Gradient Obfuscation Gives a False Sense of Security in Federated Learning

Kai Yue[1] Richeng Jin[2]

Chau-Wai Wong[1] Dror Baron[1] Huaiyu Dai[1]

[1]NC State University [2]Zhejiang University

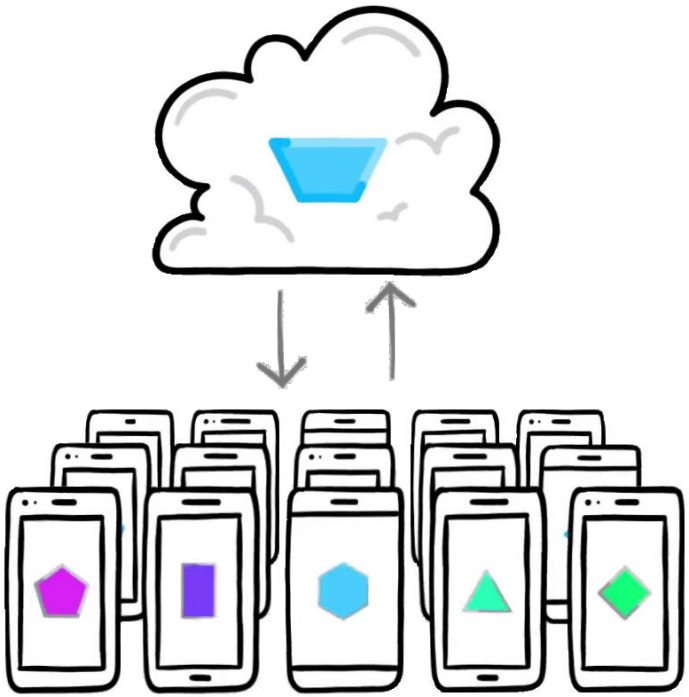
USENIX Security 2023

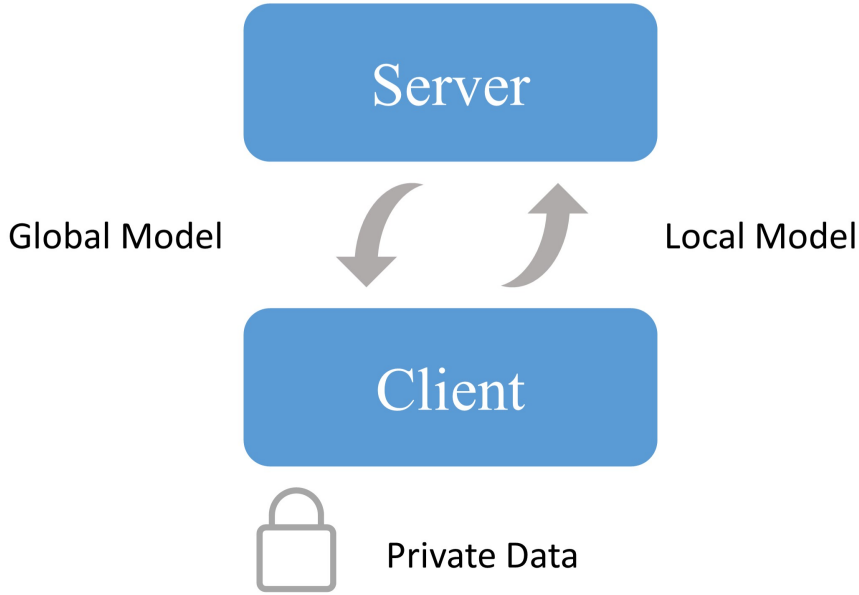
August 11, 2023

Index

Background: Federated Learning (FL)

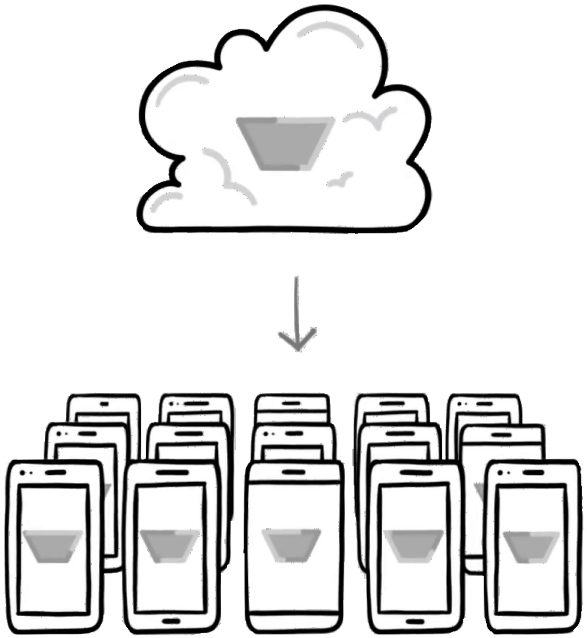
- Clients with private data jointly solve a machine learning task
- Raw data stored locally & not exchanged





[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

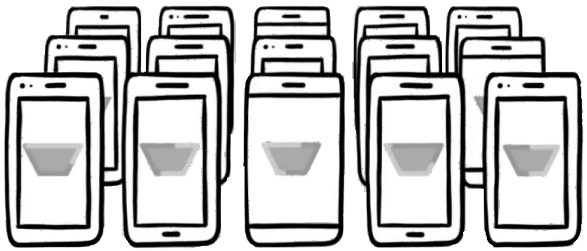
Background: Federated Averaging (FedAvg)



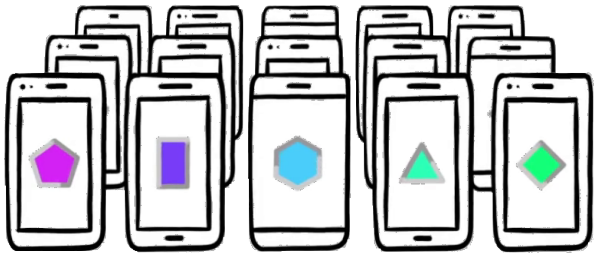
clients **download** the model

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Background: Federated Averaging (FedAvg)



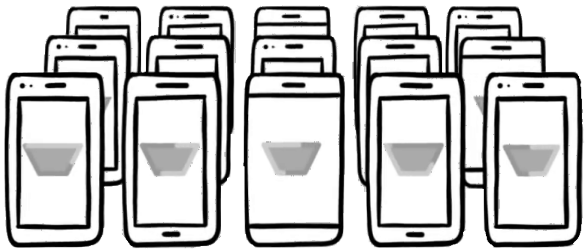
clients **download** the model



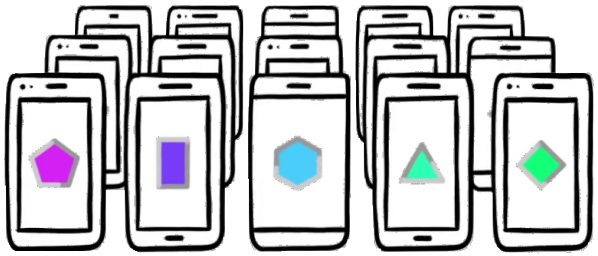
clients update **local** models

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

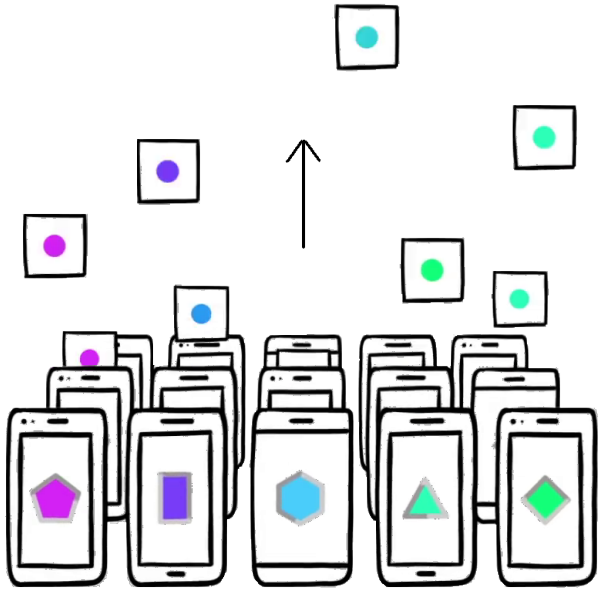
Background: Federated Averaging (FedAvg)



clients **download** the model



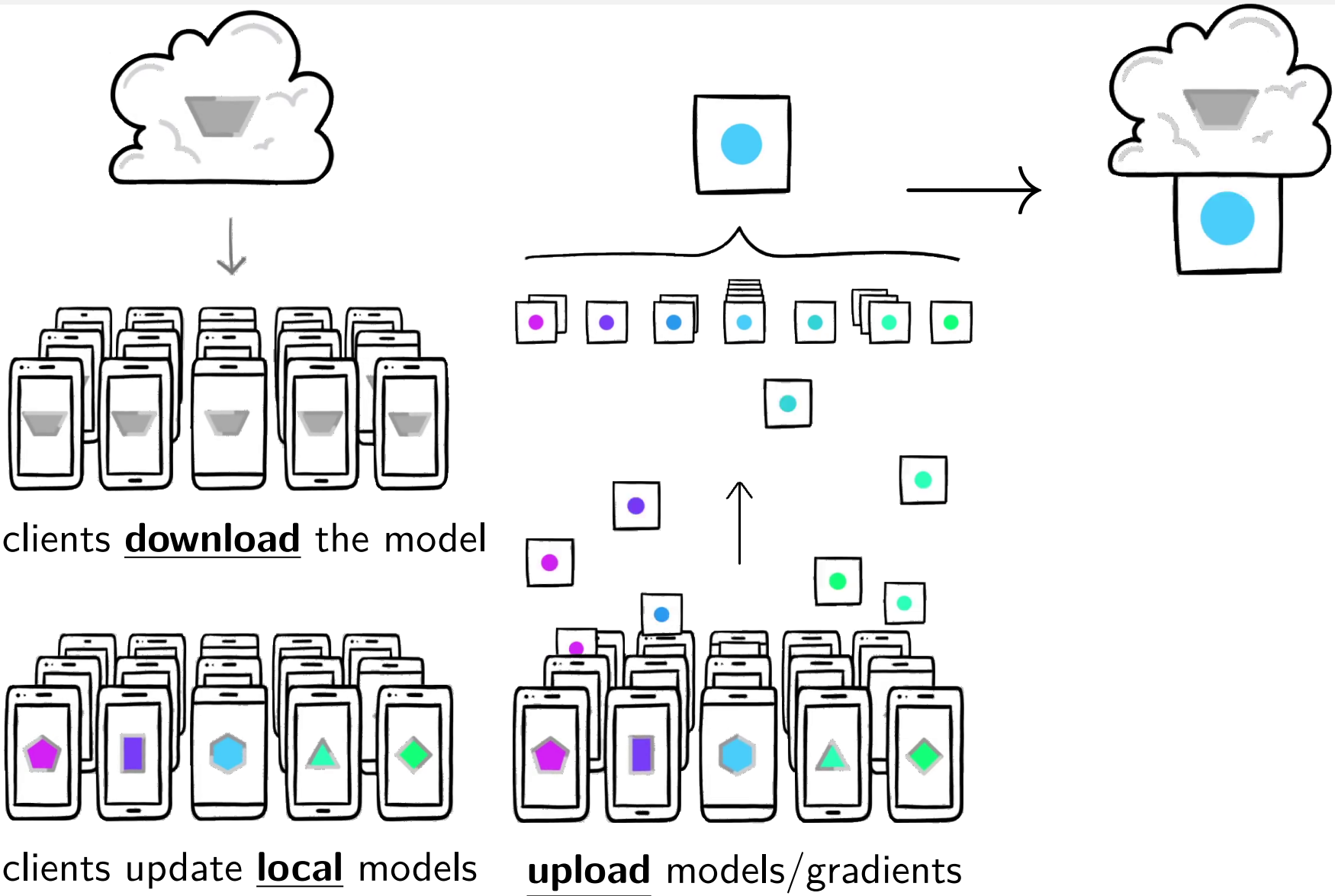
clients update **local** models



upload models/gradients

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Background: Federated Averaging (FedAvg)



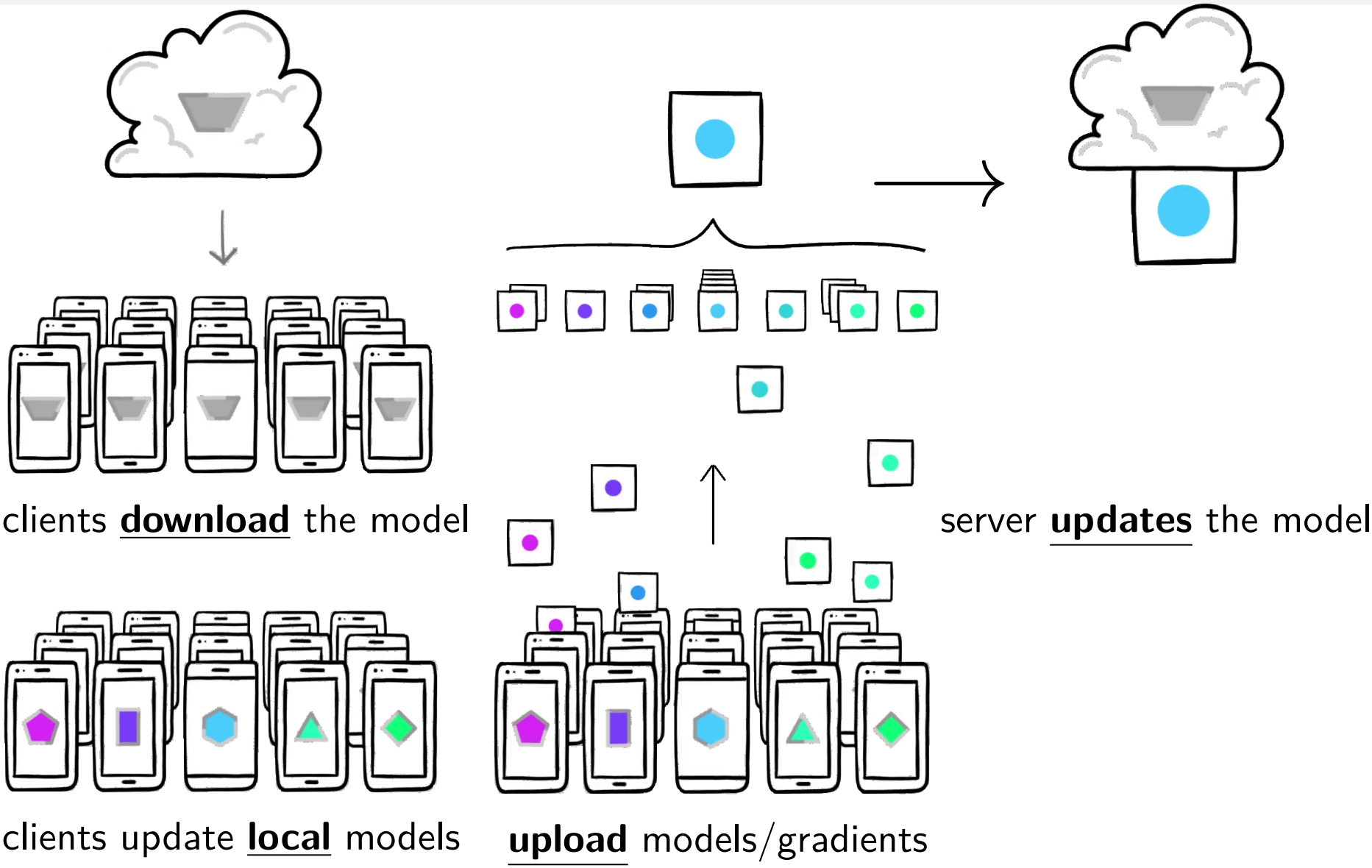
clients **download** the model

clients update **local** models

upload models/gradients

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Background: Federated Averaging (FedAvg)



clients **download** the model

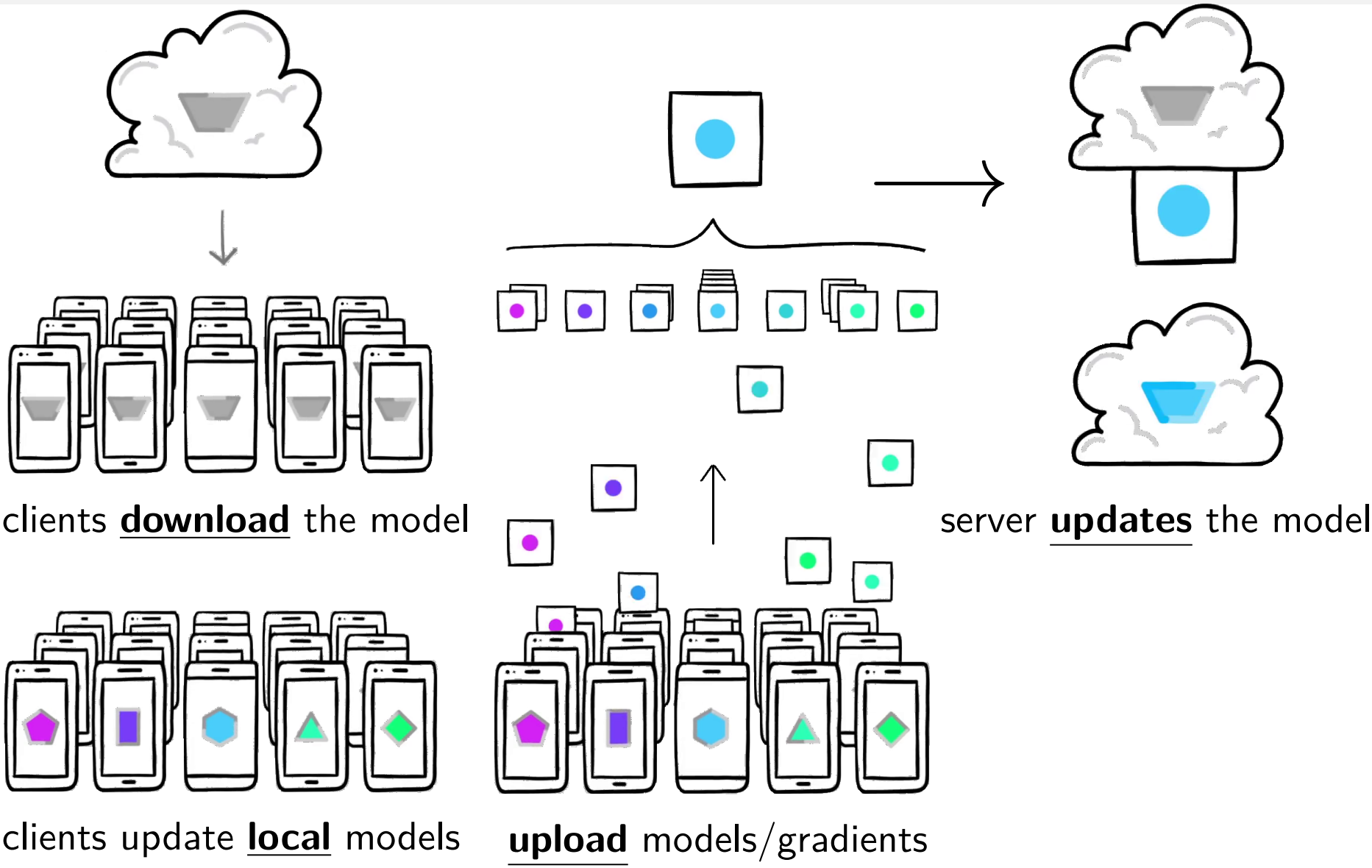
server **updates** the model

clients update **local** models

upload models/gradients

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Background: Federated Averaging (FedAvg)



clients **download** the model

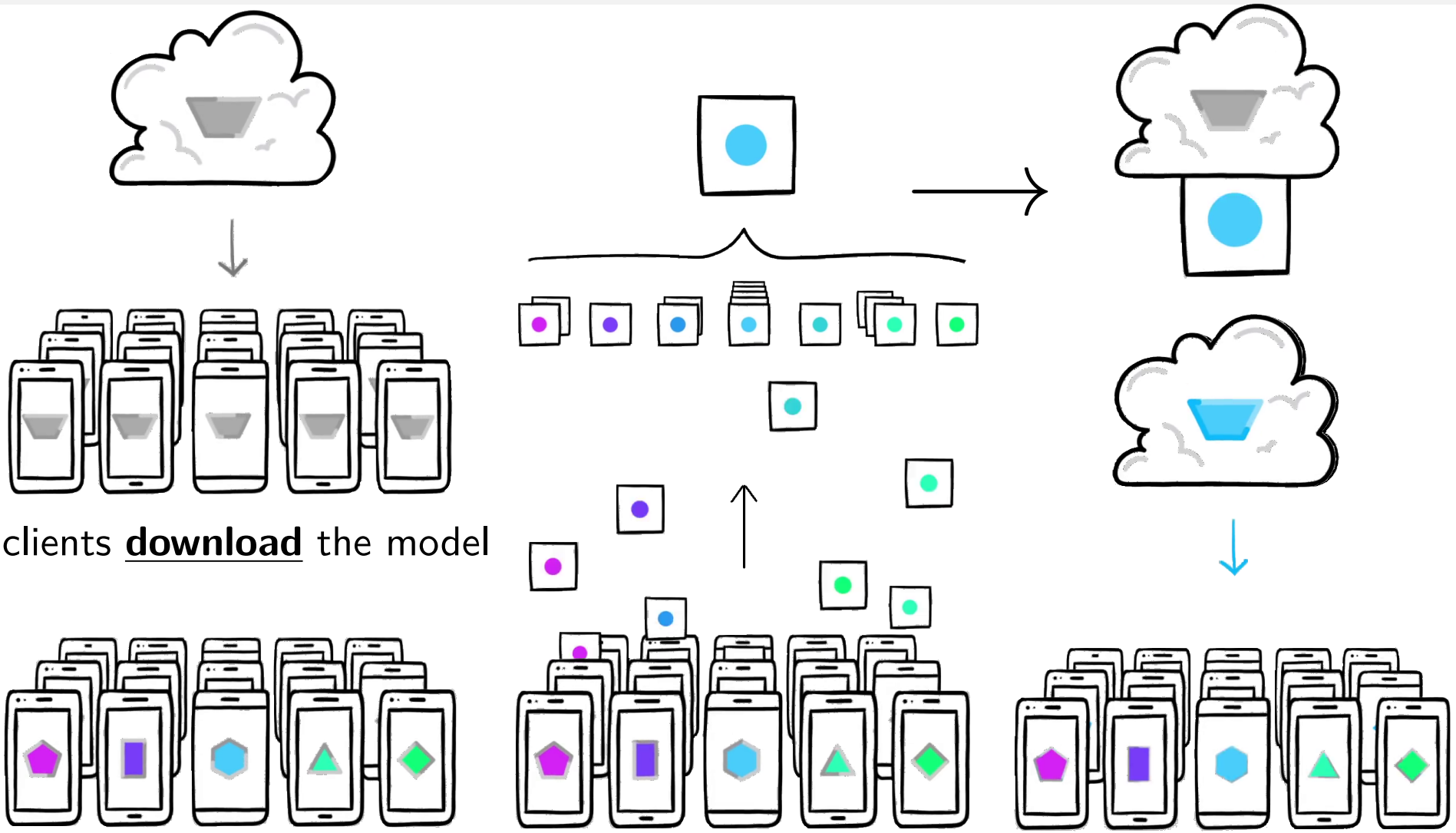
server **updates** the model

clients update **local** models

upload models/gradients

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Background: Federated Averaging (FedAvg)



clients **download** the model

clients update **local** models

upload models/gradients

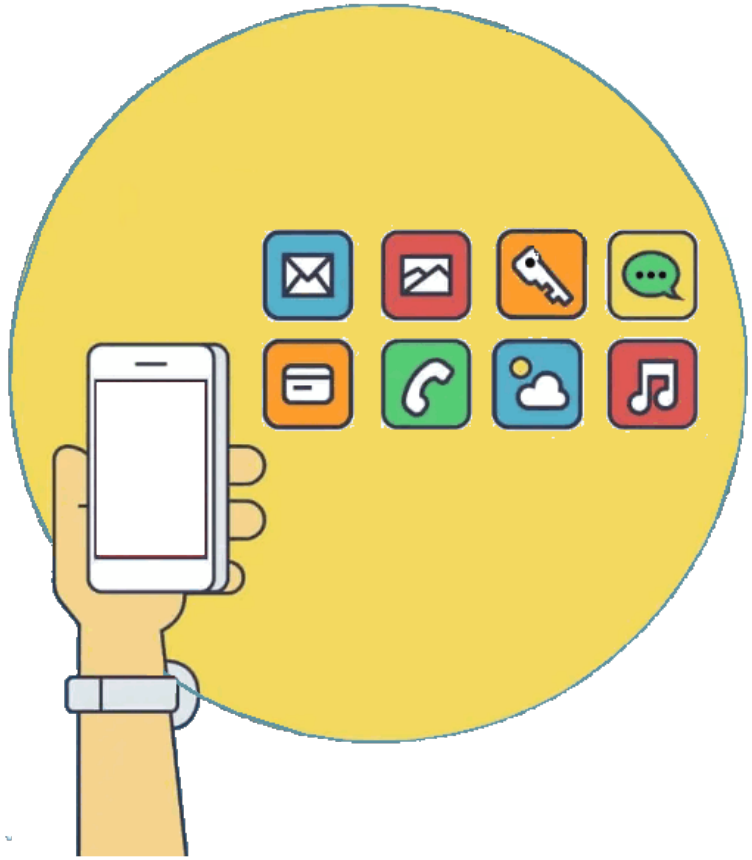
server **deploys** the model

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Background: Privacy Leakage in FL

- Can privacy be leaked in federated learning?

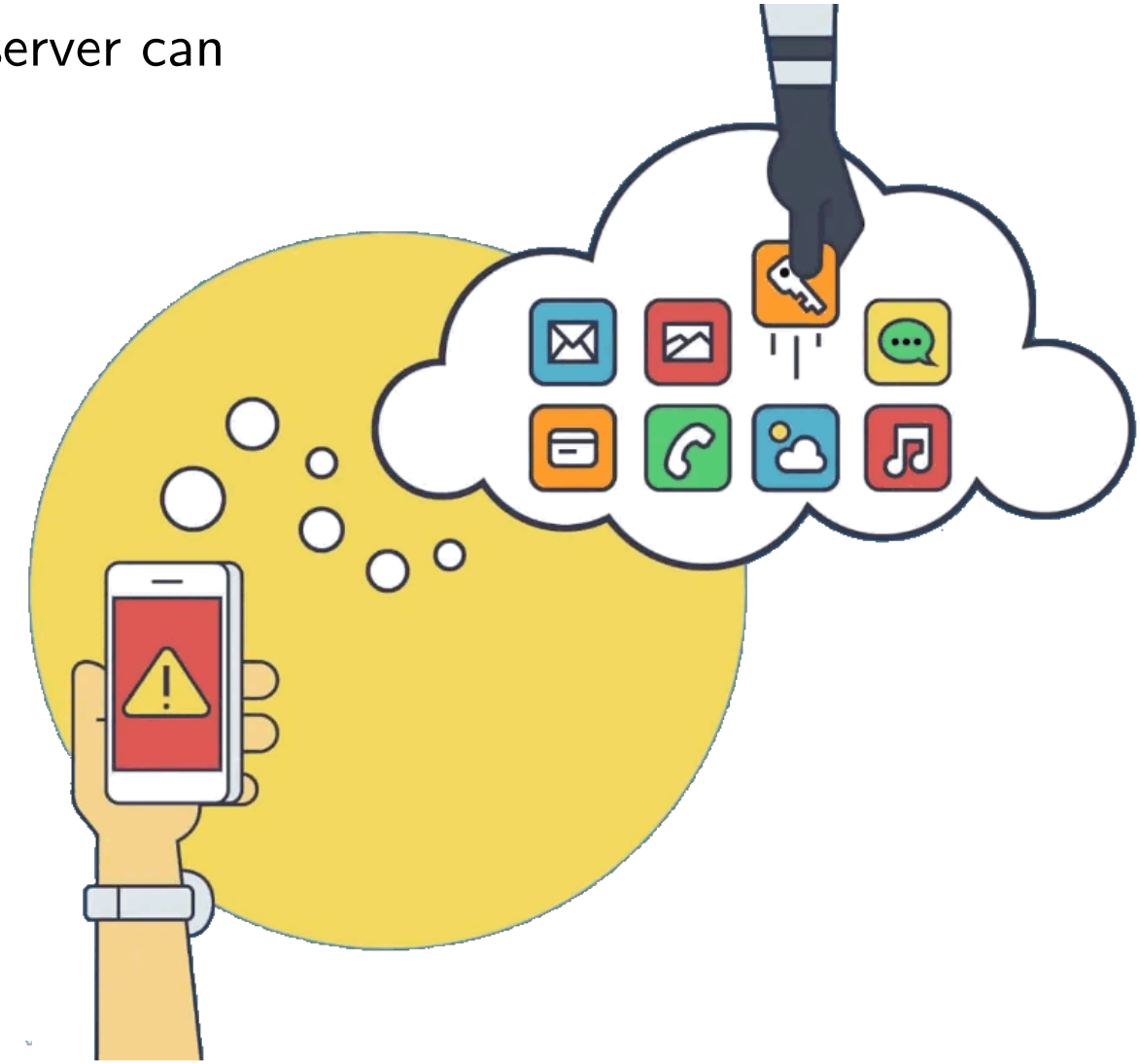
client: my data is kept locally



Background: Privacy Leakage in FL

- An honest-but-curious server can recover private data[2]

client: my data has been stolen!

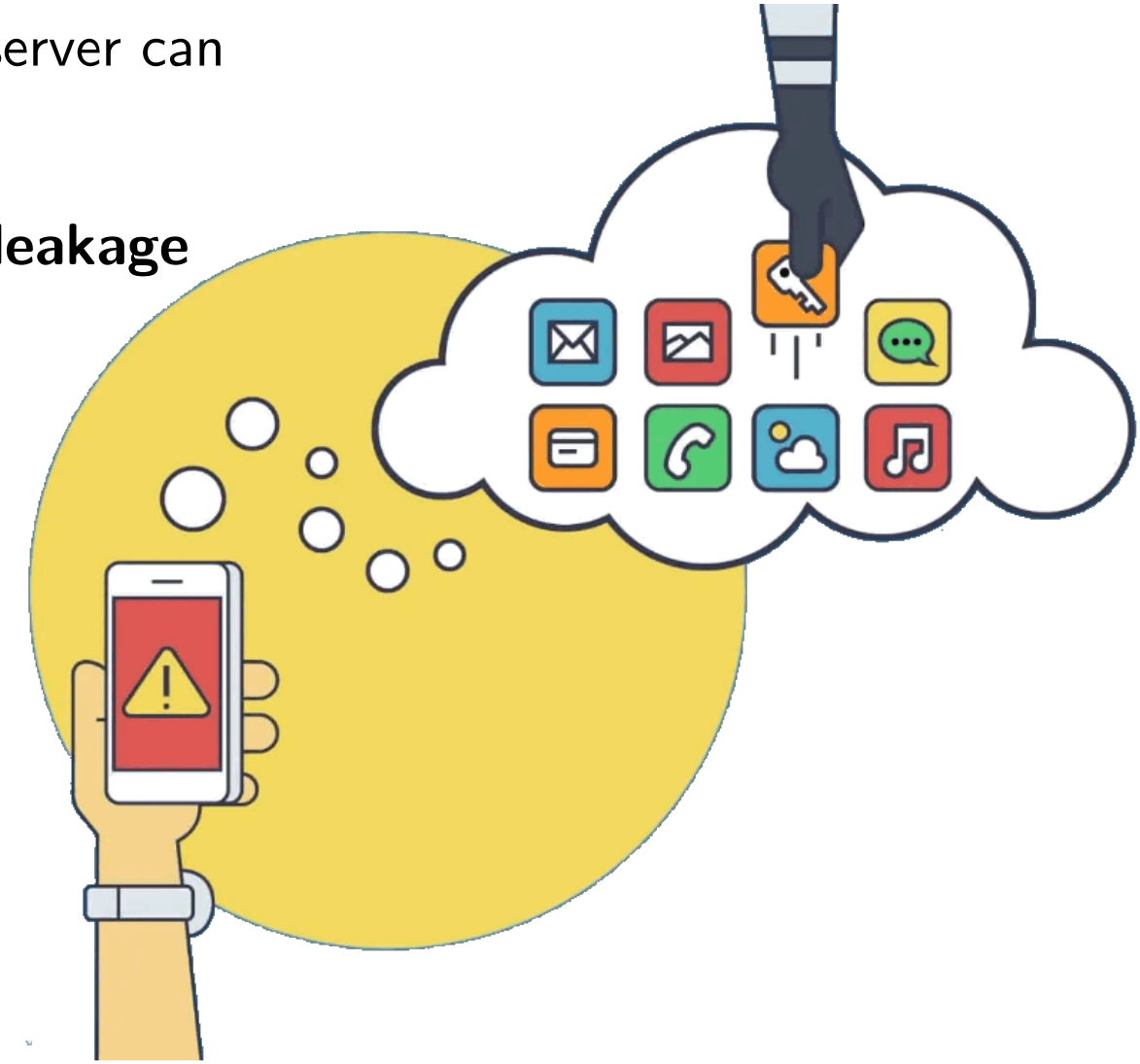


[2] Zhu, Ligeng, et al. "Deep Leakage from Gradients." Advances in Neural Information Processing Systems, 2019.

Background: Privacy Leakage in FL

- An honest-but-curious server can recover private data[2]
- **How realistic is data leakage under obfuscation?**

client: my data has been stolen!



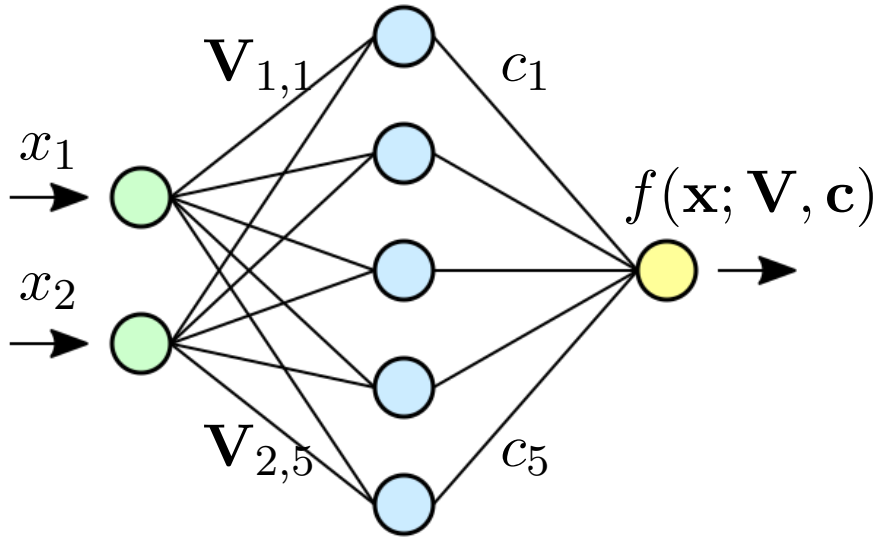
[2] Zhu, Ligeng, et al. "Deep Leakage from Gradients." Advances in Neural Information Processing Systems, 2019.

Index

Raw Data Can be Reconstructed!

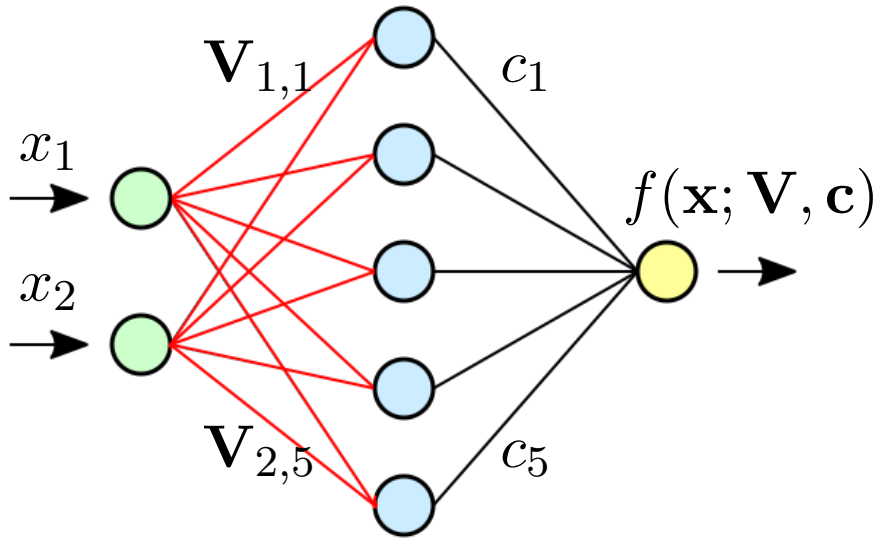
- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$

    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$



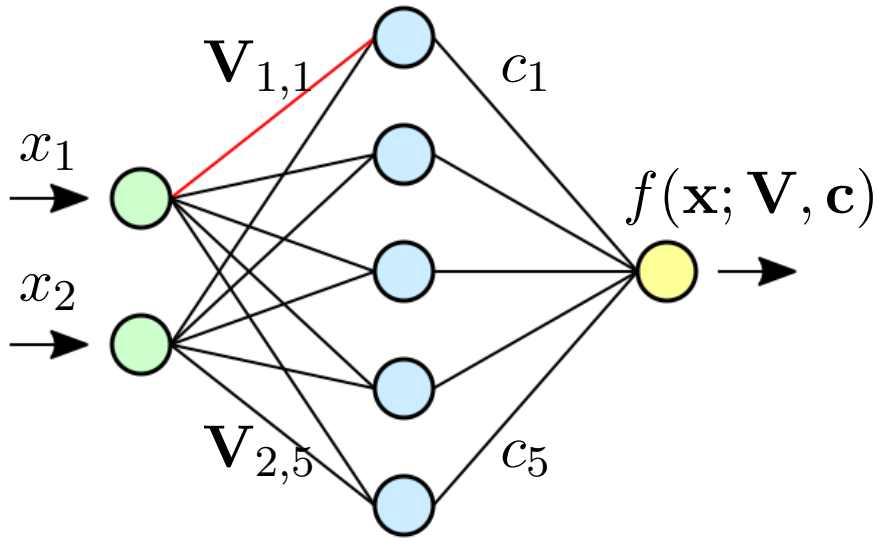
Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations



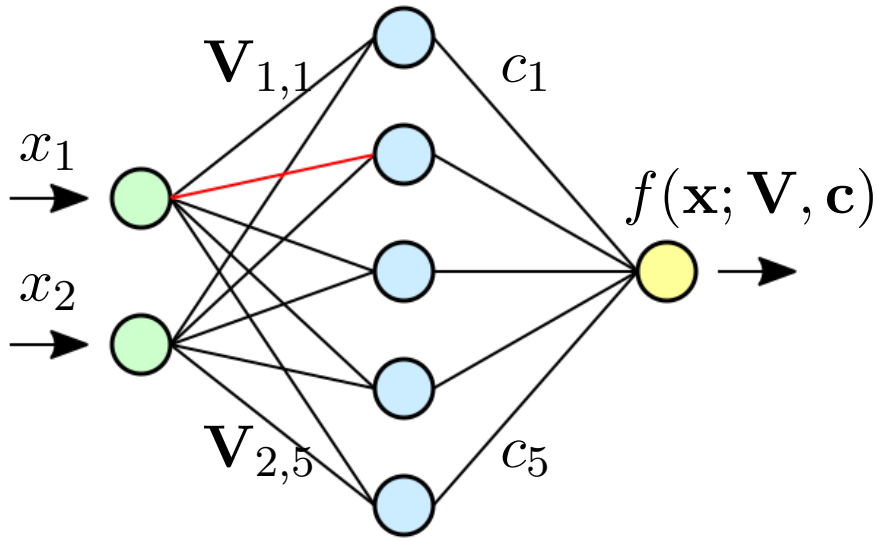
Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations



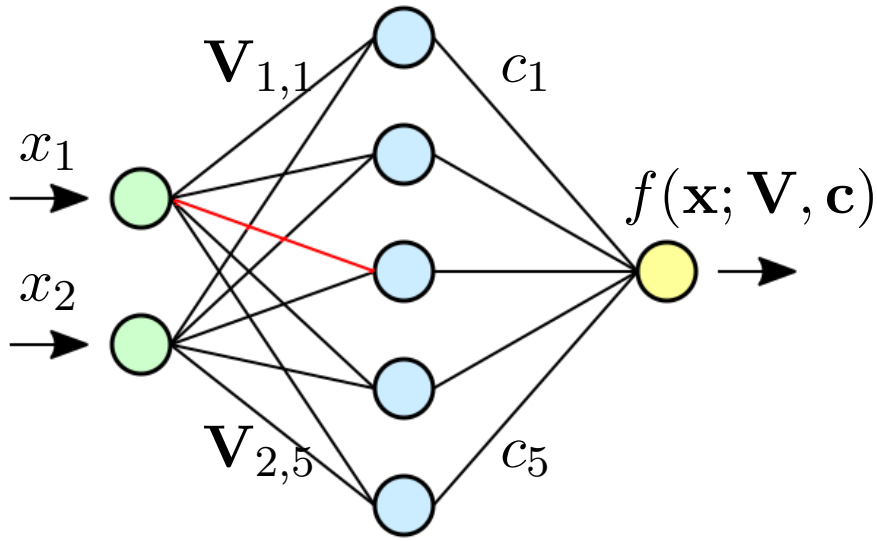
Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations



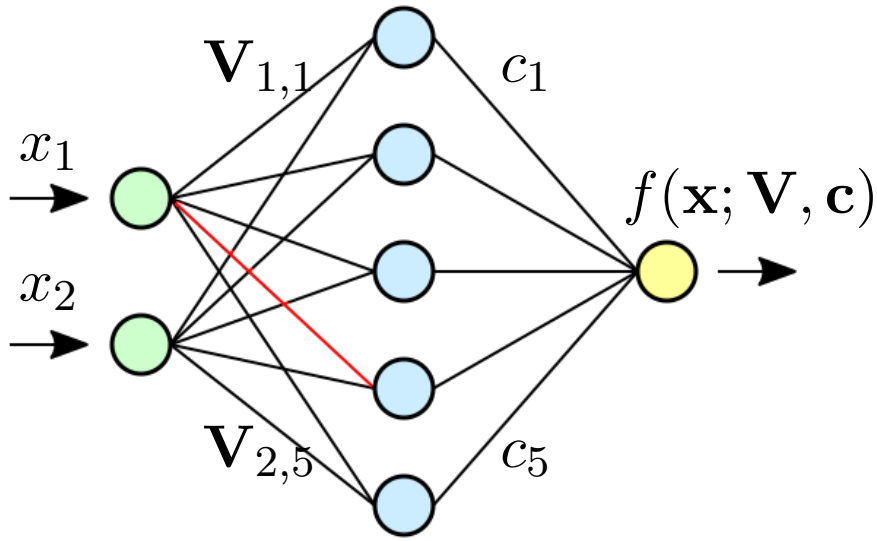
Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations



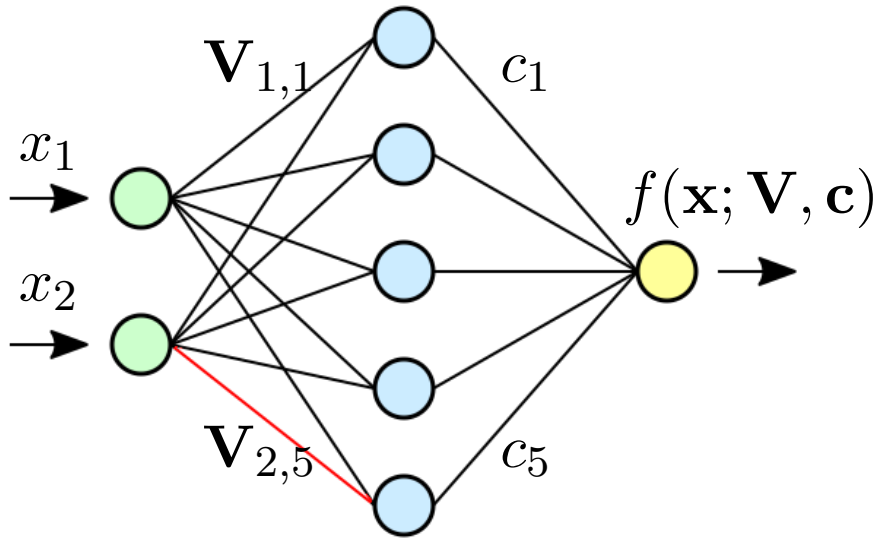
Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations



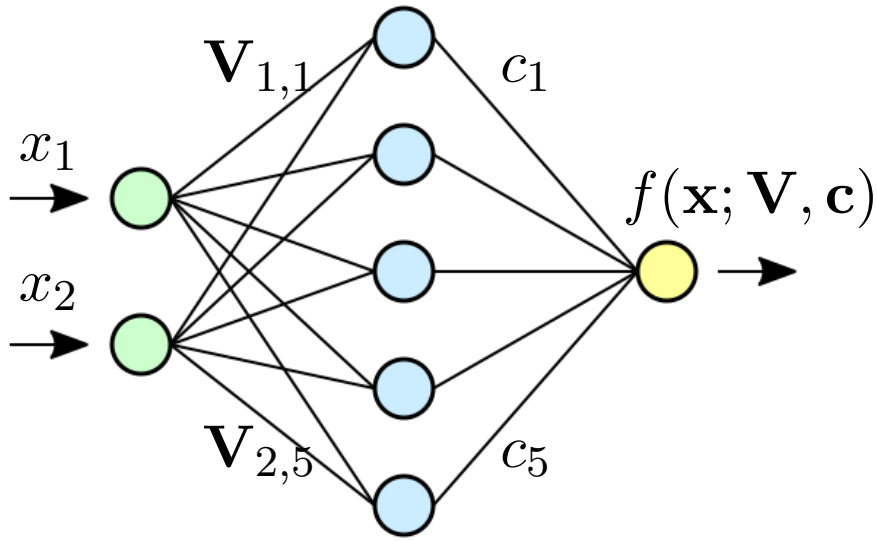
Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations
- Solve an **overdetermined** system



Raw Data Can be Reconstructed!

- Consider a neural network
  - $f(\mathbf{x}; \mathbf{V}, \mathbf{c}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} c_r \sigma(\mathbf{v}_r^\top \mathbf{x})$
  - input $\mathbf{x} \in \mathbb{R}^2$
    weight $\mathbf{V} \in \mathbb{R}^{2\times 5}$
- System of equations
  - 2 unknowns
  - 2×5 equations
- Solve an **overdetermined** system



ROG: Reconstruct from Obfuscated Gradient

client side

private data $\mathbf{X}$

- local optimization
- obtain gradient Δ
- $\widetilde{\Delta} = \phi(\Delta)$

ROG: Reconstruct from Obfuscated Gradient

client side | **server side**

private data **X**

- local optimization
- obtain gradient Δ
- $\widetilde{\Delta} = \phi(\Delta)$

upload →

ROG: Reconstruct from Obfuscated Gradient

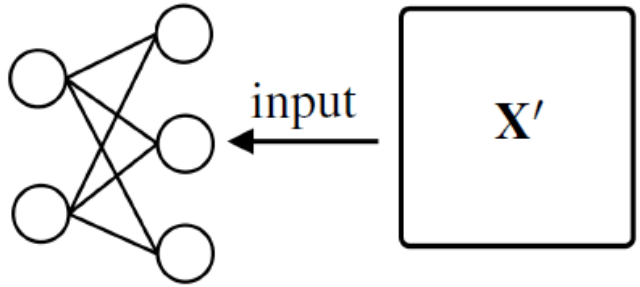
client side | **server side**

private data **X**

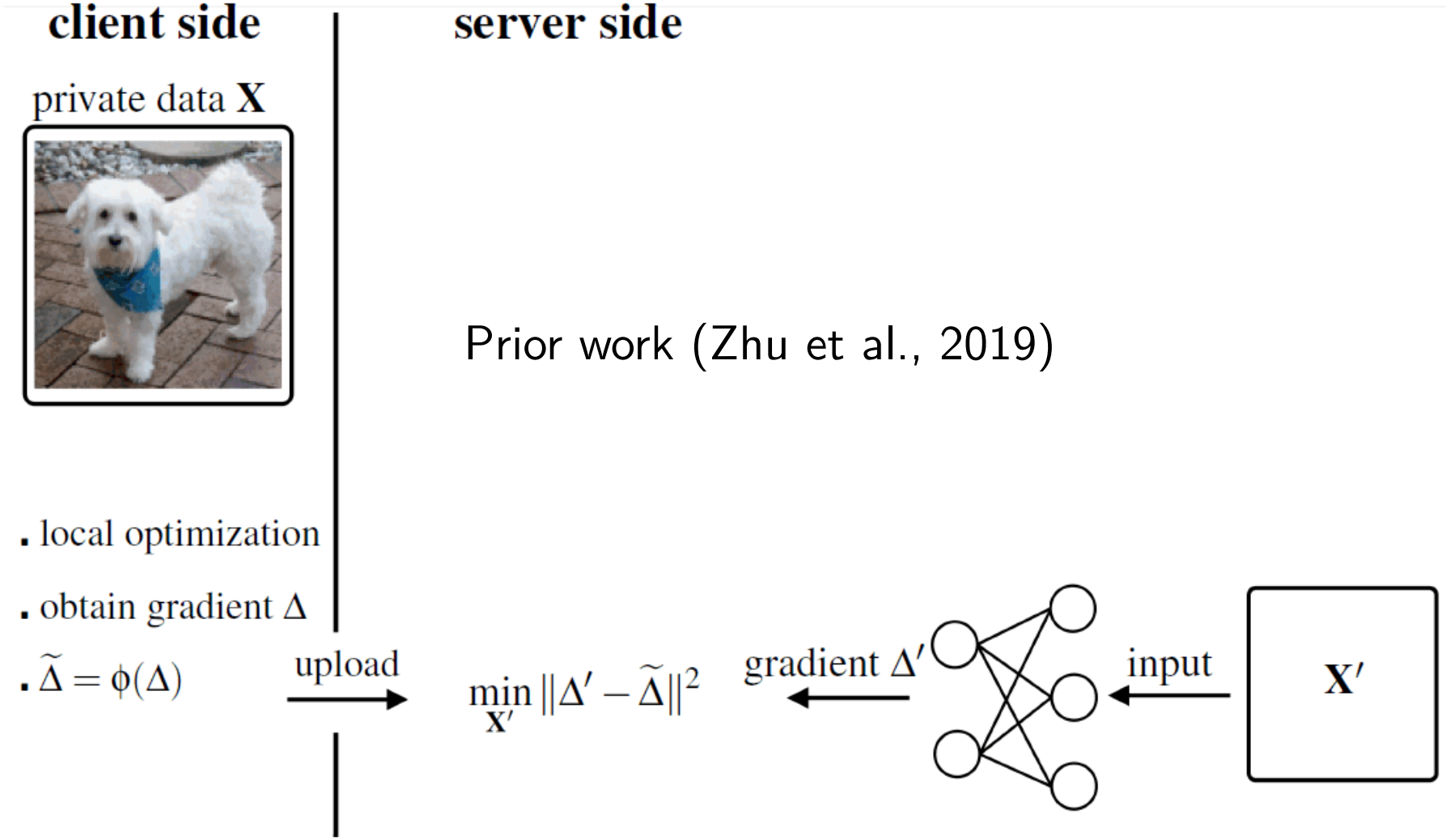
- local optimization
- obtain gradient Δ
- $\widetilde{\Delta} = \phi(\Delta)$

upload →

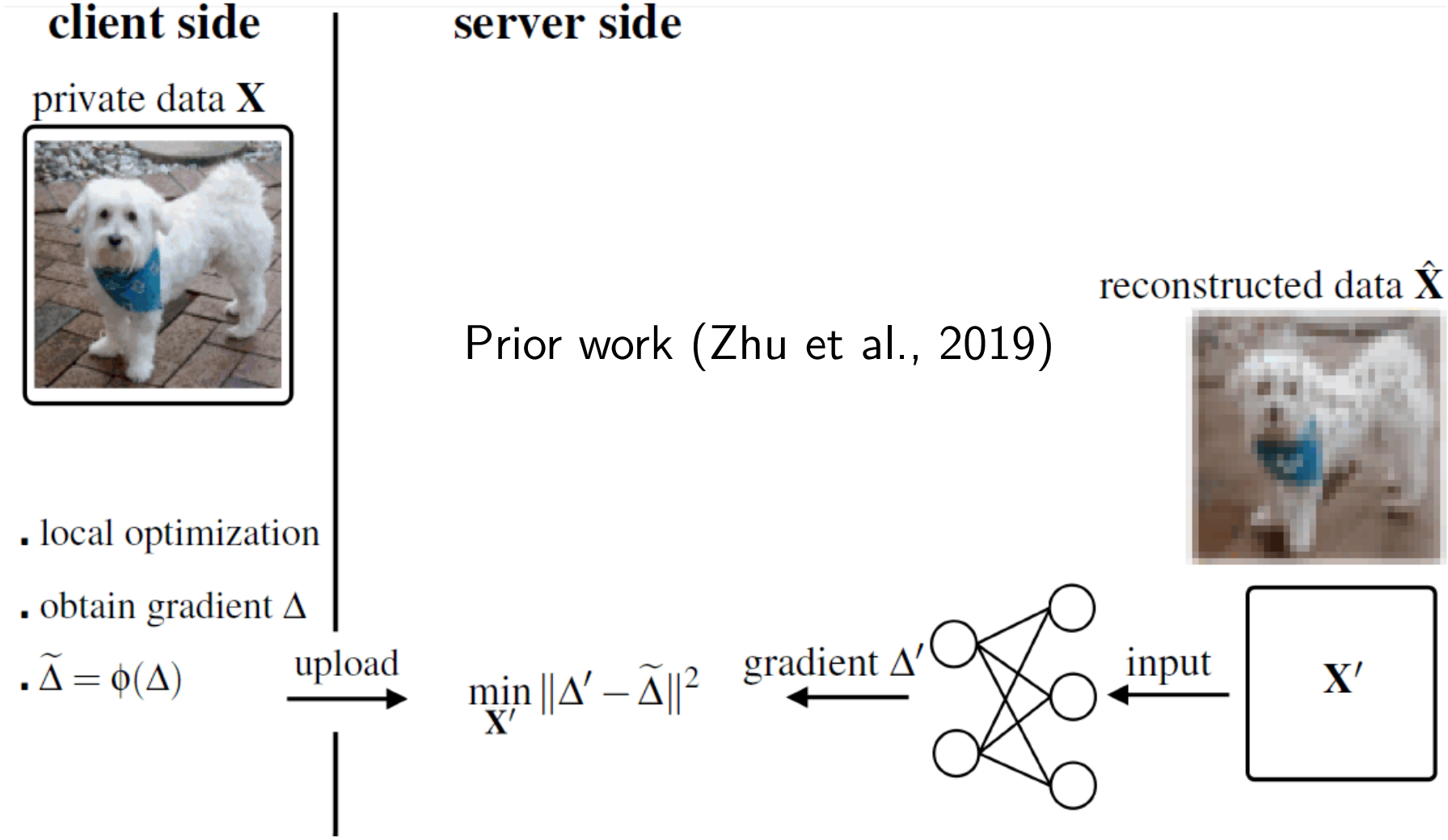
Prior work (Zhu et al., 2019)



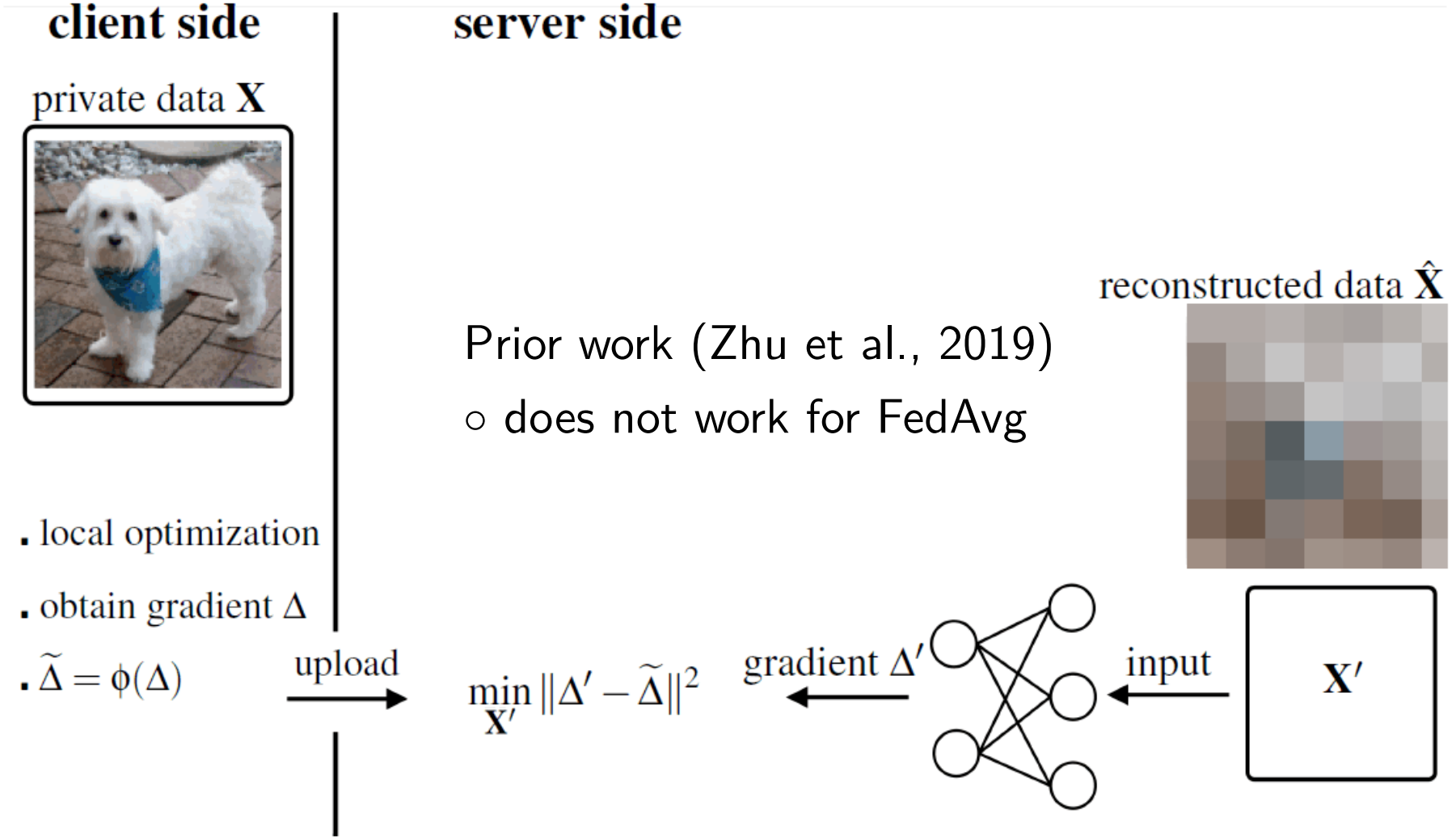
ROG: Reconstruct from Obfuscated Gradient



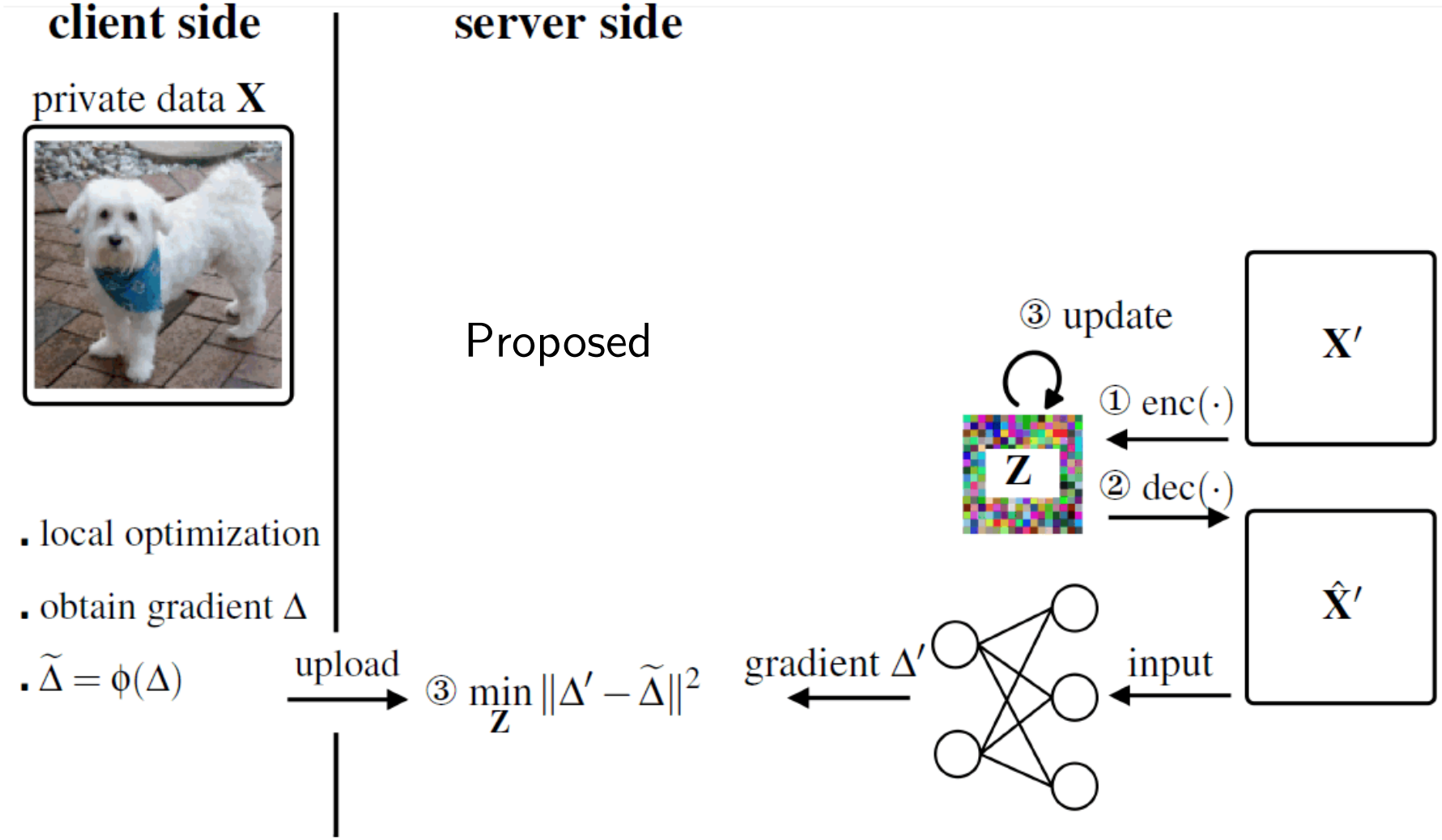
ROG: Reconstruct from Obfuscated Gradient



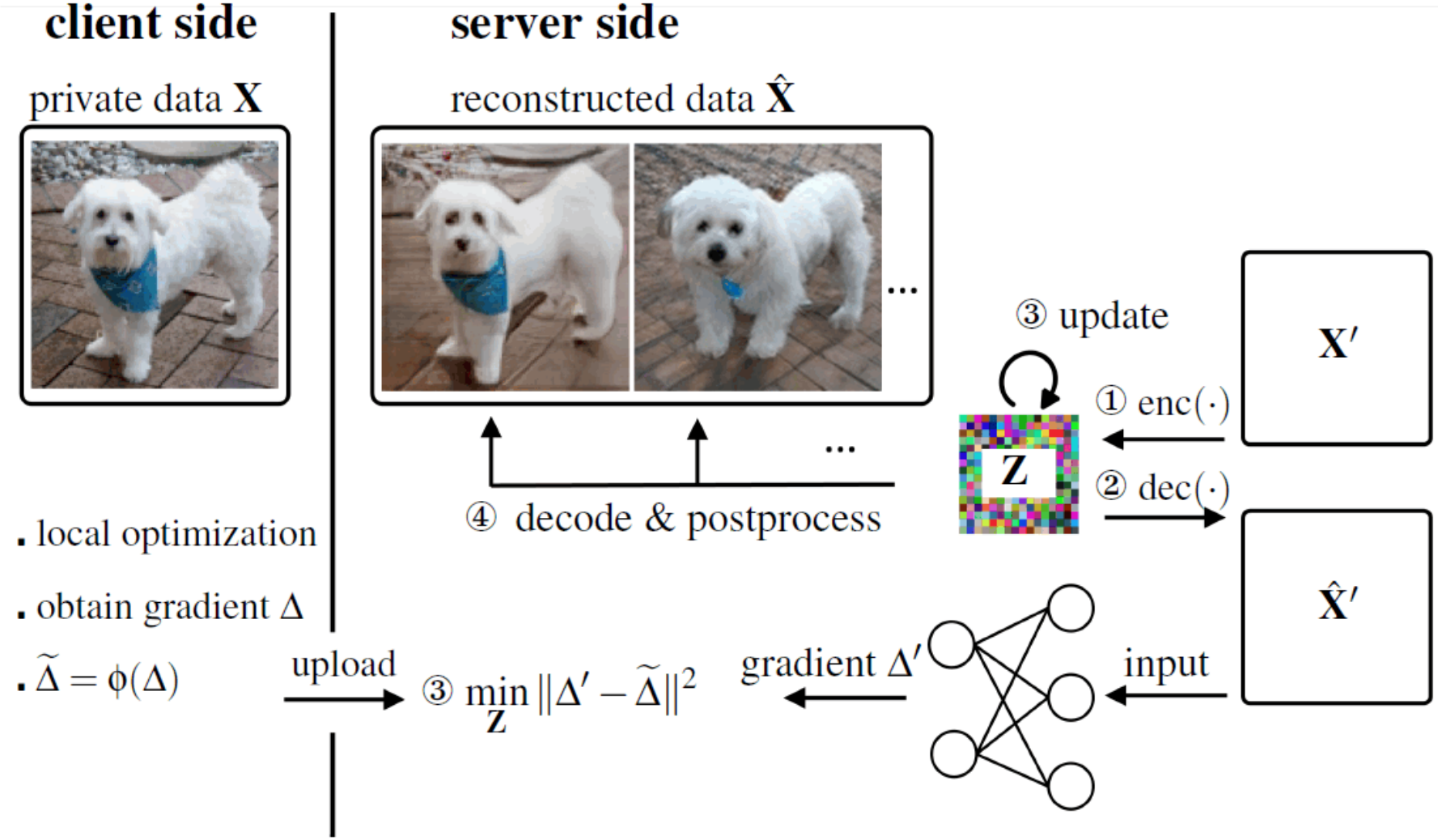
ROG: Reconstruct from Obfuscated Gradient



ROG: Reconstruct from Obfuscated Gradient



ROG: Reconstruct from Obfuscated Gradient



Index

Case Study: Does Gradient Compression Imply Privacy?

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

Case Study: Does Gradient Compression Imply Privacy?

- Quantization[3]
  - FP32 $\rightarrow$ 3 bits

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

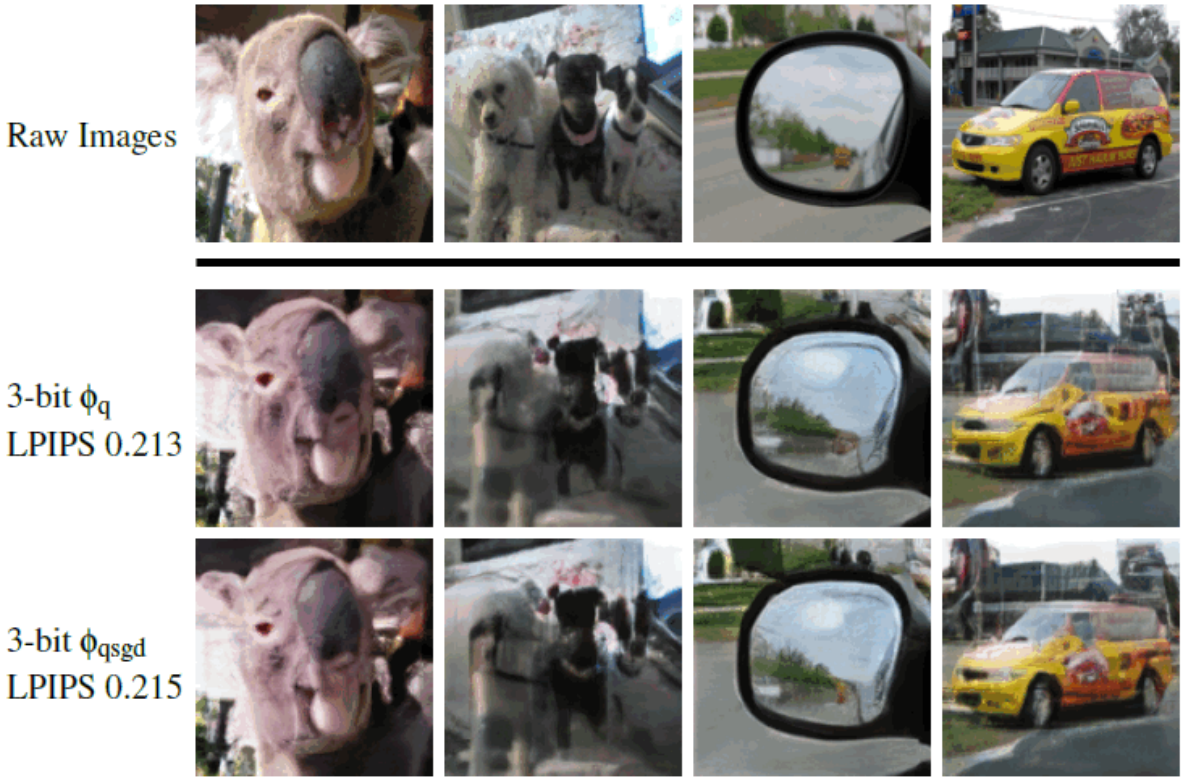
Case Study: Does Gradient Compression Imply Privacy?

- Quantization[3]
  - FP32 $\rightarrow$ 3 bits

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

Case Study: Does Gradient Compression Imply Privacy?

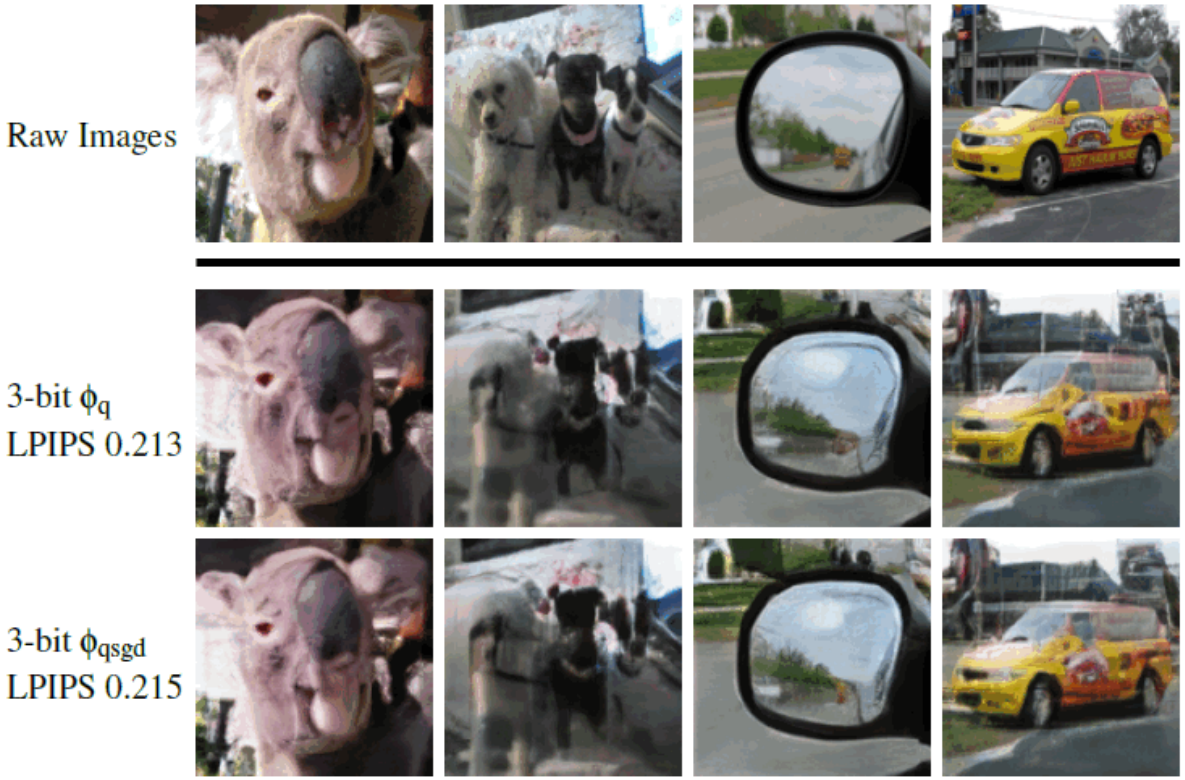


- Quantization[3]
  - FP32 $\rightarrow$ 3 bits

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

Case Study: Does Gradient Compression Imply Privacy?

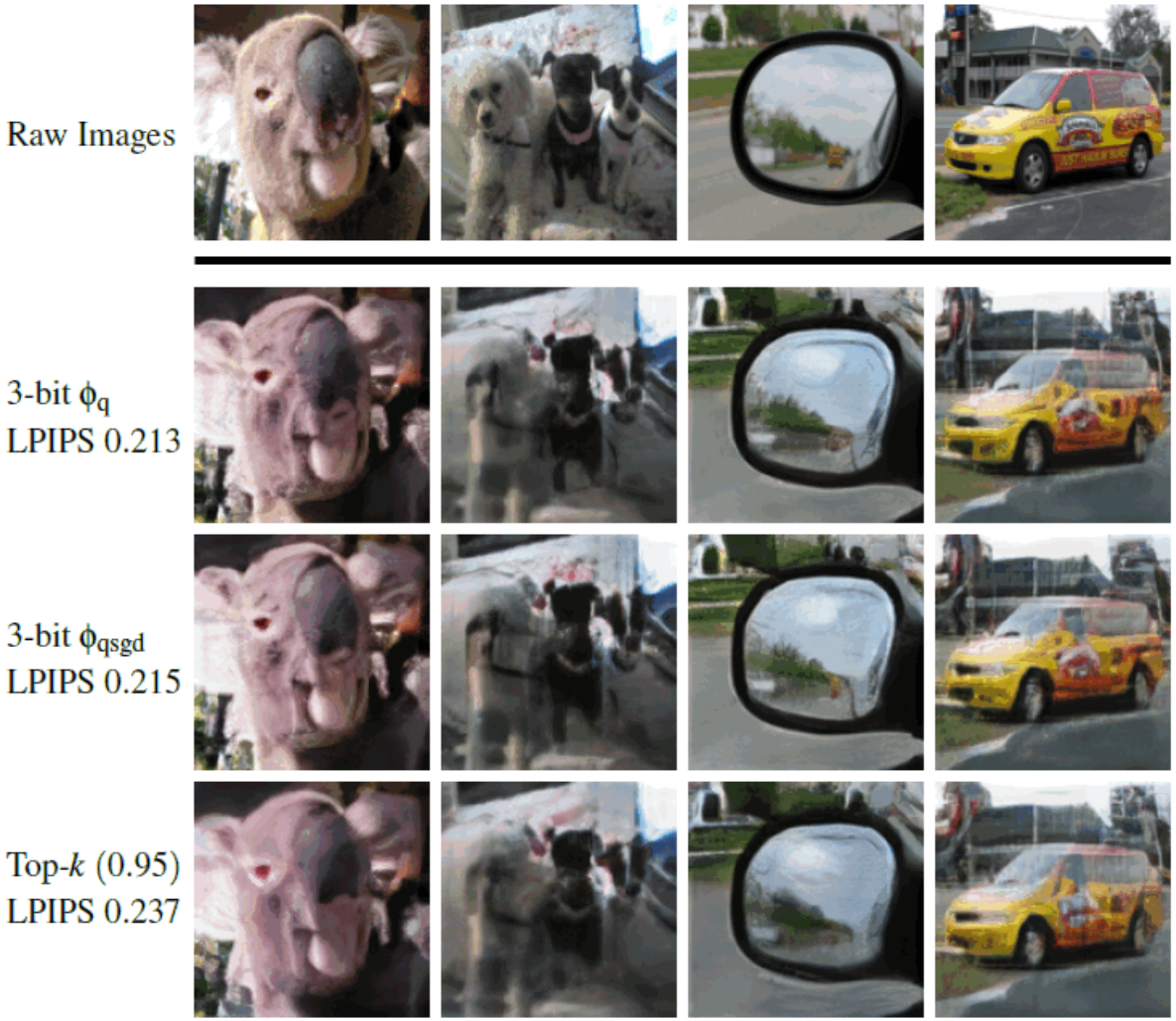


- Quantization[3]
  - FP32 $\rightarrow$ 3 bits
- Gradient sparsification
  - zero out smaller components

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

Case Study: Does Gradient Compression Imply Privacy?

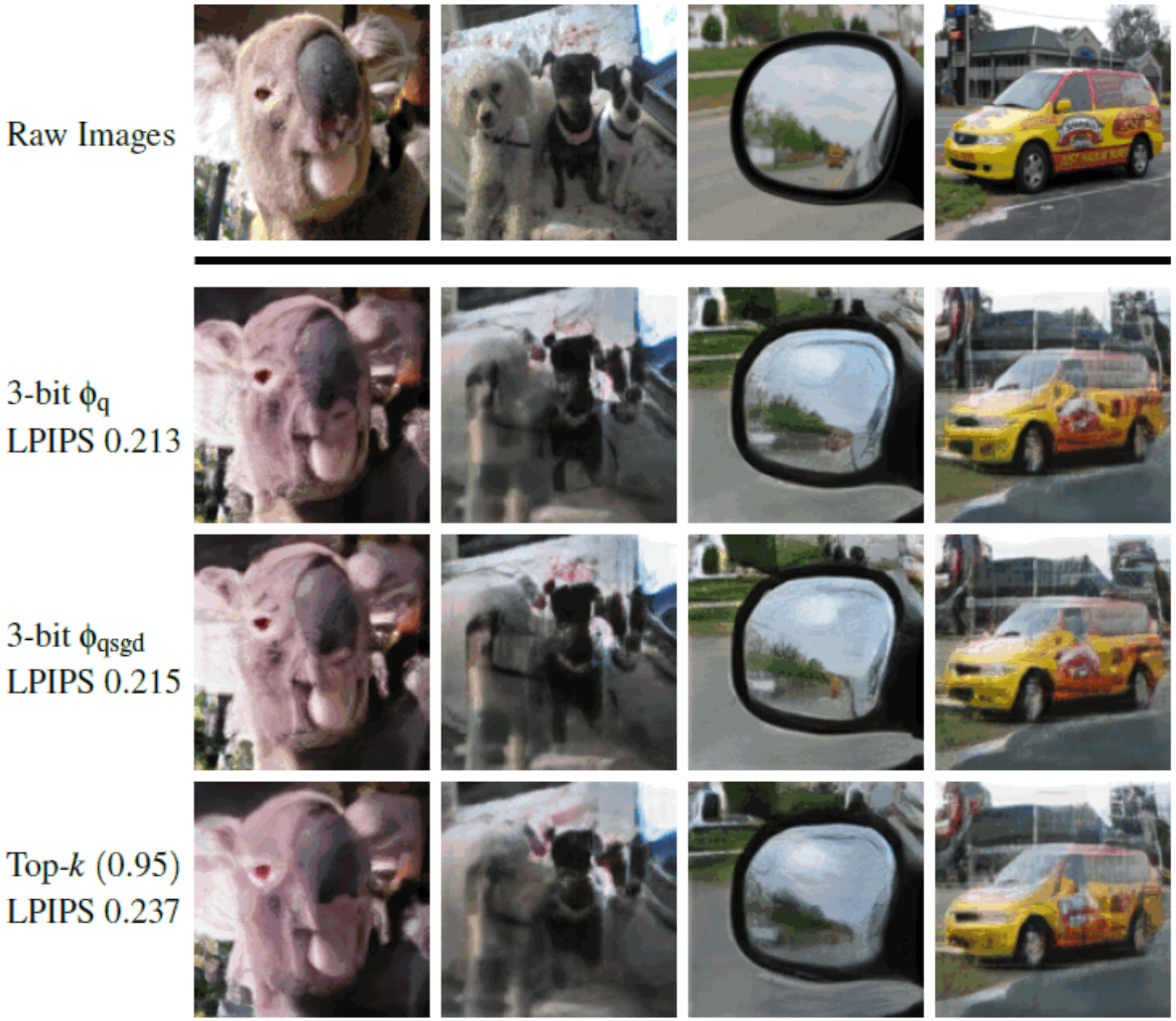


- Quantization[3]
  - FP32 $\rightarrow$ 3 bits
- Gradient sparsification
  - zero out smaller components

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

Case Study: Does Gradient Compression Imply Privacy?



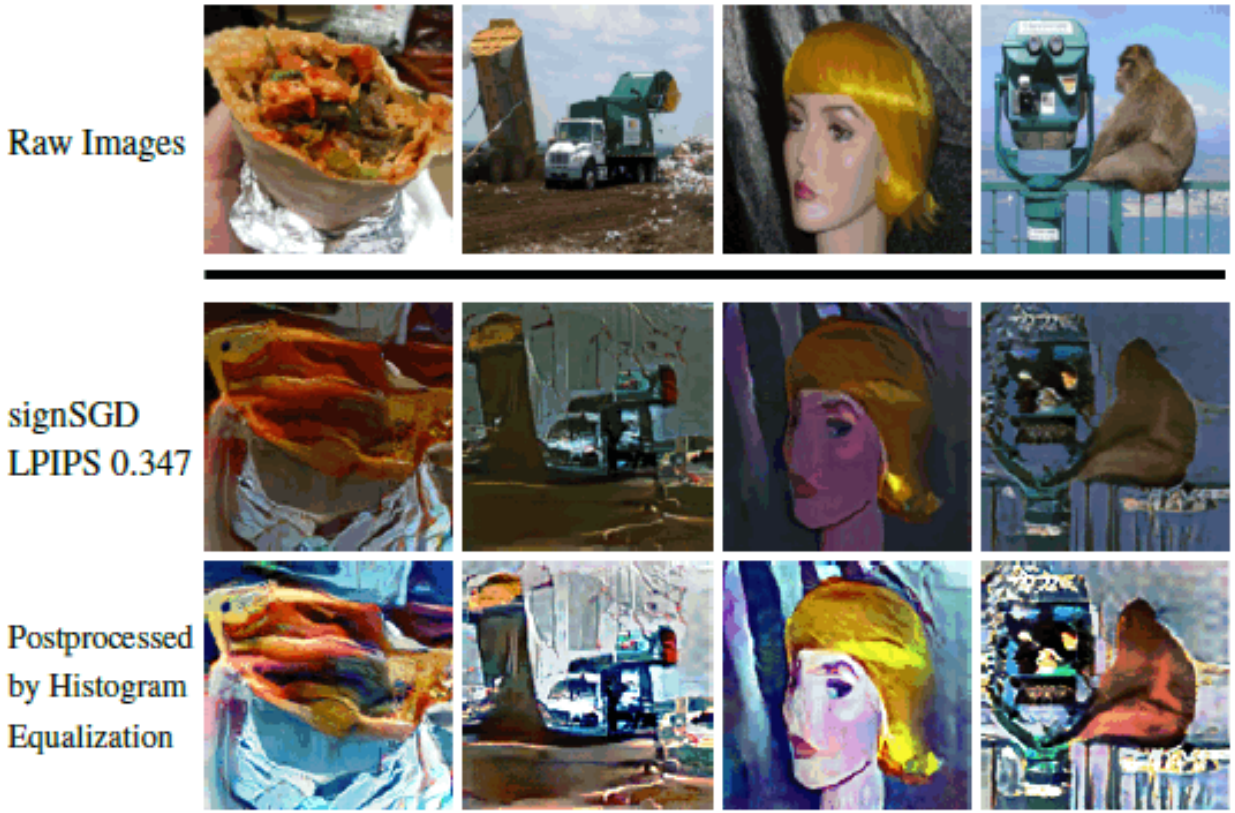
- Quantization[3]
  - FP32 $\rightarrow$ 3 bits
- Gradient sparsification
  - zero out smaller components
- LPIPS[4] (Learned Perceptual Image Patch Similarity) $\downarrow$

[3] Alistarh et al. "The Convergence of Sparsified Gradient Methods." NeurIPS 2018.

[4] Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric." CVPR 2018.

Case Study: Binary Quantization

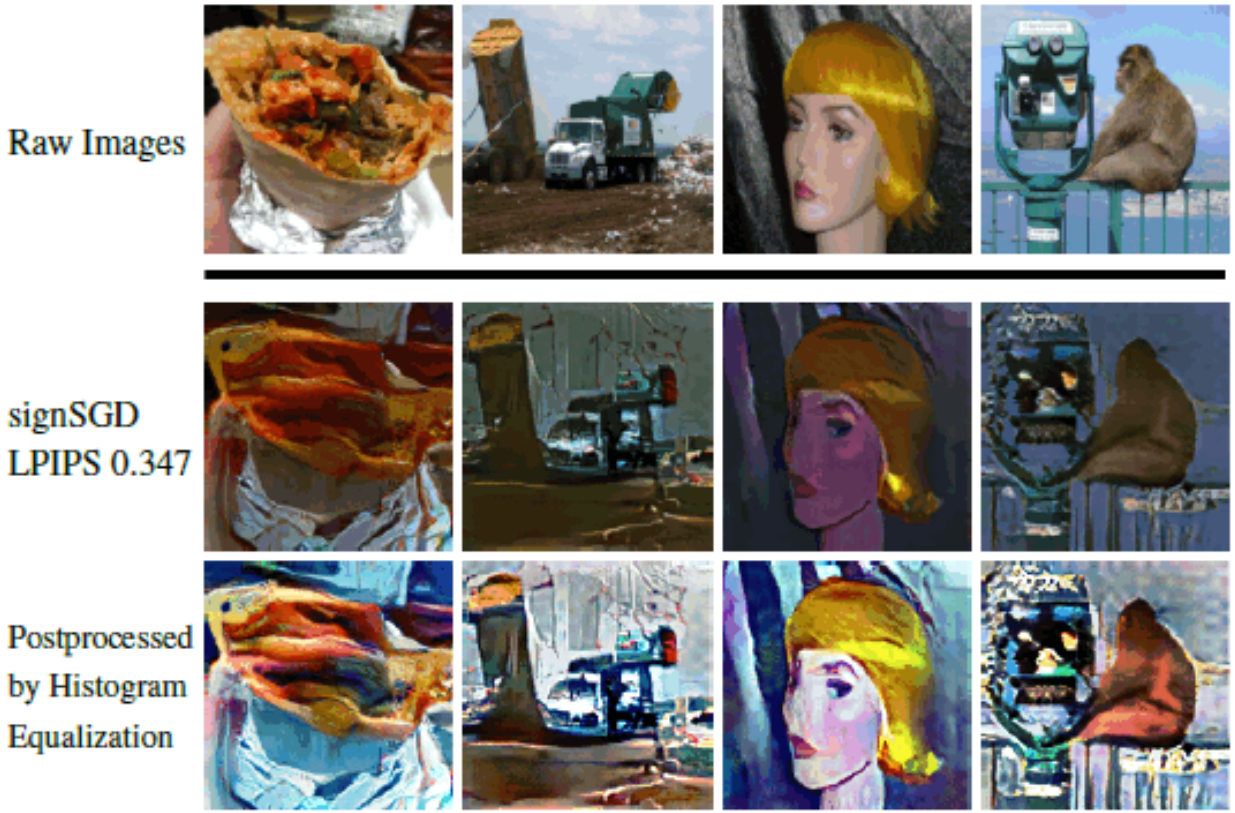
- One bit quantization: signSGD [5]
  - Each entry of the gradient is represented by its sign



[5] Bernstein, Jeremy, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. 2018. "SignSGD: Compressed Optimisation for Non-Convex Problems." ICML, 2018.

Case Study: Binary Quantization

- One bit quantization: signSGD [5]
  - Gradient compression does **not** imply privacy



[5] Bernstein, Jeremy, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. 2018. "SignSGD: Compressed Optimisation for Non-Convex Problems." ICML, 2018.

Defense: Differential Privacy[6]

- Attack differentially private training

Table 1:

Test Accuracy on CIFAR-10

SNR (dB)	Acc. (%)	LPIPS
∞	69	0.18

[6] Wei et al. "Gradient-leakage resilient federated learning." ICDCS 2021.

Defense: Differential Privacy[6]

- Attack differentially private training

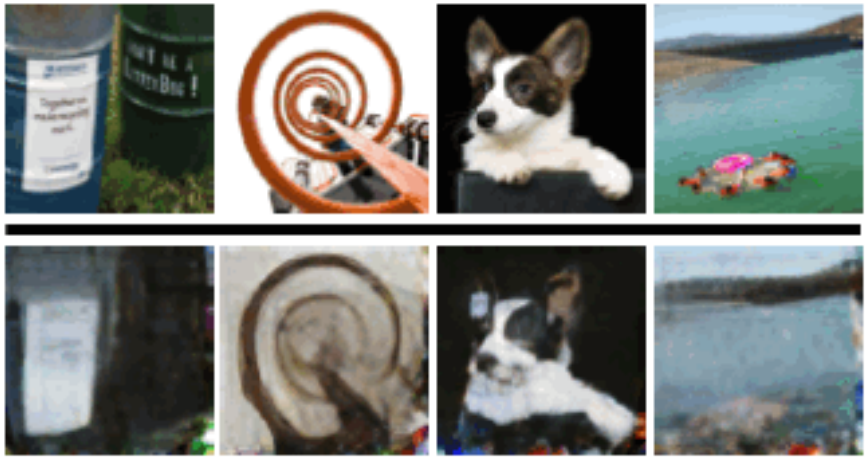


Table 1:

Tradeoff Between Utility & Privacy

SNR (dB)	Acc. (%)	LPIPS
∞	69	0.18
0	55	0.38

[6] Wei et al. "Gradient-leakage resilient federated learning." ICDCS 2021.

Defense: Differential Privacy[6]

- Attack differentially private training

Table 1:

Tradeoff Between Utility & Privacy

SNR (dB)	Acc. (%)	LPIPS
∞	69	0.18
0	55	0.38
-5	52	0.46

[6] Wei et al. "Gradient-leakage resilient federated learning." ICDCS 2021.

Defense: Differential Privacy[6]

- Attack differentially private training

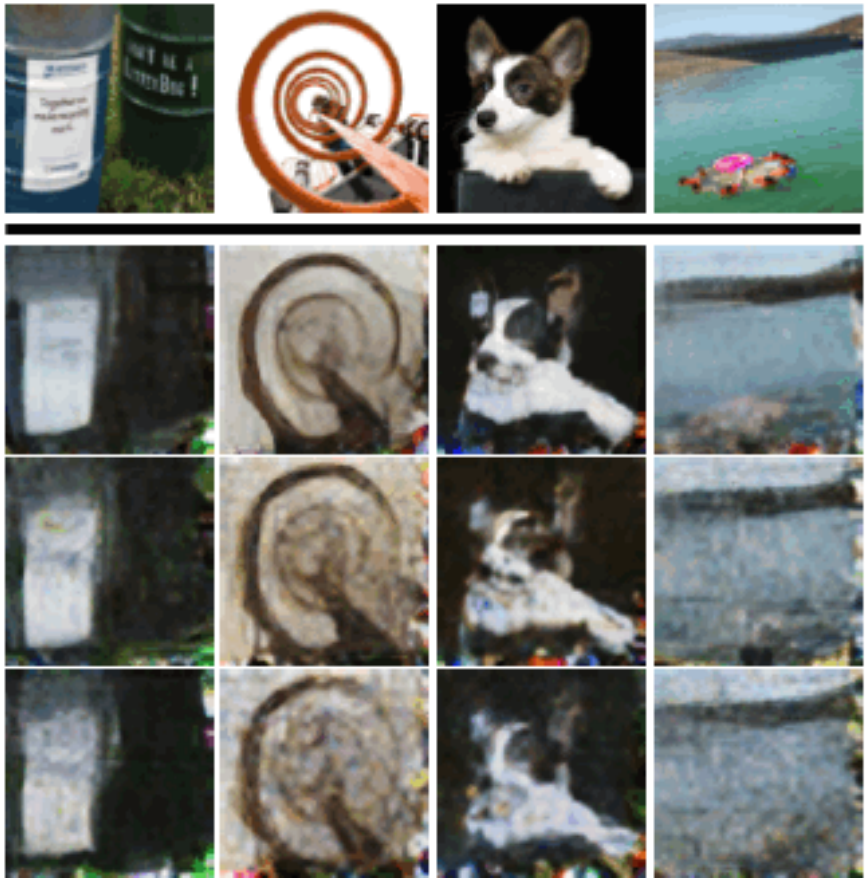


Table 1:

Tradeoff Between Utility & Privacy

SNR (dB)	Acc. (%)	LPIPS
∞	69	0.18
0	55	0.38
-5	52	0.46
-10	50	0.55

[6] Wei et al. "Gradient-leakage resilient federated learning." ICDCS 2021.

Defense: Differential Privacy[6]

- Attack differentially private training

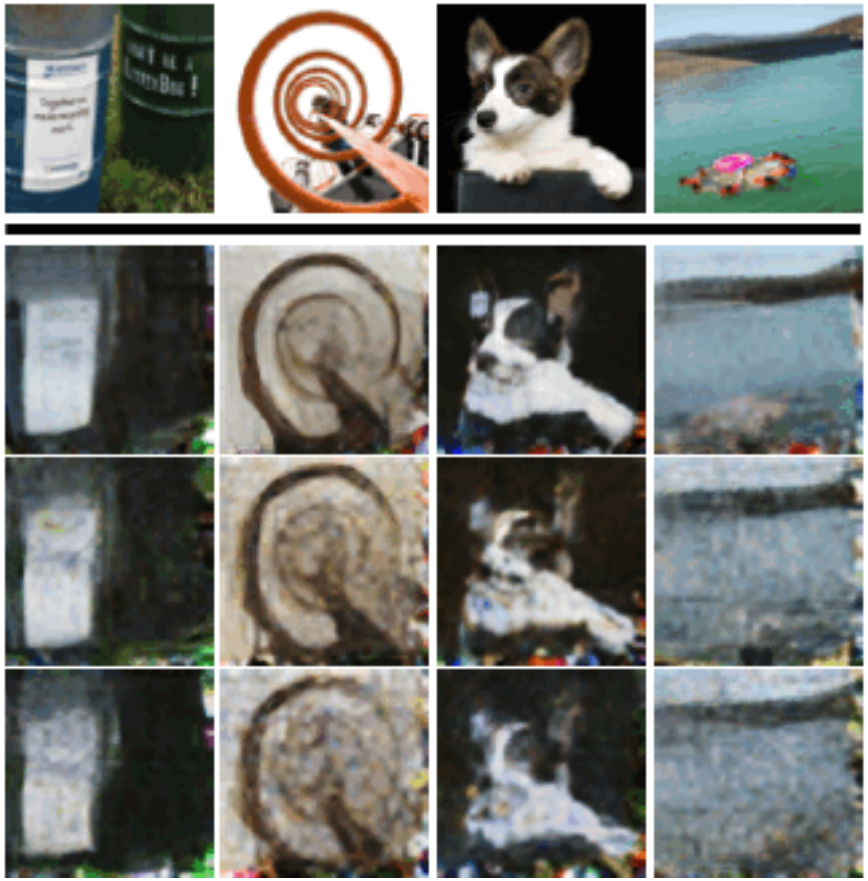


Table 1:

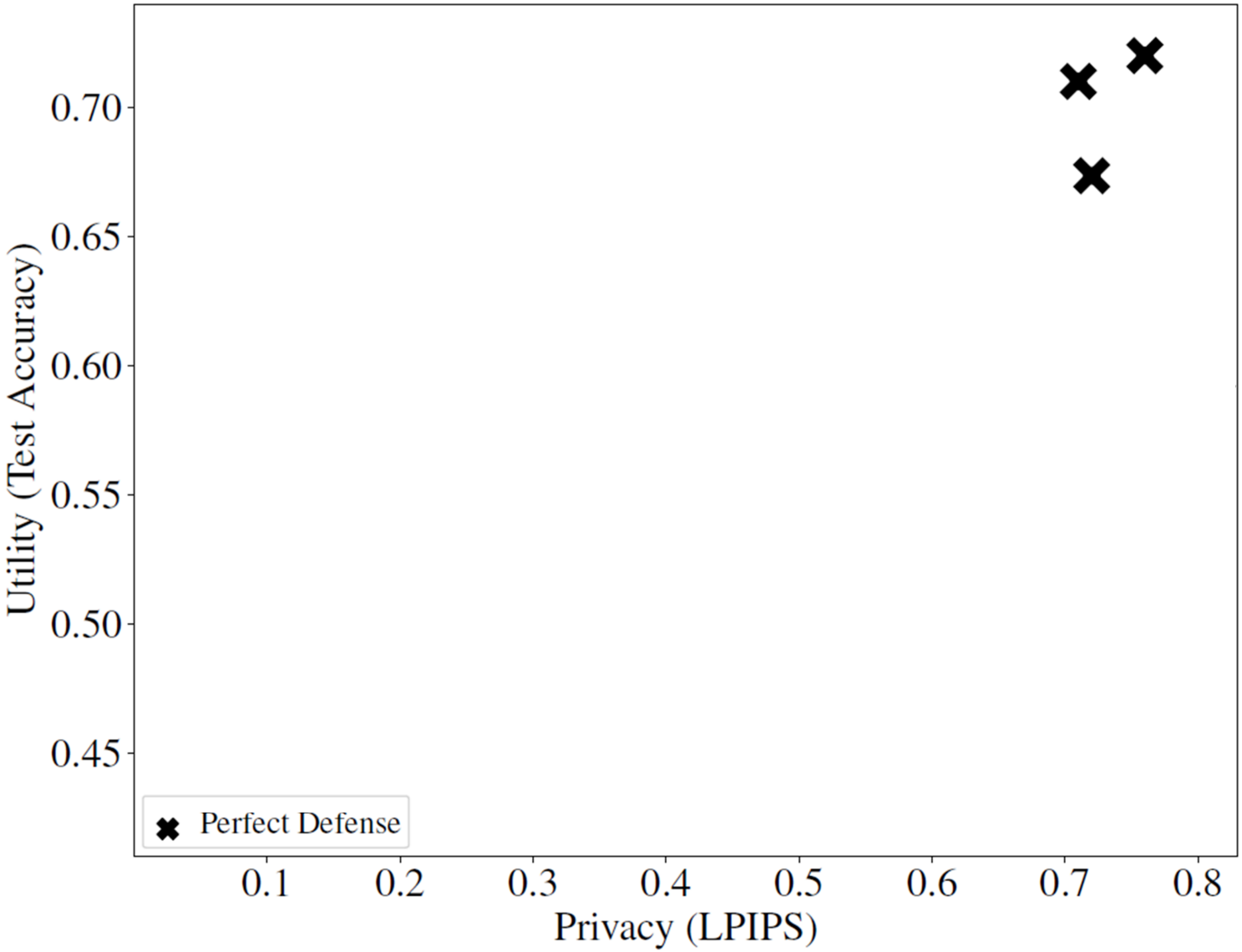
Tradeoff Between Utility & Privacy

SNR (dB)	Acc. (%)	LPIPS
∞	69	0.18
0	55	0.38
-5	52	0.46
-10	50	0.55

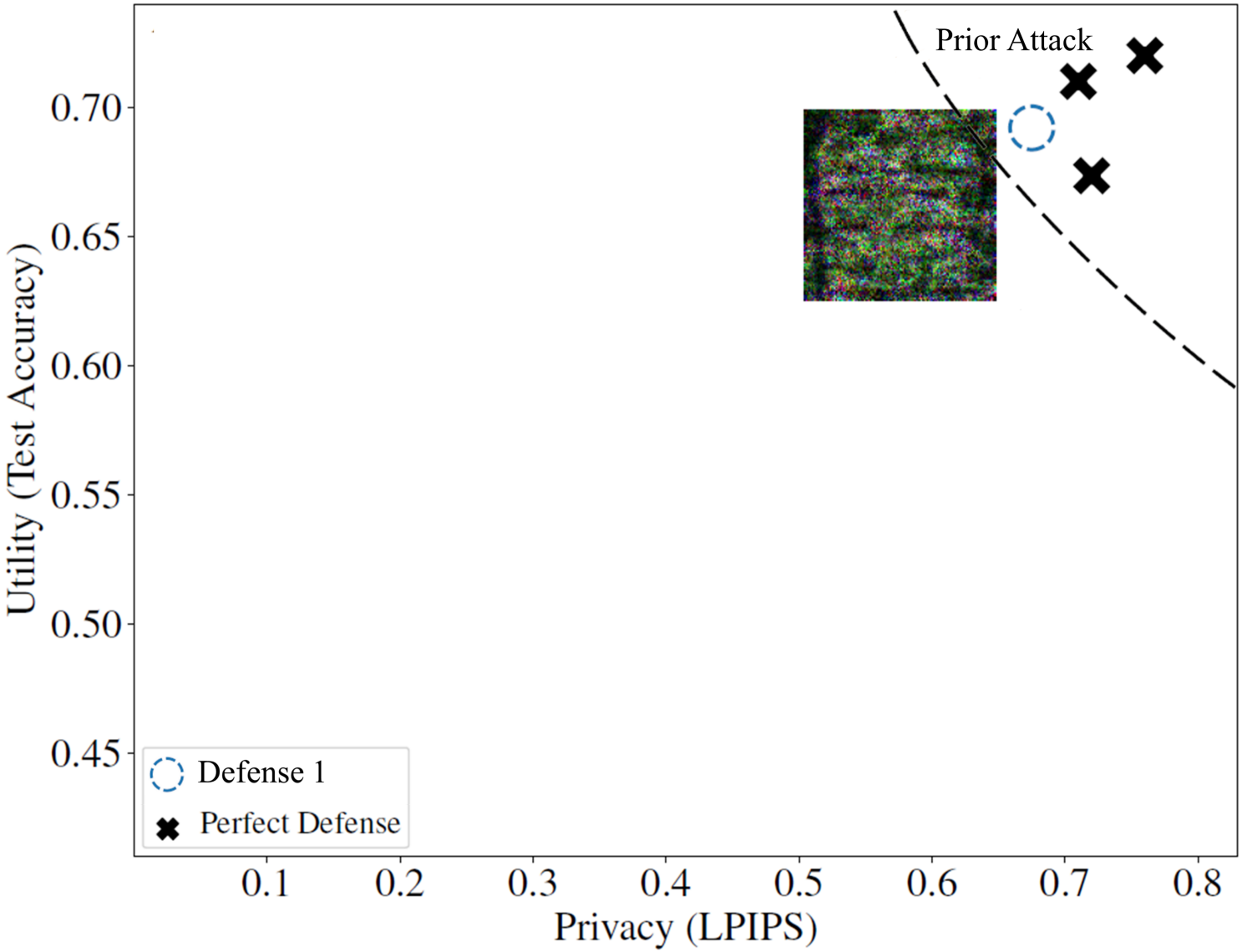
privacy $\rightarrow$ **utility** $\downarrow\downarrow$ ☹

[6] Wei et al. "Gradient-leakage resilient federated learning." ICDCS 2021.

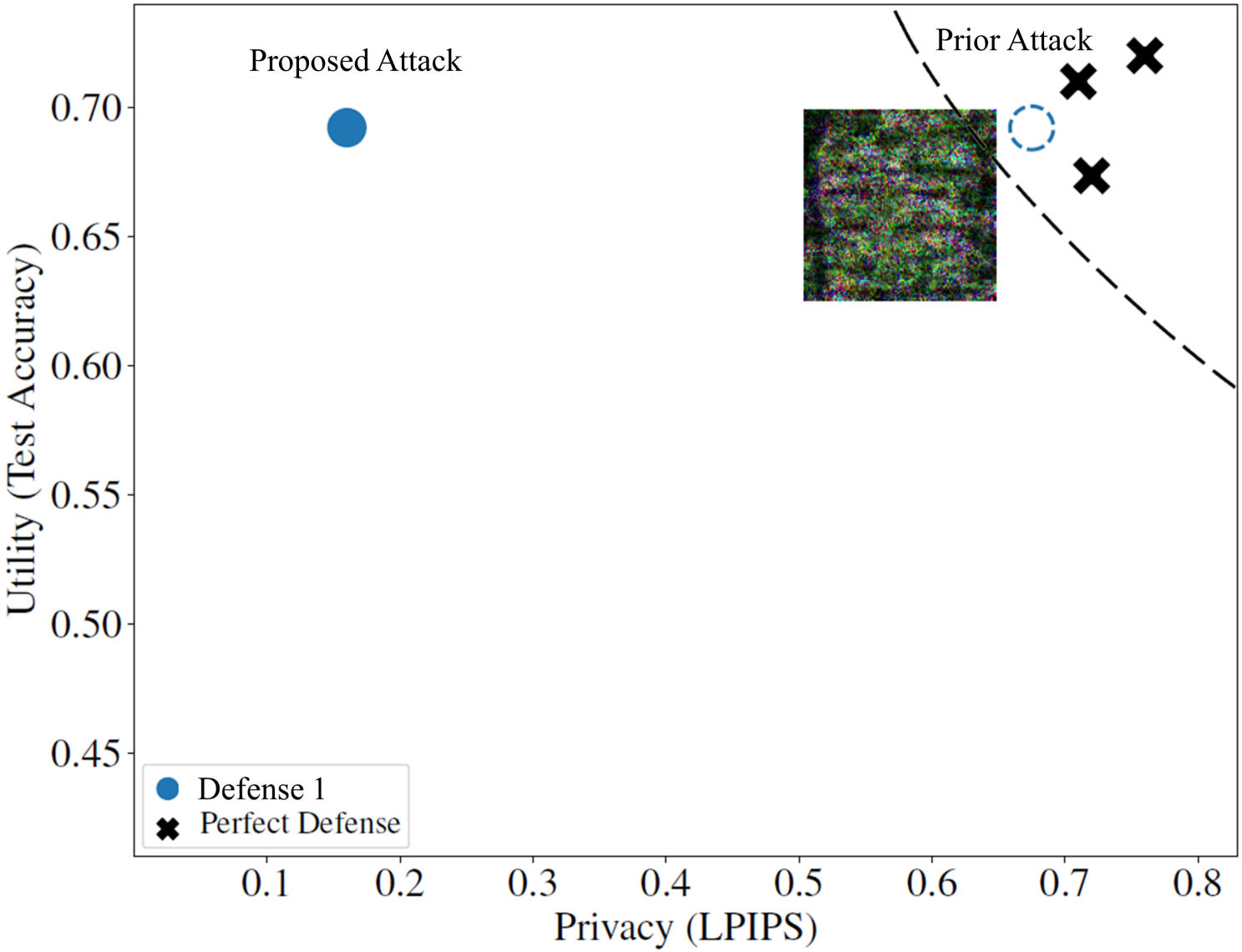
Improved Attack Efficiency



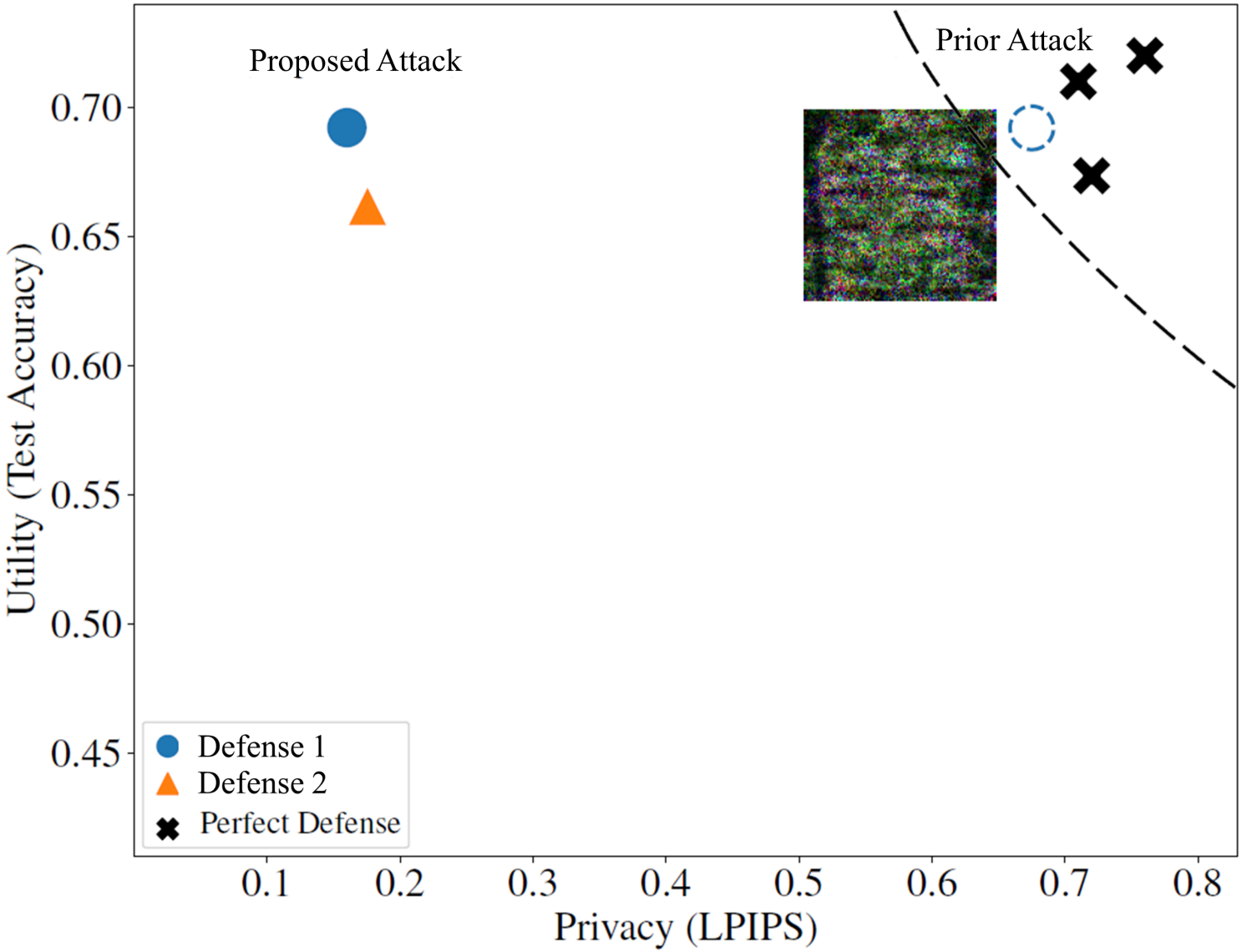
Improved Attack Efficiency



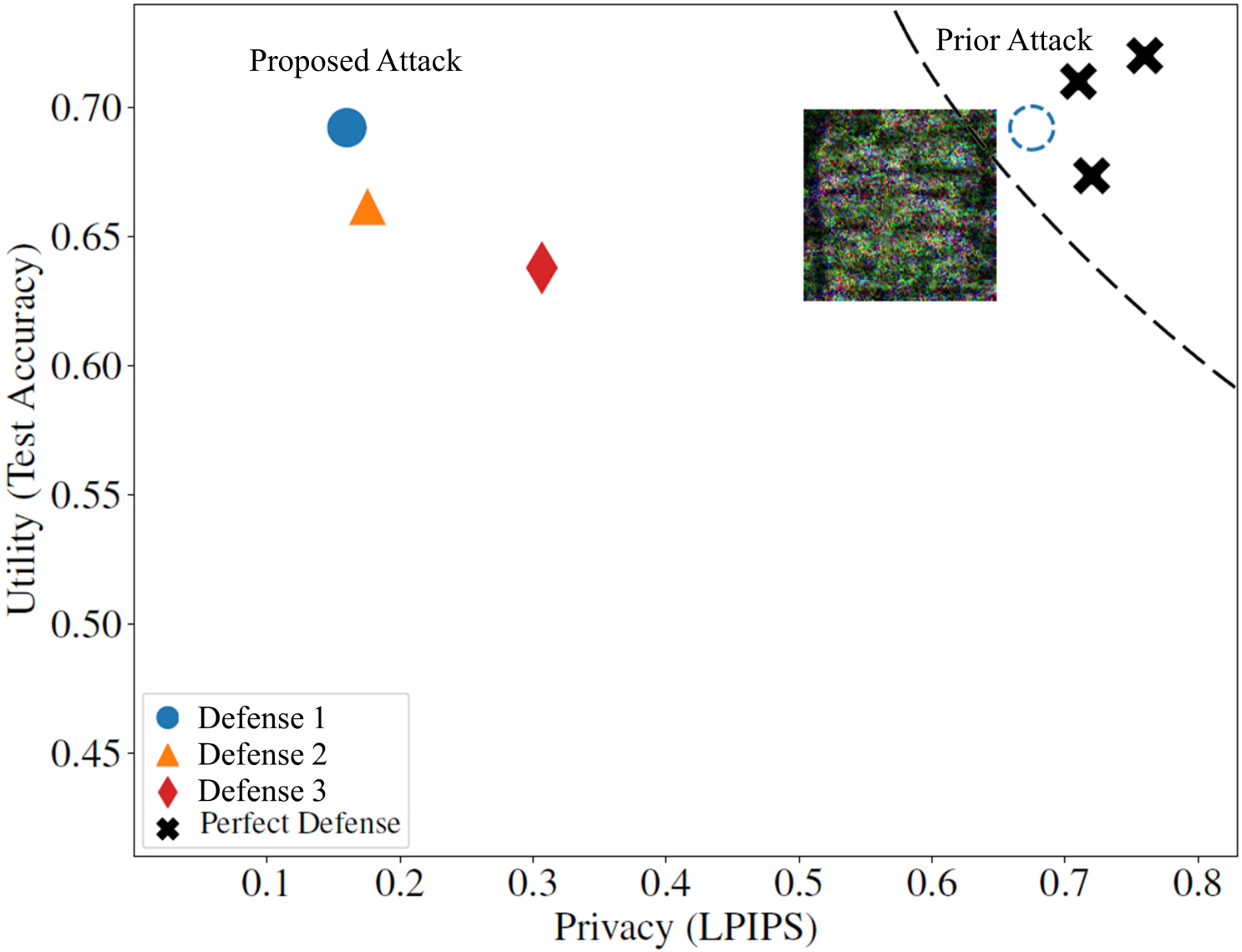
Improved Attack Efficiency



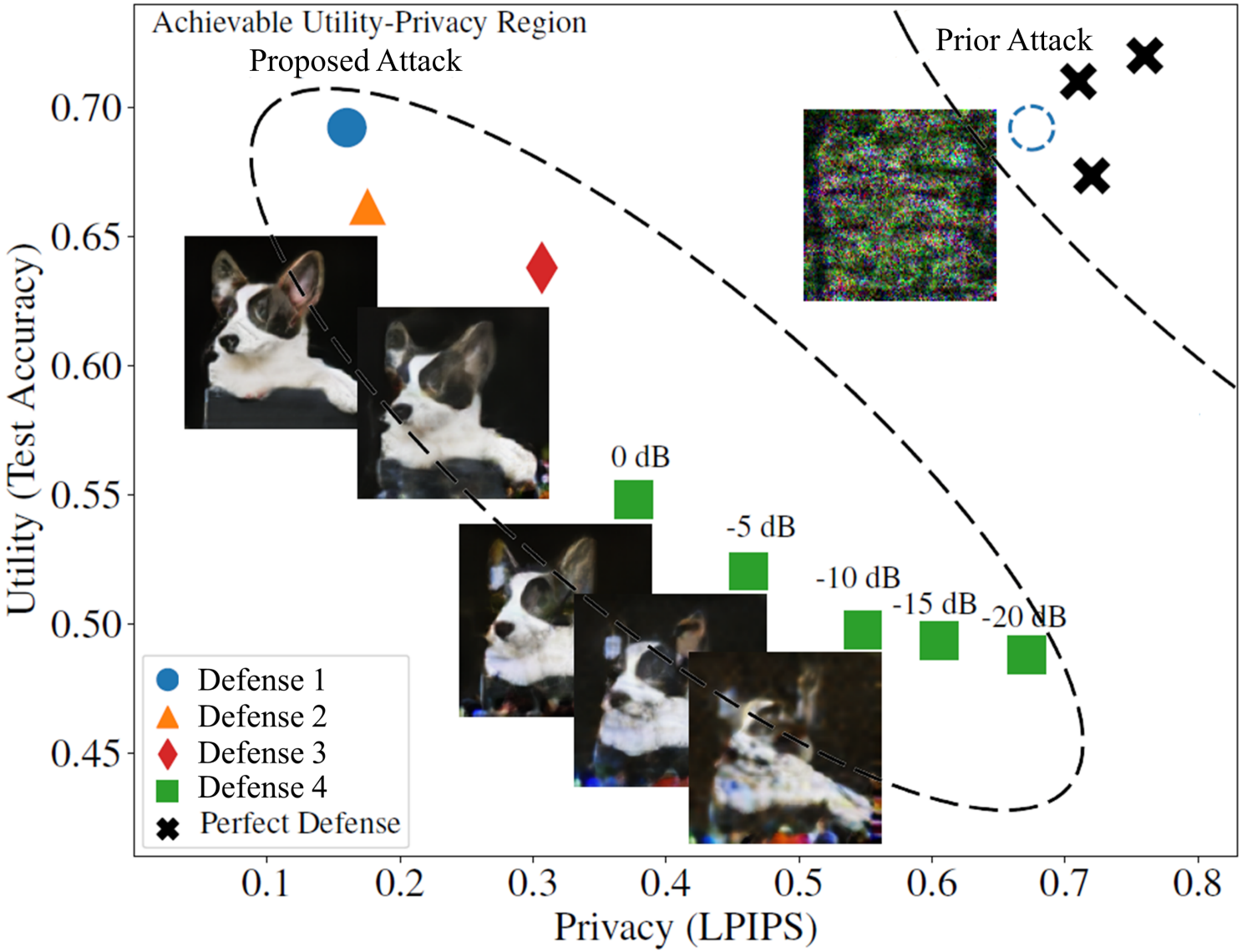
Improved Attack Efficiency



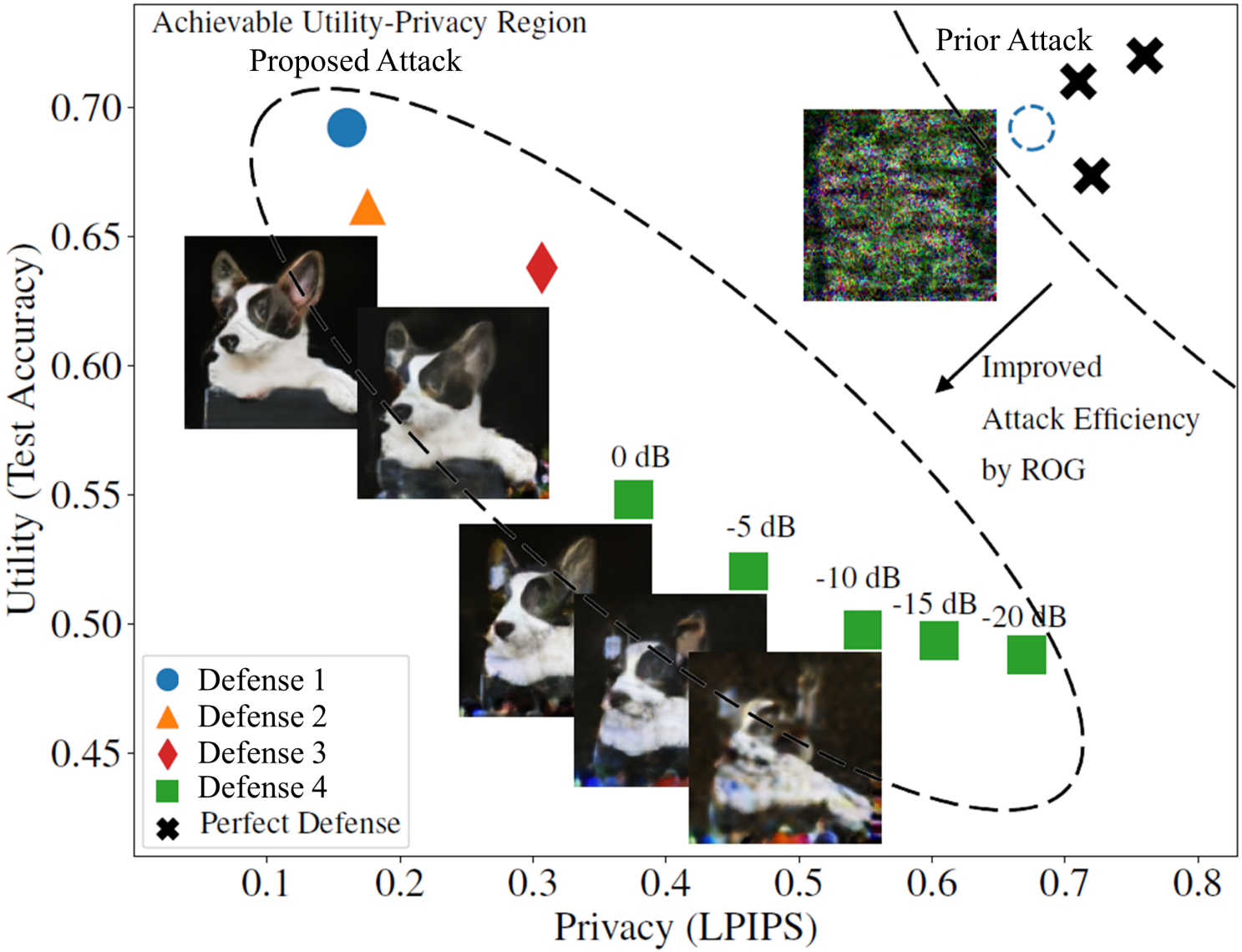
Improved Attack Efficiency



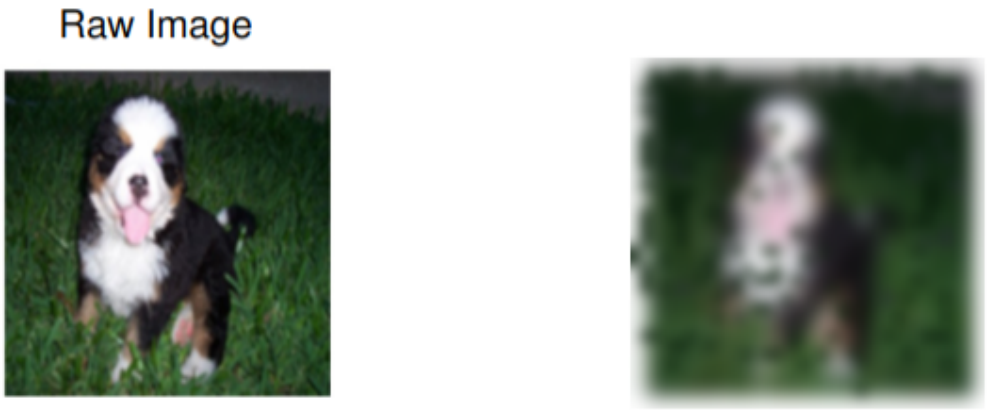
Improved Attack Efficiency



Improved Attack Efficiency



Semantic-Level Attack Variant: ROGS

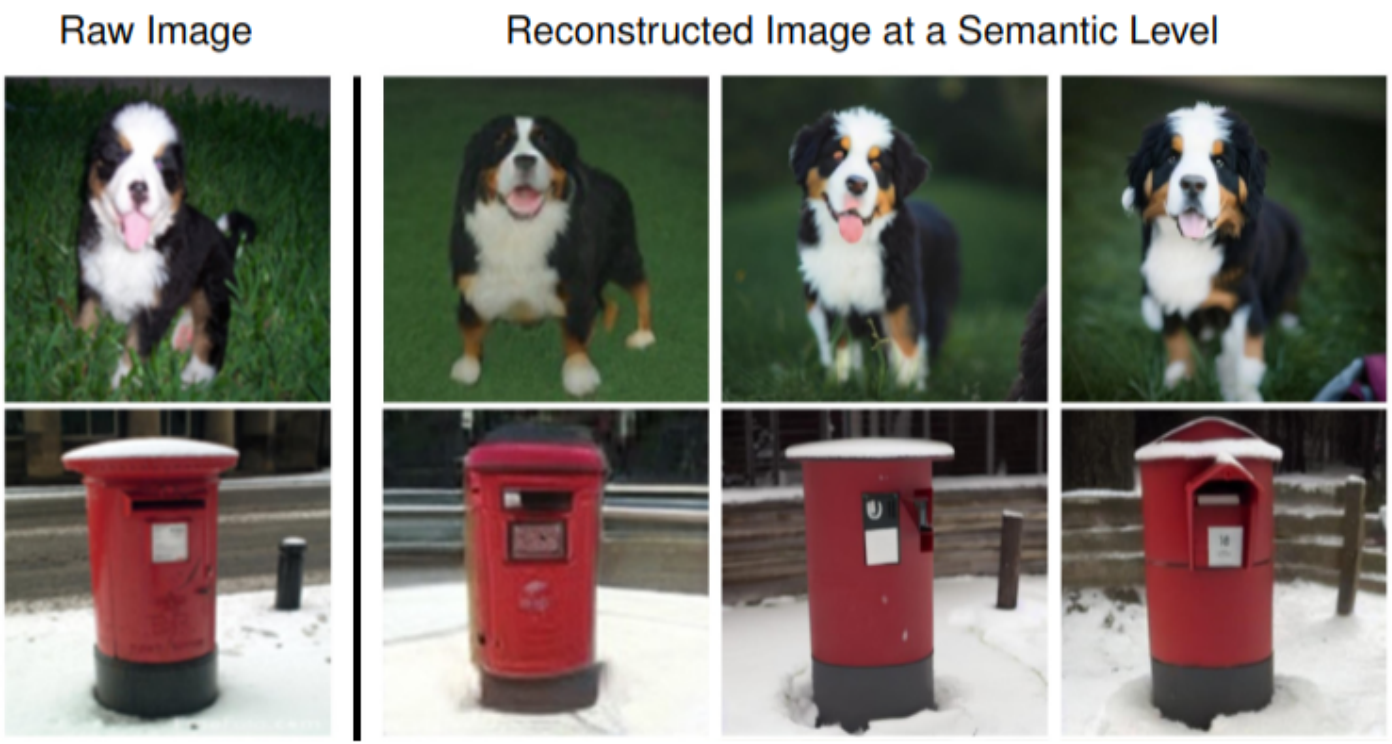


What if the reconstruction is not clear enough?

Semantic-Level Attack Variant: ROGS

Semantic-Level Attack Variant: ROGS

Semantic-Level Attack Variant: ROGS



Find more details in our paper!

Index

Conclusion

- We have studied privacy leakage under various obfuscation mechanisms in federated learning

Conclusion

- We have studied privacy leakage under various obfuscation mechanisms in federated learning
- We proposed an attack framework, which shows that
  - ▷ compressed gradient: does not protect privacy inherently
  - ▷ noisy gradient (DP): bad tradeoff between utility & privacy

Conclusion

- We have studied privacy leakage under various obfuscation mechanisms in federated learning
- We proposed an attack framework, which shows that
  - ▷ compressed gradient: does not protect privacy inherently
  - ▷ noisy gradient (DP): bad tradeoff between utility & privacy
- Future work: hybrid defense & more precise privacy notion

Gradient Obfuscation Gives a False Sense of Security in Federated Learning

Kai Yue[1] Richeng Jin[2]

Chau-Wai Wong[1] Dror Baron[1] Huaiyu Dai[1]

[1]NC State University [2]Zhejiang University

USENIX Security 2023

Thanks!

August 11, 2023