# Individual Property Inference over Collaborative Learning in Deep Feature Space

**Haoxin Yang [1,2], Yi Wang [2*], Bin Li [1,3]**

[1] Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China

[2] Dongguan University of Technology, Dongguan, China

[3] Shenzhen Institute of Articial Intelligence and Robotics for Society, Shenzhen, China


July 21, 2022 ICME 2022

# Motivation

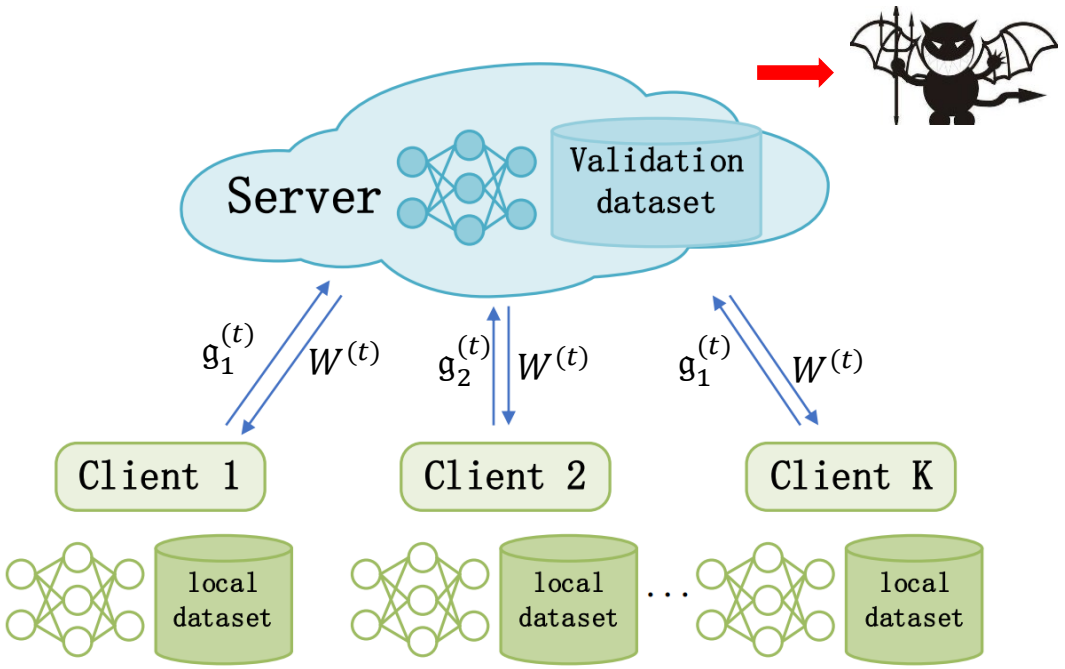


Figure 1: Framework of collaborative learning(CL).

- Client doesn't share local data with others.
- CL framework is supposed to protect privacy.
- However, information still leak from gradient.

For each round CL, the participants perform synchronous SGD, the format of s-SGD:

$$W^{(t+1)} = W^{(t)} - \eta \sum_{k=1}^{K} \frac{m_k}{M} \cdot \mathbf{g}_k^{(t)}$$

# Previous Works

The framework of CL is still subjected to

- Membership Inference: Label-only membership inference[Choquette-Choo et al.], Unintended leakage[Melis et al.]
- Property Inference: Unintended leakage[Melis et al.] , Honest-but-Curious Nets[Malekzadeh et al.]
- Sample Reconstruction: DLG[Zhu et al.], IG[Geiping et al.], GI[Yin et al.]

Specially, for property inference, these methods are failed to infer the property of *individual sample.* And for using sample reconstruction as a stepping stone of property inference, these methods are limited to the *batchsize* of training data.

# Privacy Leakage in Deep Space

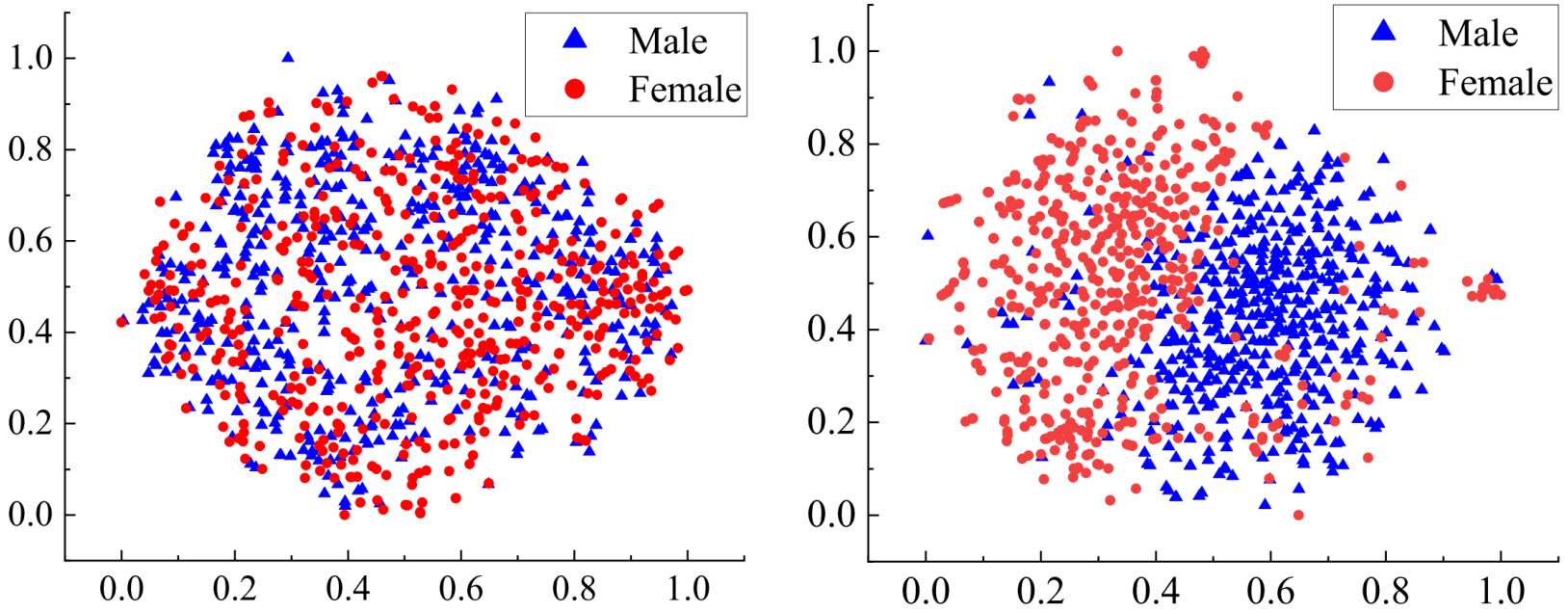


Figure 2: Visualizations of 1,000 face images from the CelebA dataset produced by t-SNE of deep features extracted from a shared model.
(a) at the beginning with random initialization, and (b) at the end of CL in the s-SGD setting.

Which inspires us to make privacy inference for individual sample in the deep feature space.

# Proposed Method

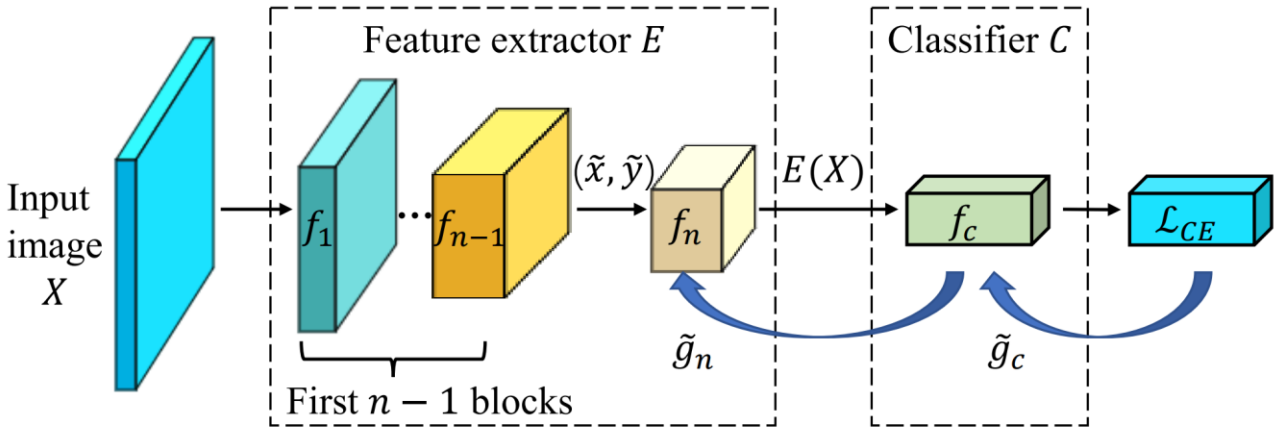


Figure 3: Proposed deep feature reconstruction on CNN.

- Target: Infer the property of individual sample $X$.
- Method: Reconstruct the deep feature $E(X)$.
- Inference: Infer the property of $X$ by $E(X)$.

# Deep Feature Reconstruction

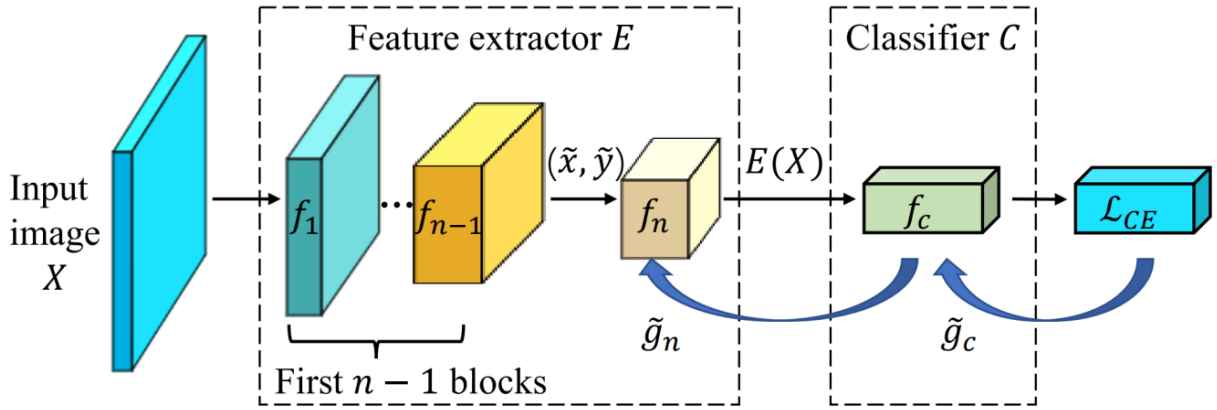


Figure 3: Proposed deep feature reconstruction on CNN.

1. Initialize a pair of dummy data $(\tilde{x}, \tilde{y})$, inject $(\tilde{x}, \tilde{y})$ into $f_c$ and $f_n$, compute the gradient i.e.,

$$\tilde{g}_c = \nabla_{f_c} \mathcal{L}_{CE}[f_c(f_n(\tilde{x})), \tilde{y}]$$
$$\tilde{g}_n = \nabla_{f_n} \mathcal{L}_{CE}[f_c(f_n(\tilde{x})), \tilde{y}] \quad (1)$$

2. Optimize $(\tilde{x}, \tilde{y})$ by objective function

$$\mathcal{L} = \lambda d(g_n, \tilde{g}_n) + d(g_c, \tilde{g}_c) \quad (2)$$

Where

$$d(g, \tilde{g}) = \left(1 - \frac{\langle g, \tilde{g} \rangle}{||g|| \cdot ||\tilde{g}||}\right) + (1 - \exp(\frac{-||g - \tilde{g}||^2}{\sigma^2})) \quad (3)$$

3. Obtain the best reconstructed feature by

$$E(X) := f_n(\tilde{x}^*) \quad (4)$$

Where $\tilde{x}^*$ is the best optimized $\tilde{x}$.

# Deep Feature-based Property Inference Attacks

- The server has an auxiliary dataset $D_{aux}$ labeled with his interested property $p$.
- The server collects uploaded gradient and reconstructs the deep feature $E(X)$ of each sample.
- The server trains a property inference model $f_p$ by $D_{aux}$ and the shared model.
- The server performs the property inference attack by $f_p$ from the reconstructed $E(X)$.

# Performance Evaluations

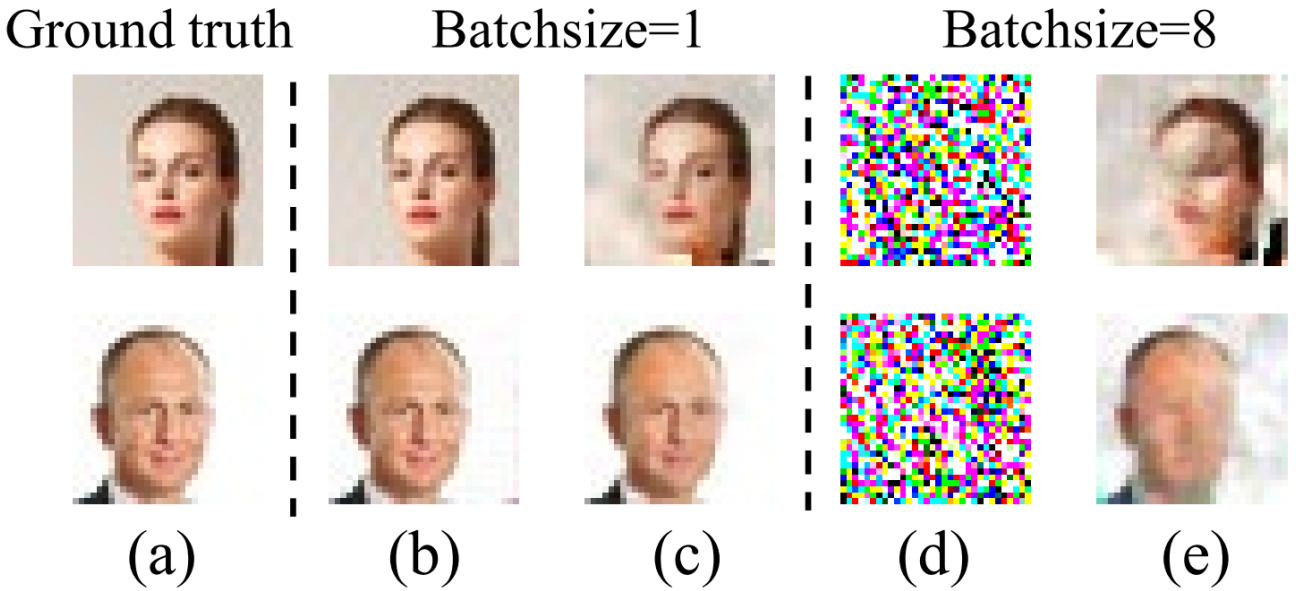


Figure 4: Image reconstruction with different batchsize:
(a) Ground truth from CelebA.
(b) and (d) results by DLG.
(c) and (e) results by GradInversion.

- The objectives of reconstruction-based methods are improving the visual quality and image fidelity at pixel level.
- This is not necessary for property inference tasks in most scenarios.
- We propose a novel inference approach by reconstructing samples in the deep feature space.

# Performance Evaluations

Table 1: Property inference accuracy (in %) by comparing with two non-reconstruction based methods.

| Dataset | Property | UL | HCN | Proposed |
|---|---|---|---|---|
| CelebA | Gender | 87.01 | 92.72 | **95.75** |
| | Glasses | 80.14 | 90.96 | **94.64** |
| | Smile | 77.53 | 85.54 | **86.93** |
| | Young | 65.83 | 73.47 | **81.18** |
| LAD | Wheel | 91.13 | 92.62 | **96.13** |
| CUB | Beak | 66.45 | 68.93 | **75.08** |

Table 2: The gender inference accuracy (in %) on CelebA at different mini-batch size for training updates.

| Batchsize | DLG | IG | GradInversion | Proposed |
|---|---|---|---|---|
| 1 | 93.12 | 94.86 | 94.79 | **95.35** |
| 8 | 50.00 | 67.34 | 52.34 | **95.42** |
| 32 | 50.00 | 54.31 | 73.23 | **95.49** |

Table 3: Property inference accuracy using cross-datasets.

| Method | Gender | Glasses | Smile | Young |
|---|---|---|---|---|
| UL | 74.14 | 50.00 | 50.00 | 50.00 |
| HCN | 72.04 | 50.20 | 53.23 | 50.00 |
| Proposed | **89.63** | **82.29** | **76.44** | **65.90** |

# Inference Against Defense

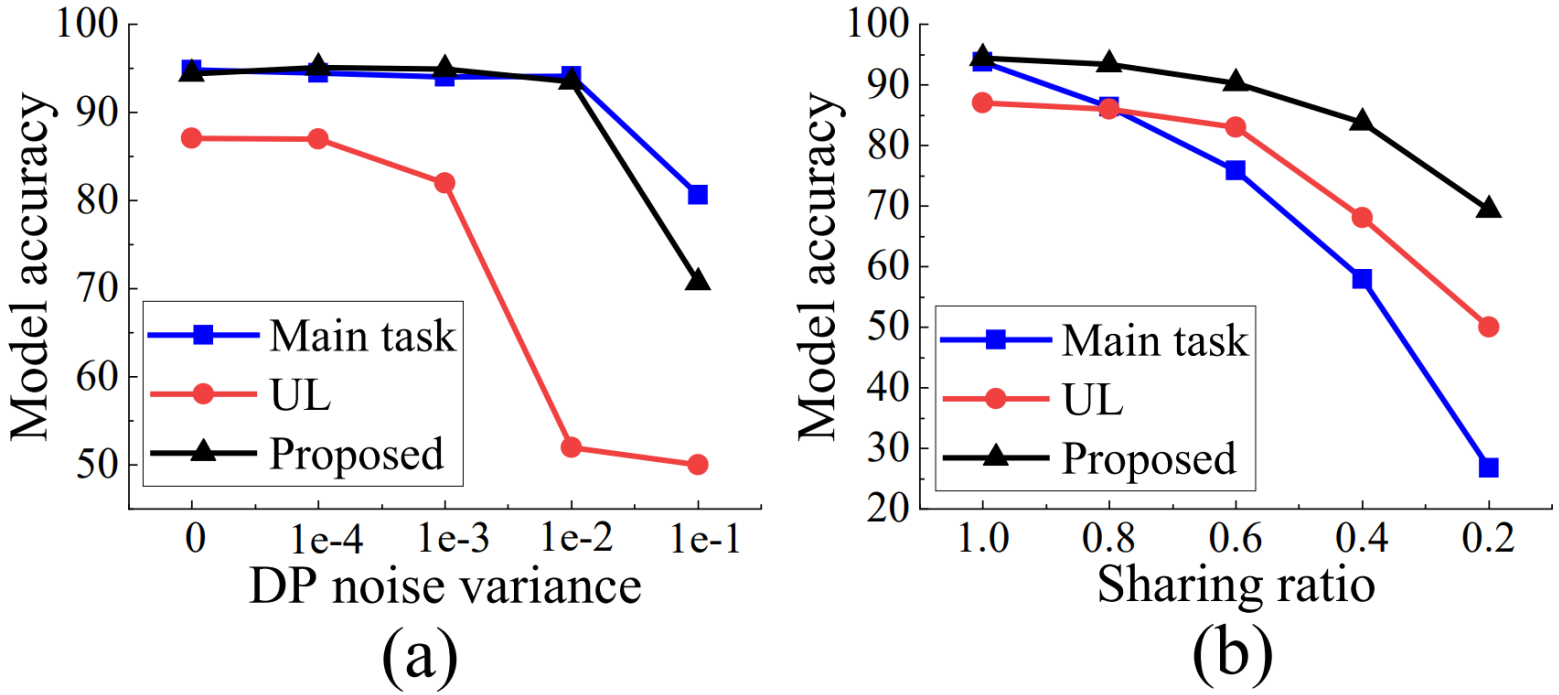


Figure 5: Inference accuracy in the presence of (a) DP by varying noise, and (b) sharing less gradients in the updates.

The improvement of privacy protection results in the reduction of model utility.

# Ablation Test

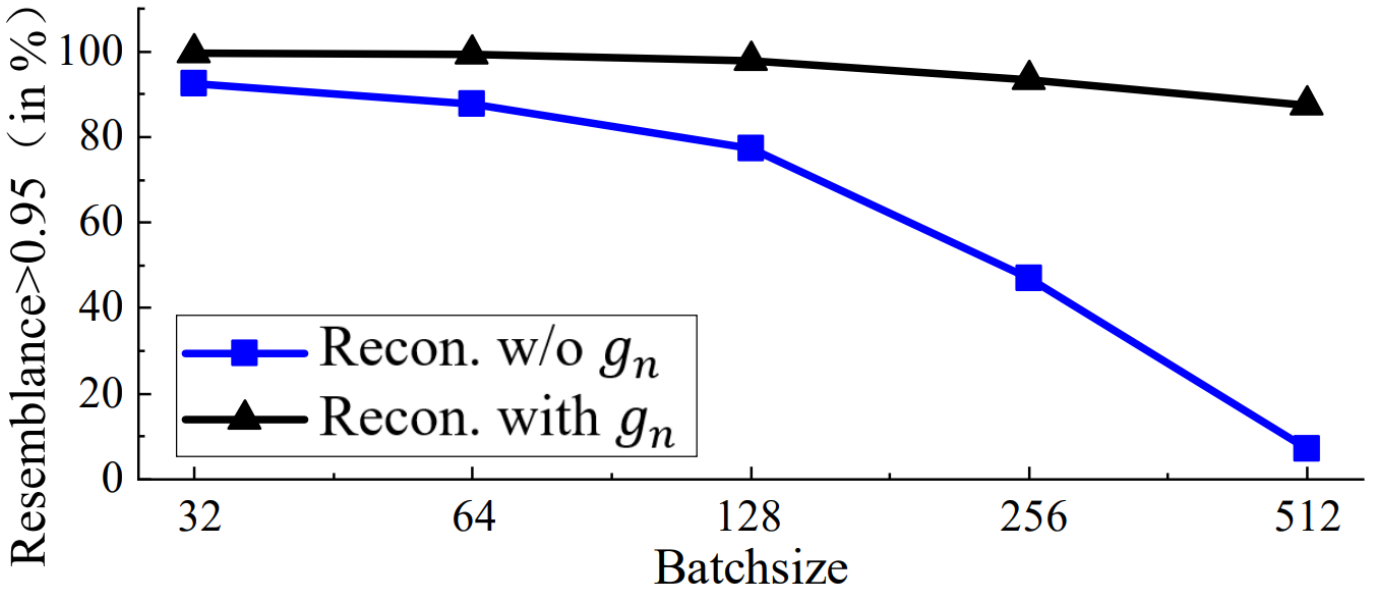


Figure 6: Feature reconstruction by client gradient updates.

$g_n$ is important by providing additional information for reconstruction especially with large batch size of updates.

# Conclusion

- We demonstrate privacy leakage in the deep feature space and high-level feature encode unintended information of training data.
- We propose a novel deep feature reconstruction method.
- We design a deep feature-based inference algorithm that perform property inference attack for individual sample.


11

Thank you!

Q&A

Contact: wangyi@dgut.edu.cn

# Reference


- Choquette-Choo C A, Tramer F, Carlini N, et al. Label-only membership inference attacks[C]//International conference on machine learning. PMLR, 2021: 1964-1974.
- Melis L, Song C, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning[C]//2019 IEEE symposium on security and privacy (SP). IEEE, 2019: 691-706.
- Malekzadeh M, Borovykh A, Gündüz D. Honest-but-Curious Nets: Sensitive Attributes of Private Inputs Can Be Secretly Coded into the Classifiers' Outputs[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 825-844.
- Zhu L, Liu Z, Han S. Deep leakage from gradients[J]. Advances in neural information processing systems, 2019, 32.
- Geiping J, Bauermeister H, Dröge H, et al. Inverting gradients-how easy is it to break privacy in federated learning?[J]. Advances in Neural Information Processing Systems, 2020, 33: 16937-16947.
- Yin H, Mallya A, Vahdat A, et al. See through gradients: Image batch recovery via gradinversion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16337-16346.