
- ➢ Proportion of patches shuffled across diff. batches

Spec



Position Embedding

before patch shuffling

Further eliminate positional correlation between patches

So that each permutation has equal prob. to occur

Frequency (c) Output of Spectral Shuffling

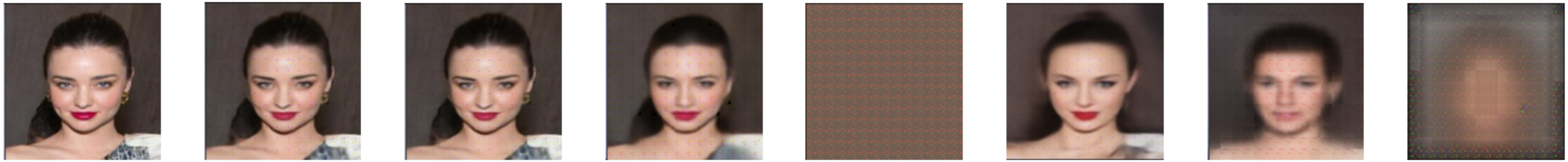
03

Evaluation

Black-Box Attack (MAE Decoder)

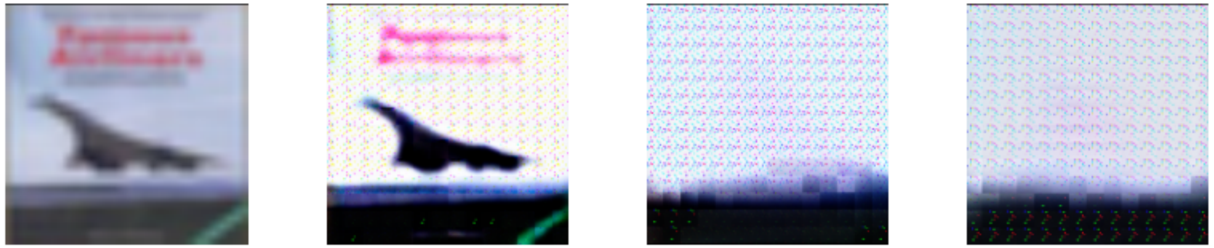
Accuracy VS Privacy: BS --- Batch Shuffling, PS --- Patch Shuffling, PS+ --- Spectral Shuffling

➢ Visualization effect of CelebA reconstruction



	(a) Input	(b) SL	(c) Adv	(d) Blur	(e) DP	(f) GN	(g) Our BS	(h) Our PS+
Accuracy(%)		91.05	90.36	89.58	80.67	87.35	89.18	88.21

➢ Visualization effect of CIFAR10 reconstruction



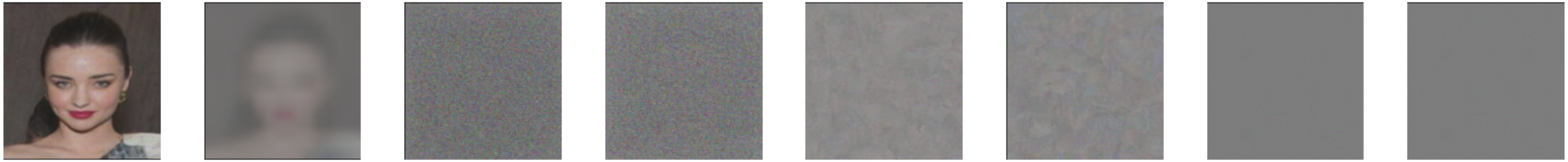
	(a) Input	(b) SL	(c) Our PS	(d) Our BS 75
Accuracy(%)		98.36	96.99	96.16

Criteo

Methods	Utility: Acc ↑	Privacy: MSE ↑
SL	**77.81**	0.0012
Our PS	77.78	**0.0015**
GN	77.28	0.0012

White-Box Attack

Attacker is aware of the model weights, but not the permutation order



(a) SL/Adv (b) Blur (c) DP (d) GN (e) Our PS (f) Our BS (g) Our PS+ (h) Jigsaw to Our BS

A stronger threat: Jigsaw solving

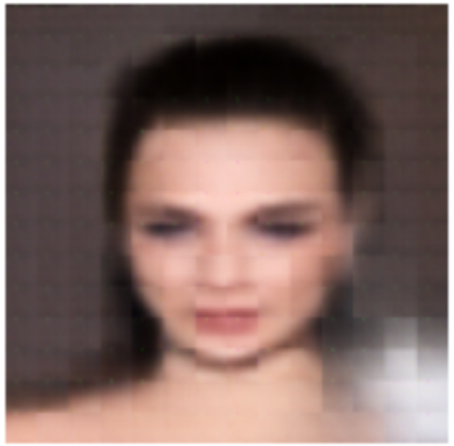
Train a model to guess the permutation order → **Failed** due to random permutation

Adaptive Attack

Attackers intercept the intermediate results throughout the whole training process

- ➢ We use 30 rounds of intermediate results to attack

(a) Input



(a) Our BS

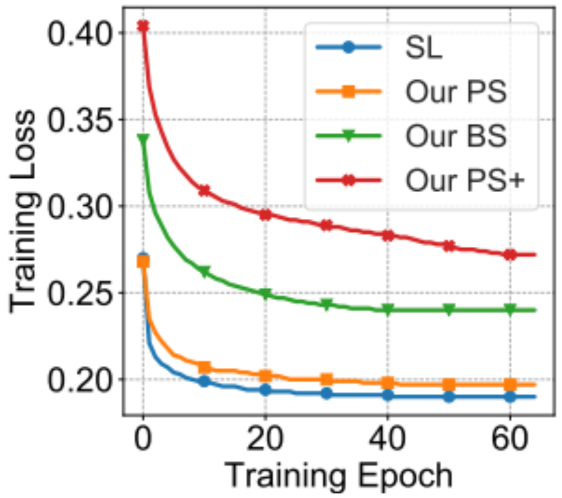
Failed to recover the original images

Privacy, Utility & Efficiency

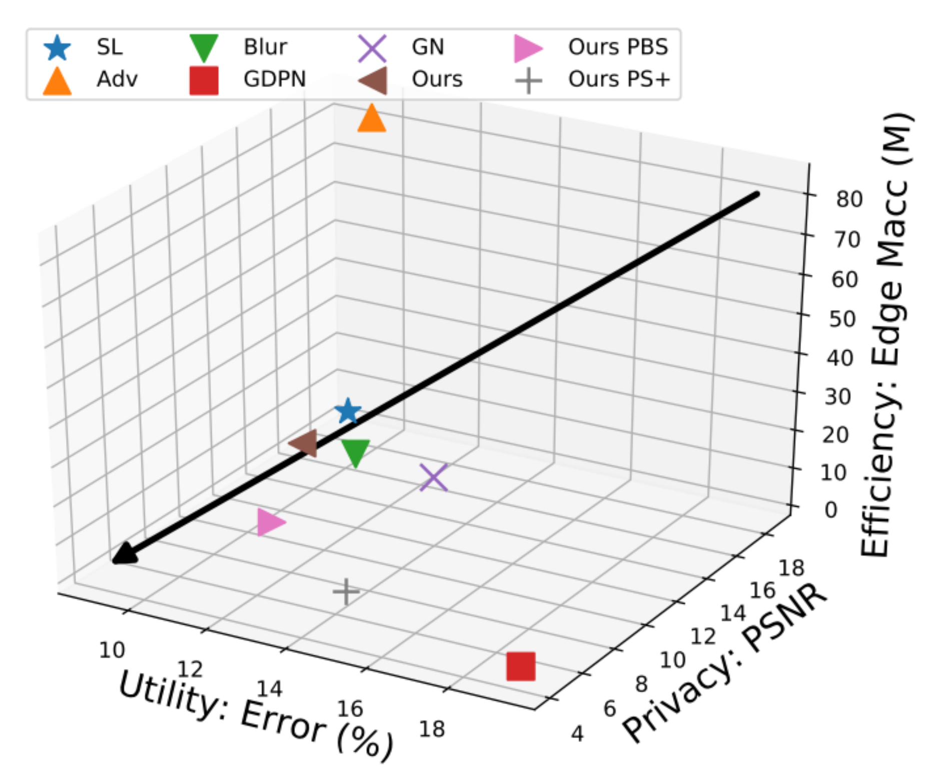
Efficiency, CelebA

Computaional and memory costs at the edge, lower is better

Methods	Macc Edge (M)↓	Mem Edge (G)↓
SL / Transform	**3.10**	**0.97**
Adv	81.63	2.43
Our PS/BS	**3.10**	**0.97**
Our PS+	1.18	1.01



Convergence curves



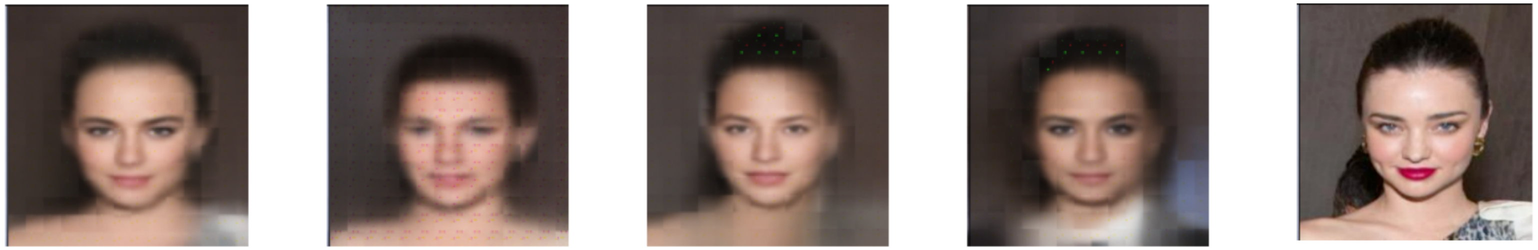
➢ Our methods have negligible impact to standard split learning

Privacy, Utility & Efficiency, CelebA:

➢ Our methods achieve ideal tradeoffs

Ablation Studies

- k: Proportion of patches shuffled across diff. images within a batch



k:	0.5	0.6	0.75	0.85	Original Input
Acc.(%):	90.29	89.18	88.54	88.76	

- k = 0.6 exhibits the best tradeoff
- a smaller k leads to better reconstruction and higher accuracy

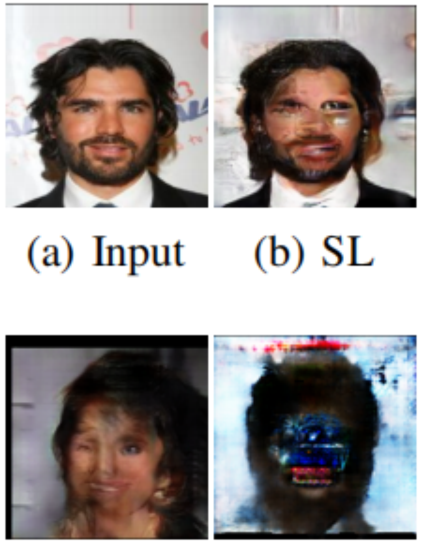
- Transferability: against black-box attacks with auxiliary datasets

(a) Input (b) SL (c) Our BS

Auxiliary set: CelebA
Private set: LFW

(a) SL (b) Our BS

Auxiliary set: LFW
Private set: CelebA



(a) Input (b) SL (c) BS (d) PS+

- Adaptability: change attack model to CNN model --- Pix2Pix

Takeaways

- An efficient privacy-preserving approach in split learning
- A formal privacy guarantee based on patch shuffling
- Eliminating positional correlation by spectral shuffling

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

Thanks!

饮 水 思 源 爱 国 荣 校