WED-AM-007
West Building Exhibit Halls ABC 392

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Privacy-preserving Adversarial Facial Features

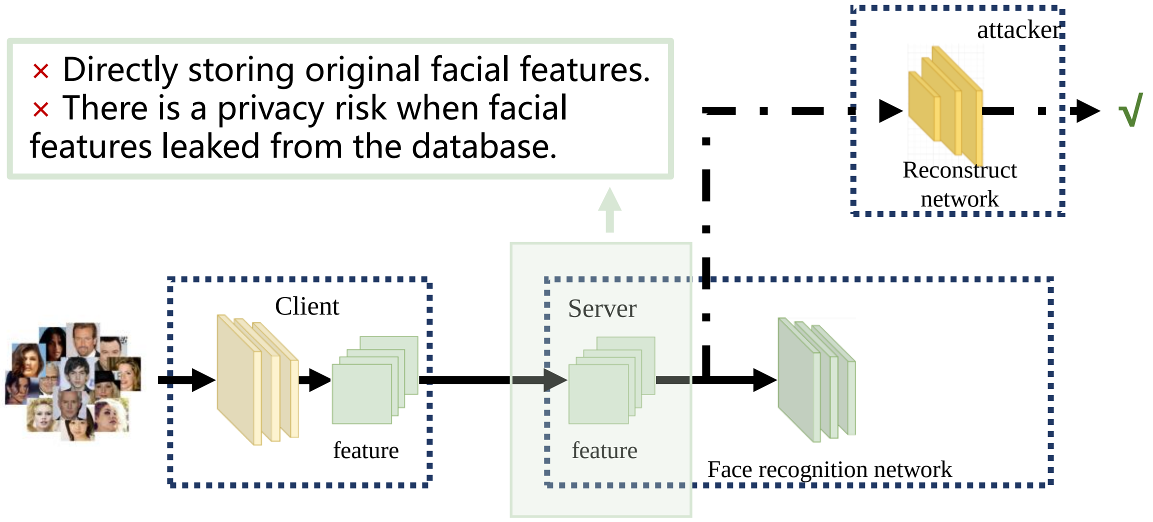
Zhibo Wang[†,≀,*], He Wang[†], Shuaifan Jin[†], Wenwen Zhang[‡], Jiahui Hu[†], Yan Wang[‡]
Peng Sun[♯], Wei Yuan[♮], Kaixin Liu[♮], Kui Ren[†]

[†]School of Cyber Science and Technology, Zhejiang University, P. R. China
[≀]ZJU-Hangzhou Global Scientific and Technological Innovation Center [‡]Alibaba Group, P. R. China
[♯]College of Computer Science and Electronic Engineering, Hunan University, P. R. China
[♮]School of Cyber Science and Engineering, Wuhan University, P. R. China

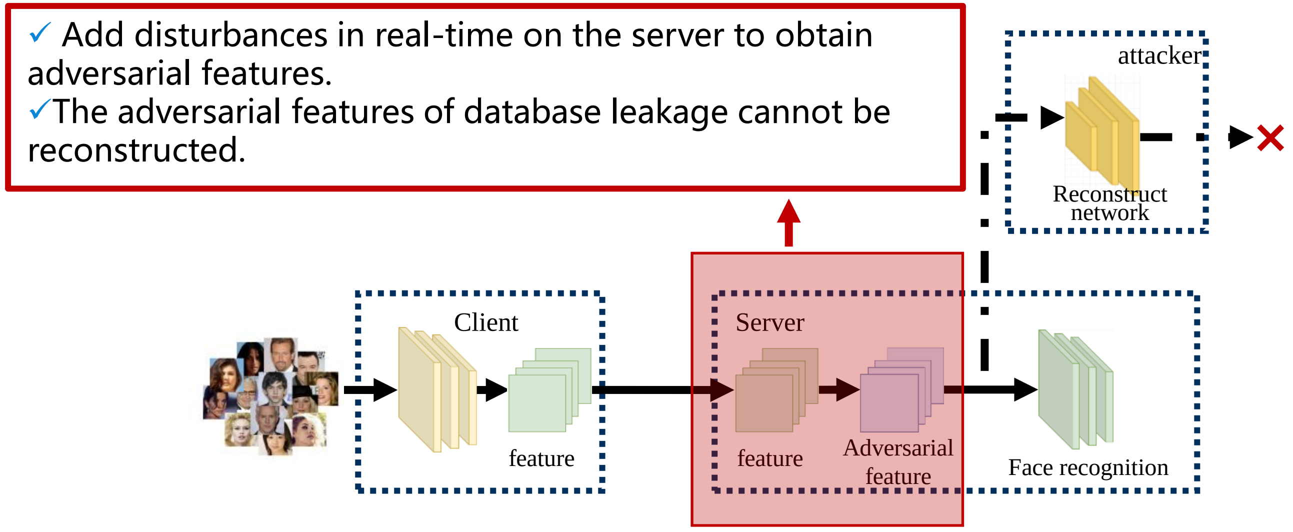
Quick Preview

❑ How to prevent privacy leakage from facial features?

- We propose a novel facial privacy-preserving method (namely AdvFace), which can generate privacy preserving adversarial features against unknown reconstruction attacks while maintaining face recognition accuracy.



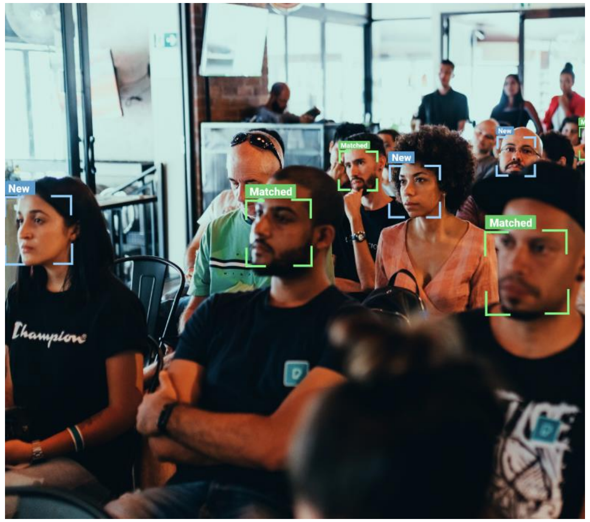
Traditional facial recognition models



an adversarial features-based face privacy protection (AdvFace)

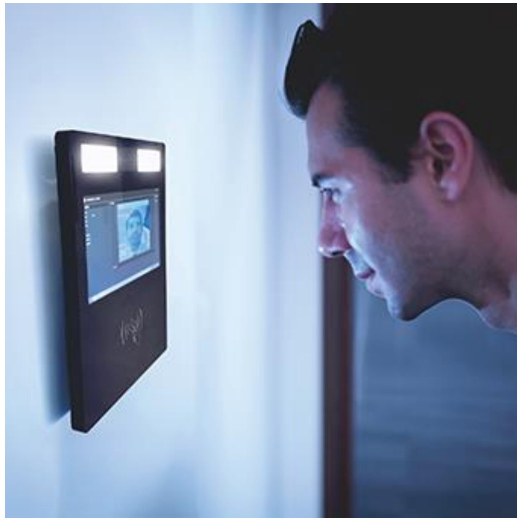
Background

❑ Face recognition system is widely used



Security

Healthcare



Attendance

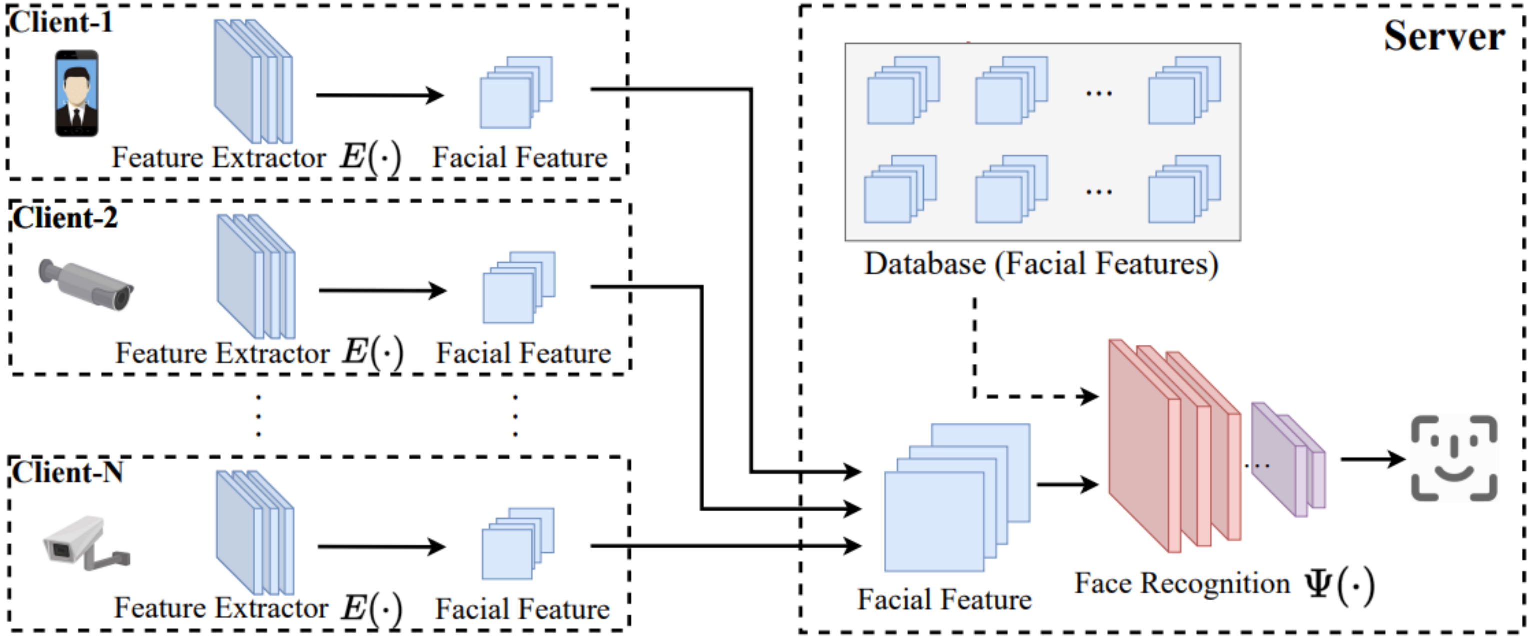
Finance

Background

❑ Typical Client-Server face recognition system

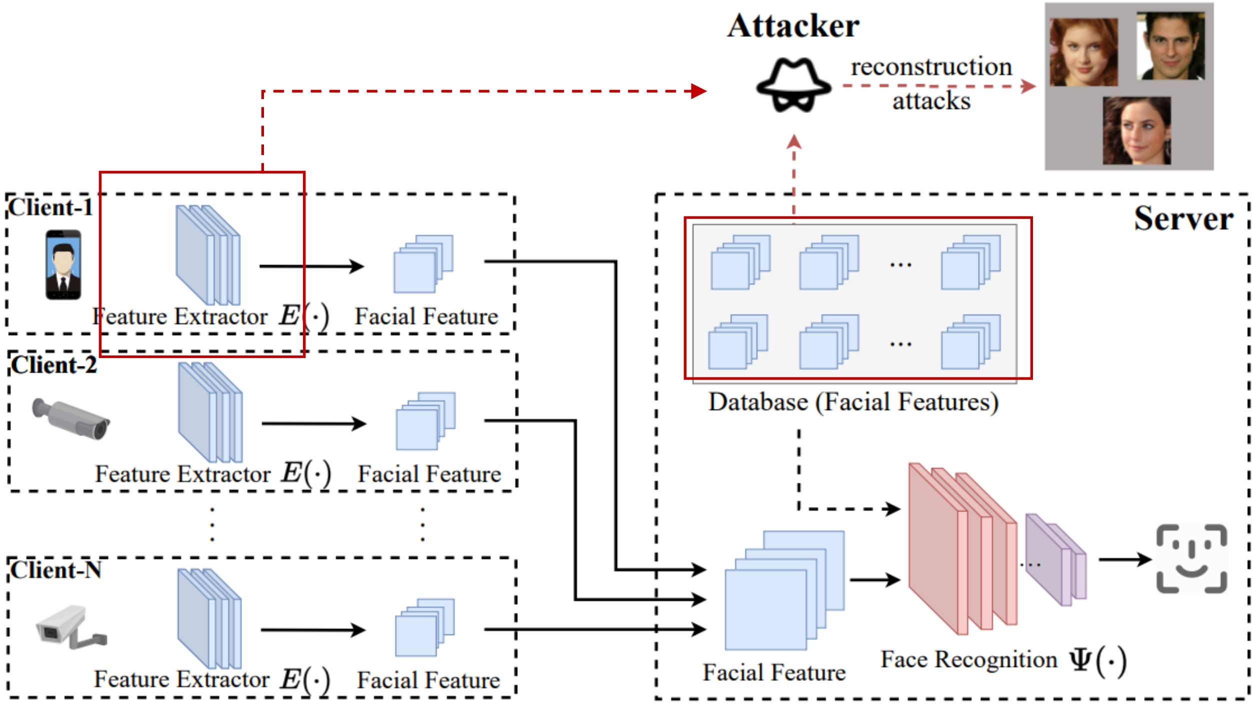
- Stored Keys: Facial image x → Facial features E(x)

A straightforward solution to prevent direct privacy leakage



Background

- **However, a powerful but realistic attacker can reconstruct the original image from the facial feature** [1~3]
  - Using Neural Networks $\emptyset$ to implement reconstruction attack: $\emptyset : \boldsymbol{z} \rightarrow \boldsymbol{x}, \quad \boldsymbol{z} = \boldsymbol{E}(\boldsymbol{x})$
  - Training Objective: Minimizing the reconstruction loss $\boldsymbol{L_R}(\boldsymbol{Z}, \boldsymbol{X})$



$$\mathcal{L}_R(Z, X) = \sum_{i=1}^{N} \|x_i - R(z_i)\|_1$$

[1] Alexey Dosovitskiy, et al. Inverting visual representations with convolutional networks.
[2] Zecheng He, et al. Model Inversion Attacks Against Collaborative Inference.
[3] Guangcan Mai, et al. Deep models under the gan: information leakage from collaborative deep learning.

Background

❑ Motivation:

- Typical protection methods:
  - Encryption[1,2] : incur large computation overhead and redeployment costs
  - DP(Differential privacy)[3,4] : cannot defend against reconstruction attacks
  - Frequency domain[5,6] : cannot defend against reconstruction attacks and incur redeployment costs
  - Adversarial training[7] : cannot maintain recognition accuracy and incur redeployment costs

❑ Goal:

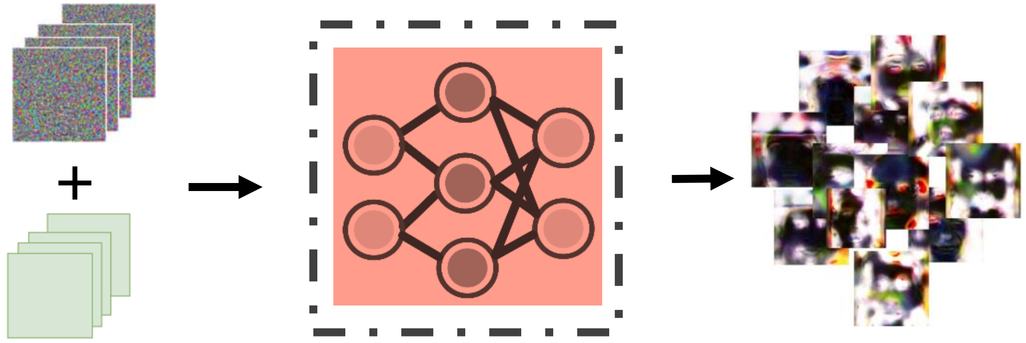
- Resist reconstruction attack
- Maintain face recognition accuracy

[1] Craig Gentry, et al. Implementing gentry's fully-homomorphic encryption scheme.
[2] Xiaoyu Kou, et al. Efficient and privacy-preserving distributed face recognition scheme via facenet.
[3] Mahawaga Arachchige Pathum Chamikara , et al. Privacy preserving face recognition utilizing differential privacy.
[4] Yunlong Mao, et al. A privacy-preserving deep learning approach for face recognition with edge computing.
[5] Jiazhen Ji , et al. Privacy-preserving face recognition with learnable privacy budgets in frequency domain.
[6] Yuxi Mi , et al. Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain.
[7] Ang Li , et al. Deepobfuscator: Obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones.

Key Idea of AdvFace

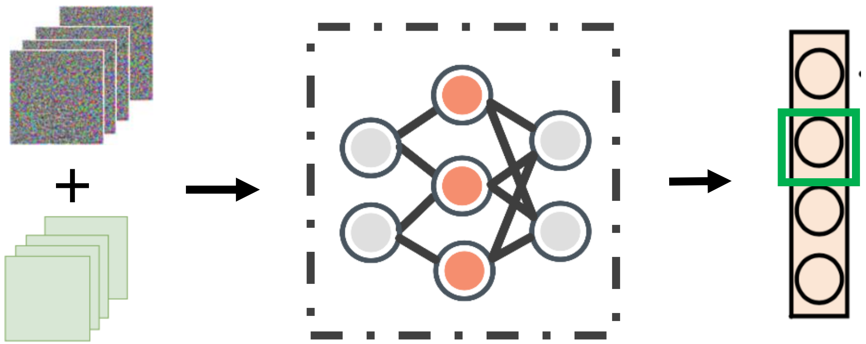
❑ Design adversarial latent noise to minimize the impact on facial recognition networks and maximum the impact on attack networks

- impact on attack networks:



- Noise addition on facial features lightly.
- The added noise serves as an adversarial sample to induce attack errors.

- impact on facial recognition networks:

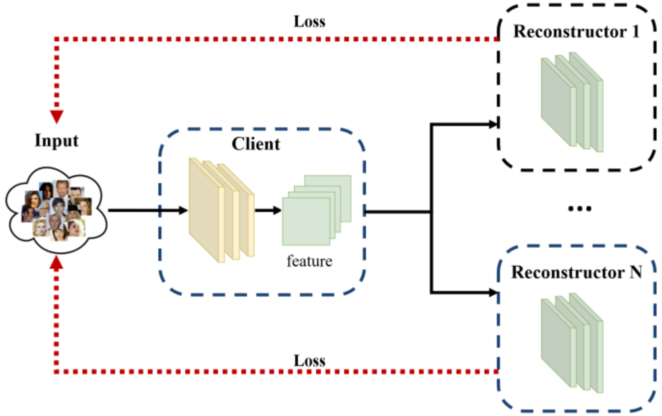


- The noise also affects the facial recognition network.
- The added noise can be considered as a normal disturbance with only a slight impact.

Key Idea of AdvFace

How to resist reconstruction attack?

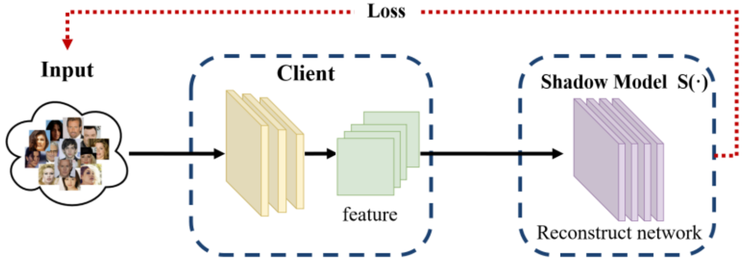
- **Key idea: Disrupt the mapping from facial feature to facial image to defend reconstruction attacks**



Observation: Even if the attacker uses a different reconstruction network structure, the mapping relationships learned by the network are similar.

- **Build a shadow model $S()$**

$$\mathcal{L}_S(Z, X) = \sum_{i=1}^{N} \|x_i - S(z_i)\|_1 . \quad (1)$$



Simulate the practical attacker to learn the mapping from facial feature to facial image, and use adversarial perturbation to disrupt the mapping.

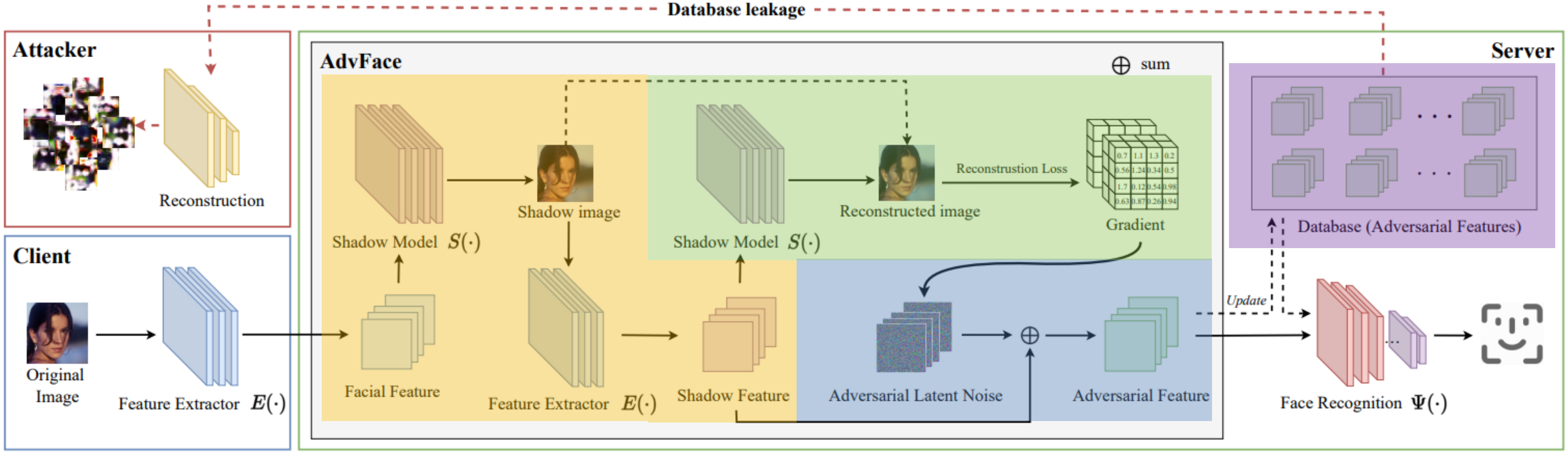
How to maintain face recognition accuracy?

- **Key idea: Strictly limit the magnitude of perturbation to and reduce its impact on face recognition**

$$z_{t+1} = z_t + \alpha \cdot sign(grad(S, z_t, \tilde{x})), \; z_0 = \tilde{z}, \quad \text{s.t.} \; \|z_{t+1} - z_t\| < \boxed{\varepsilon,} \quad (2)$$

Privacy Protection

- Pipeline: How to replace facial feature by adversarial feature



- Generate shadow images $\tilde{x}$ to replace unobtainable original images x, and obtain its feature $\tilde{z}$

$$z = E(x),\ \tilde{x} = S(z), \tilde{z} = E(\tilde{x}), \qquad (3)$$

- Obtain the reconstructed image and calculate the gradient

$$grad(S, \tilde{z} + \delta, \tilde{x}) = \nabla_\delta \| S(\tilde{z} + \delta) - \tilde{x} \|_1, \qquad (4)$$

- Generate adversarial features from gradient

$$z_{t+1} = z_t + \alpha \cdot sign(grad(S, z_t, \tilde{x})),\ z_0 = \tilde{z}, \quad \text{s.t. } \|z_{t+1} - z_t\| < \varepsilon, \qquad (5)$$

- Replace facial feature by adversarial feature in database

Privacy Protection

■ Maintaining face recognition accuracy

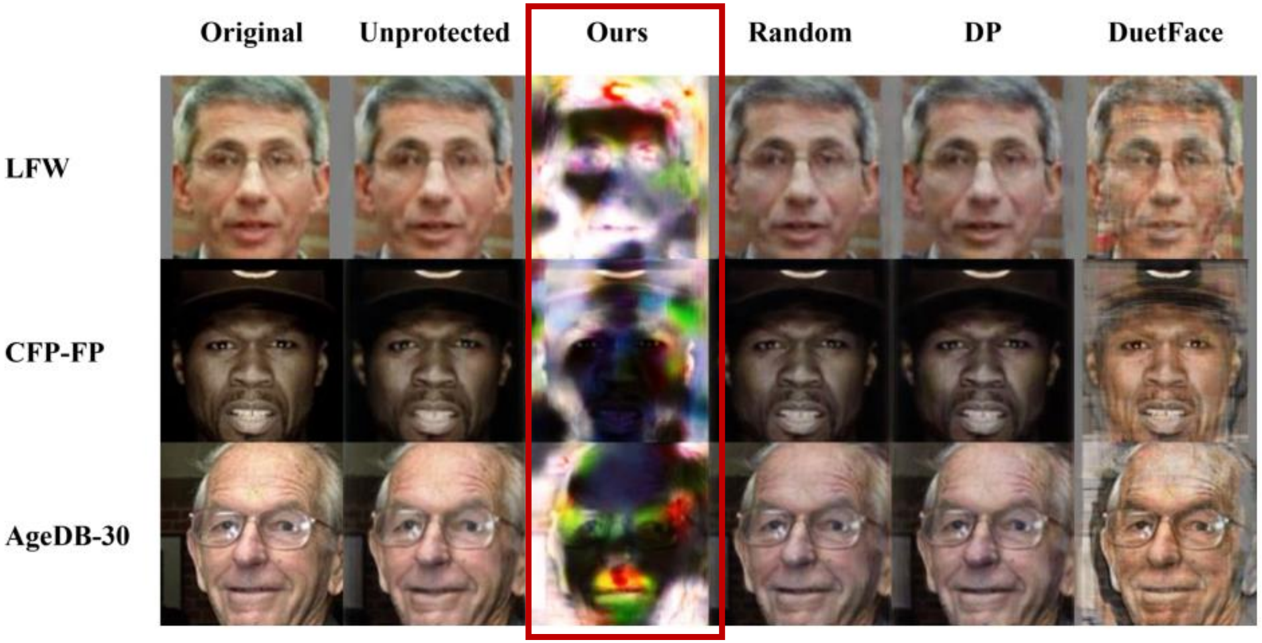
Mehods	LFW	CFP-FP	AgeDB-30
Unprotected	98.13%	93.16%	87.50%
Random	97.07%	91.71%	86.83%
DP	97.38%	91.66%	86.40%
DuetFace	98.02%	84.37%	87.10%
Ours(online)	96.73%	**91.89%**	86.32%
Ours(offline)	**98.05%**	91.64%	**87.37%**

■ Defense against malicious identity inference

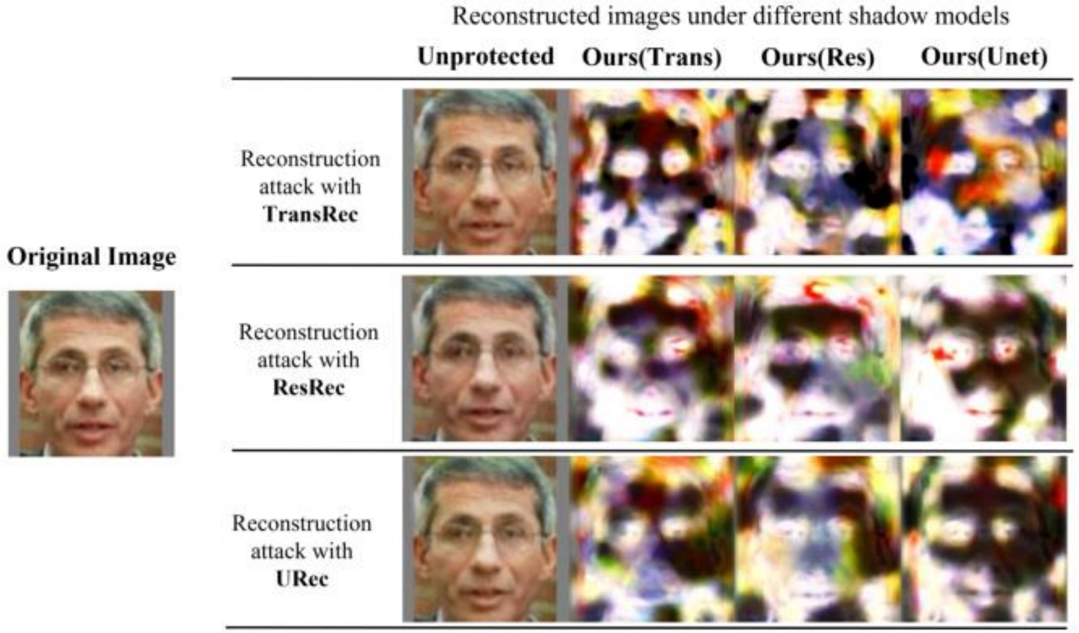
Dataset	**Ours(res)**	**Random**	**DP**	**DuetFace**	**Unprotected**
LFW	**3.37%**	94.83%	94.00%	95.07%	97.40%
CFP-FP	**17.89%**	85.60%	88.20%	74.74%	89.71%
AgeDB-30	**24.97%**	80.13%	79.97%	87.80%	84.53%

Privacy Protection

- Visualization of resistance to reconstruction attack



- Transferability to different attack settings



Conclusion

- **We propose a novel facial privacy-preserving method (namely AdvFace).**
  - AdvFace can generate privacy preserving adversarial features against unknown reconstruction attacks while maintaining face recognition accuracy.
  - AdvFace can be easily integrated into deployed face recognition systems as a plug-in privacy-enhancing module.

- **We unveil the rationale of the reconstruction attack.**
  - Breaking the mapping from features to facial images has strong transferability in defense against reconstruction attacks, as different attack models learn the same mapping.

Thank You

ZHEJIANG UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA