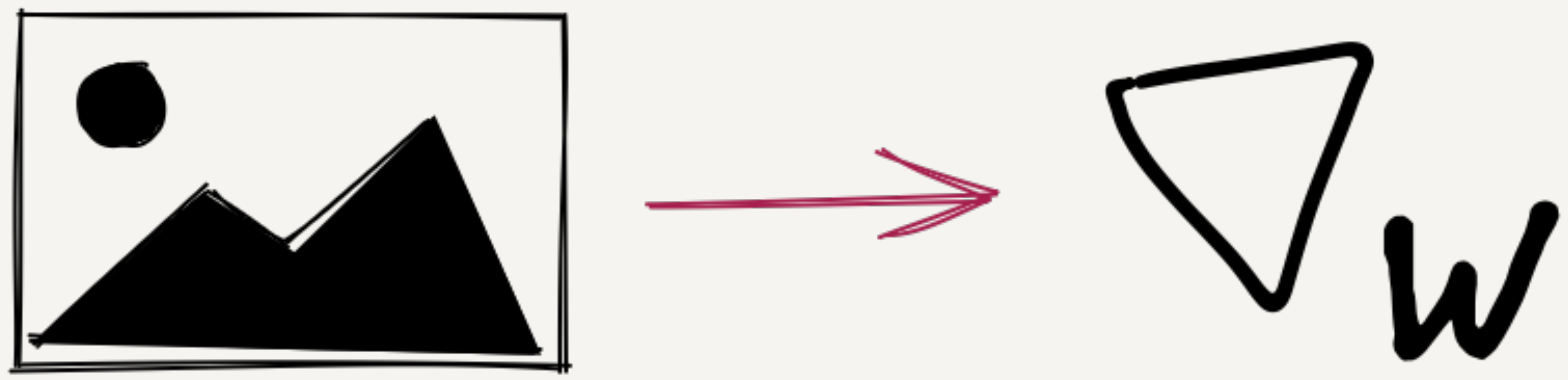
The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

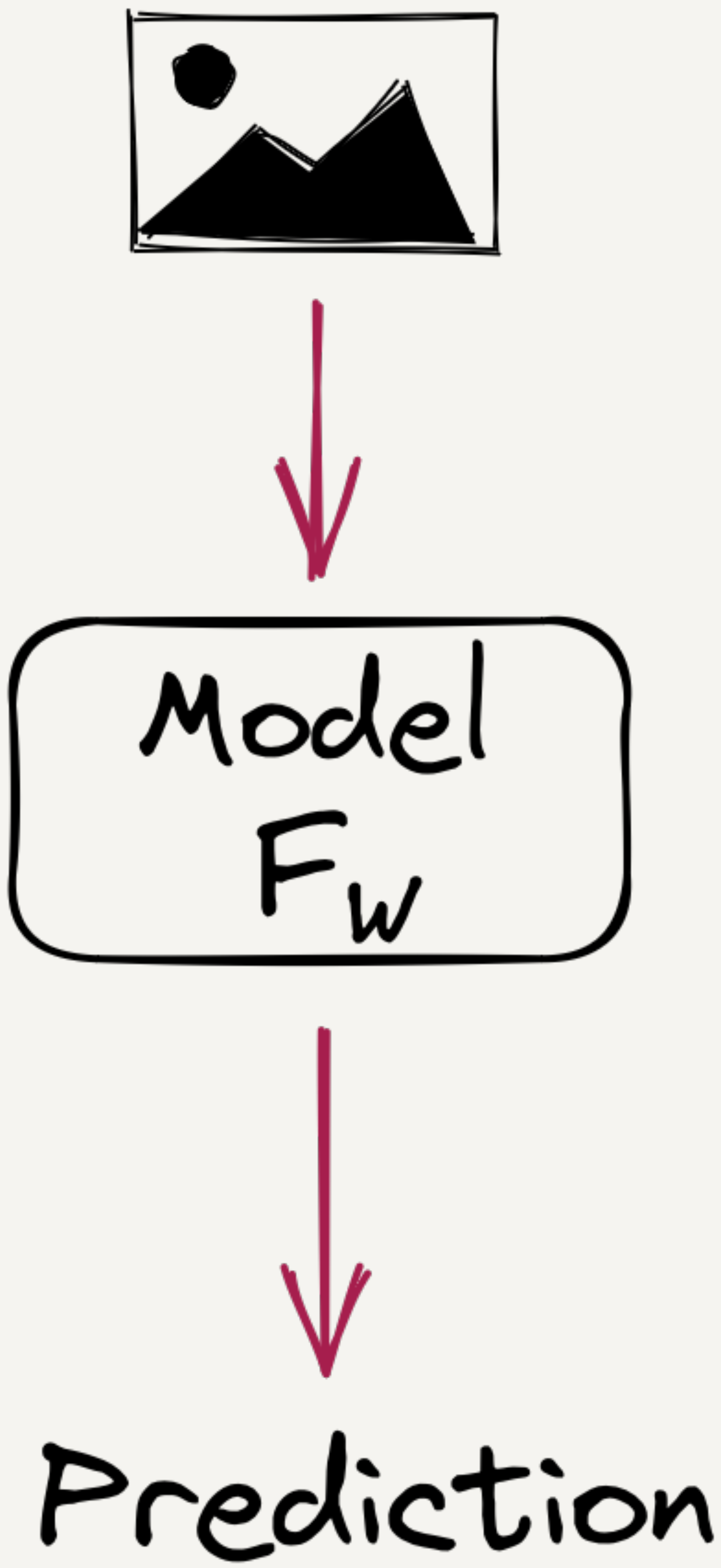
More than Enough is Too Much: Adaptive Defenses against Gradient Leakage in Production Federated Learning

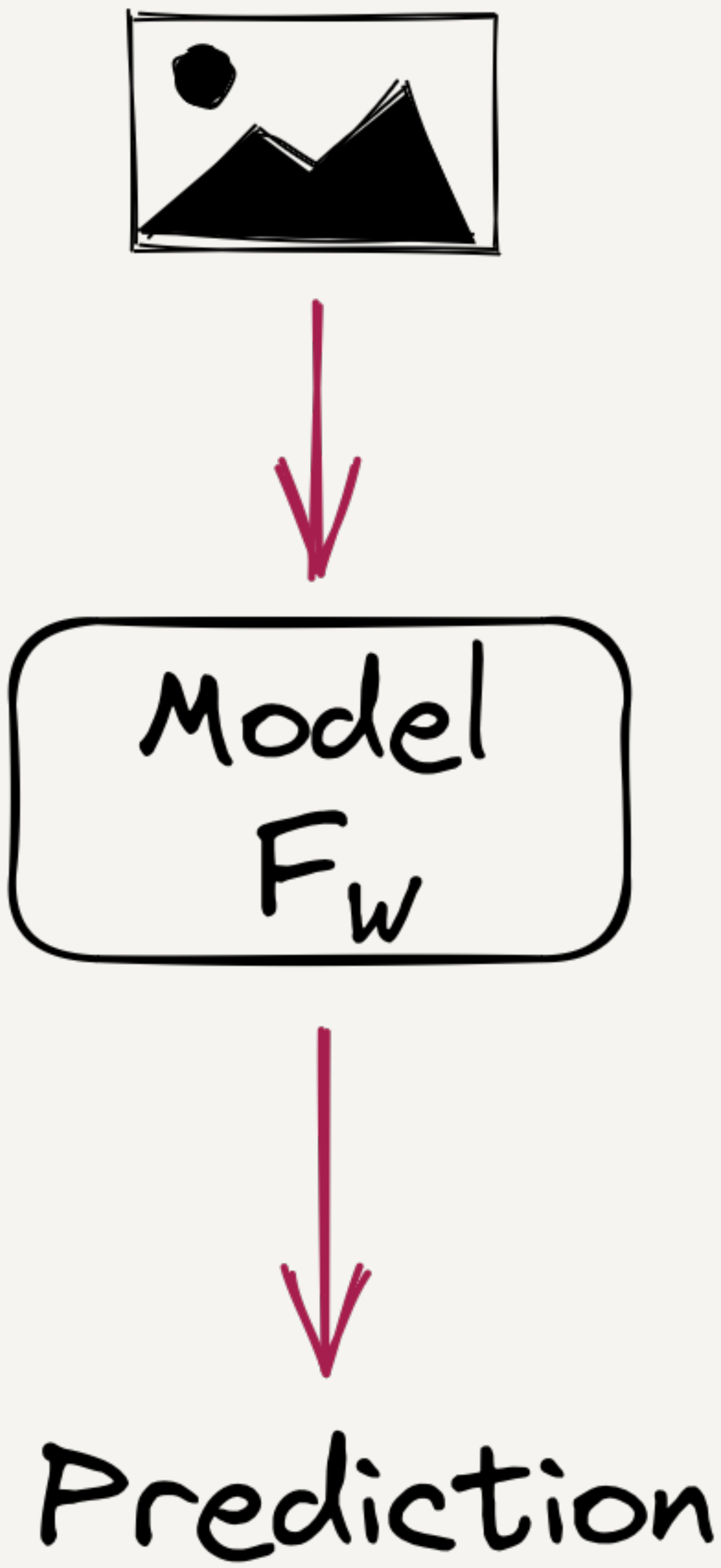
Fei Wang, Ethan Hugh, Baochun Li
Department of Electrical and Computer Engineering
University of Toronto

INFOCOM'23

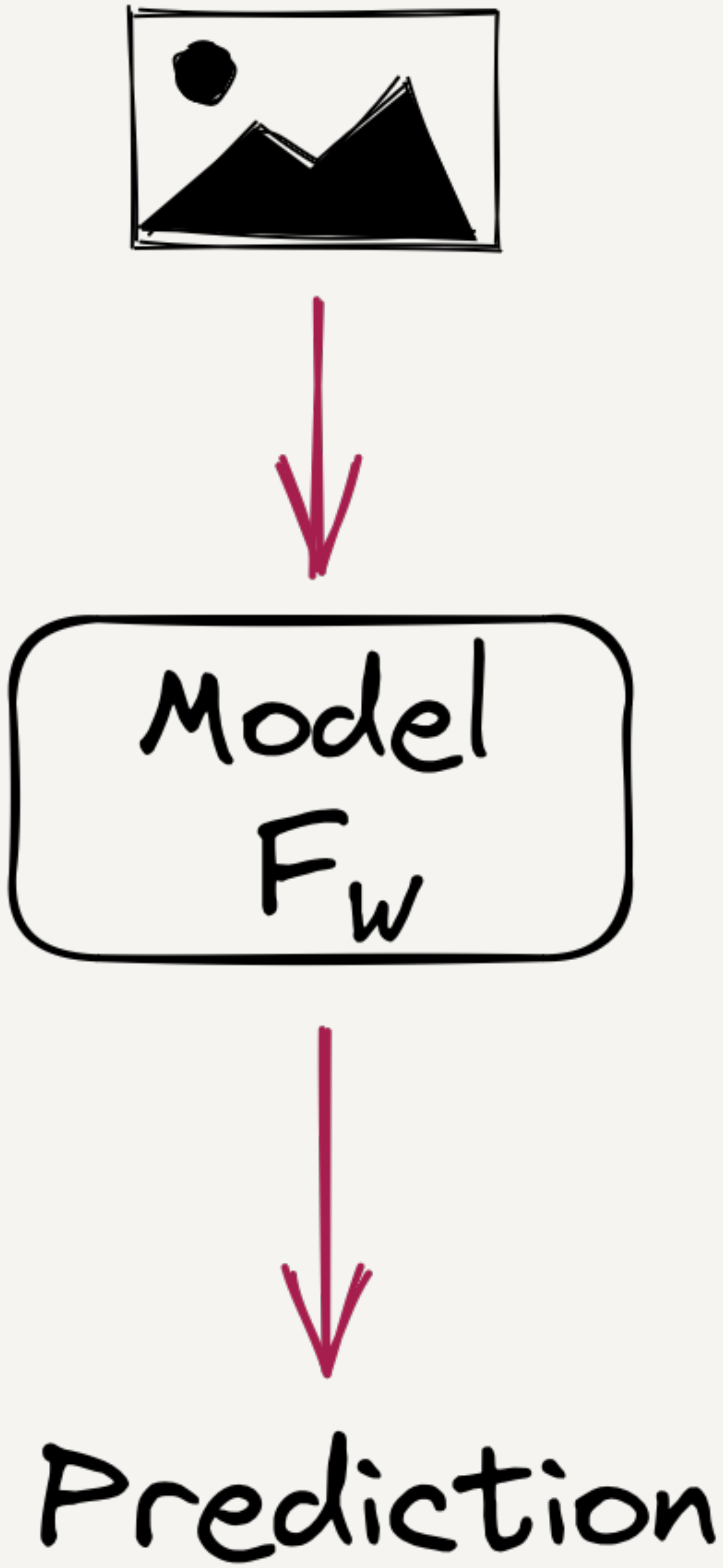
A Glance at Gradient Leakage Attacks


∇_w


Model
F_W
Prediction

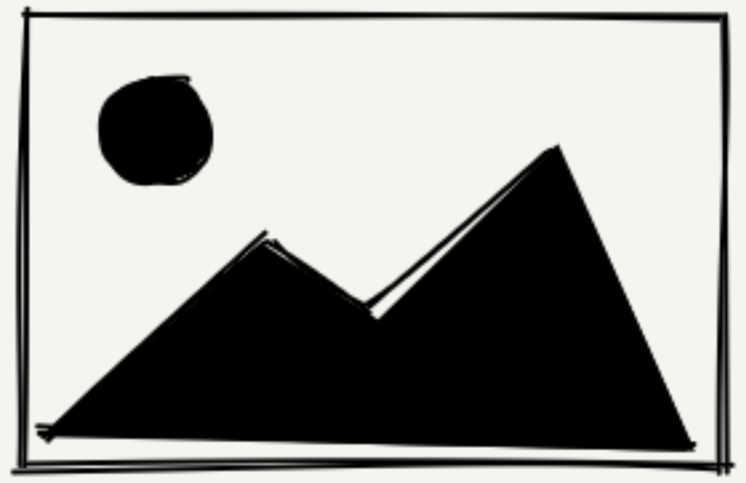


$$\nabla_W = \frac{\partial}{\partial W} \text{Loss}(\text{Prediction}, \text{Label})$$

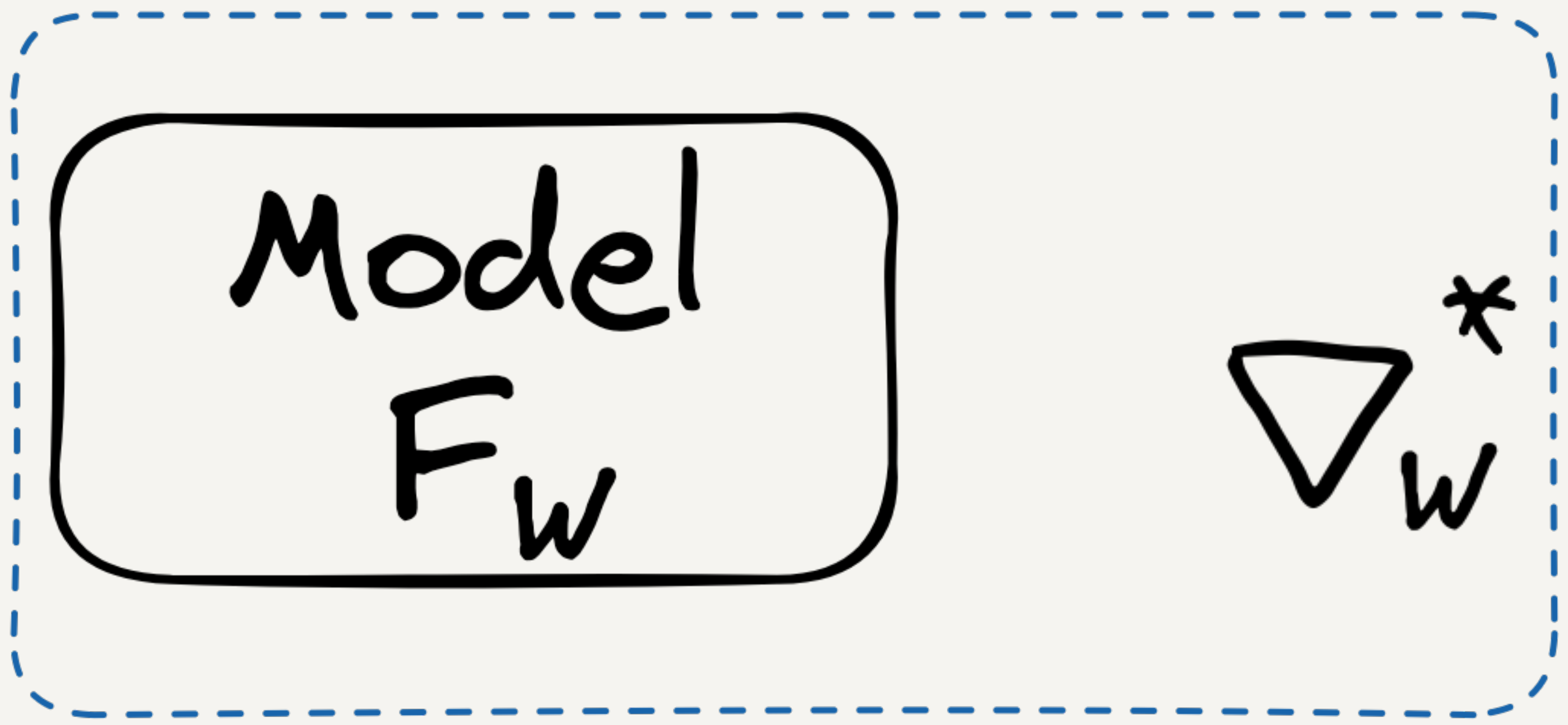


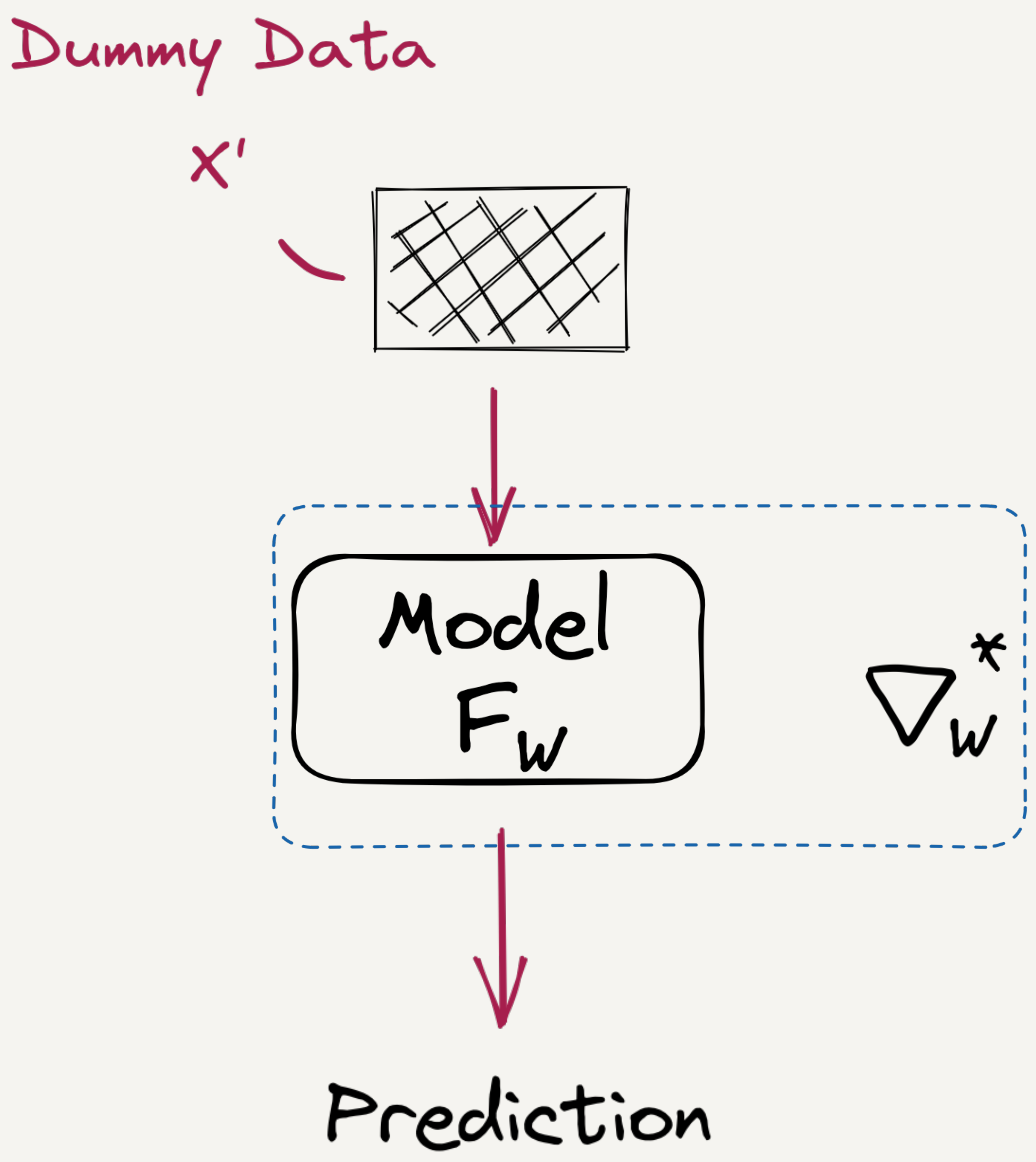
$$\nabla_w = \frac{\partial}{\partial w} \text{Loss}\left(\text{Prediction}, \text{Label}\right)$$

$$w' = w - \eta \nabla_w$$

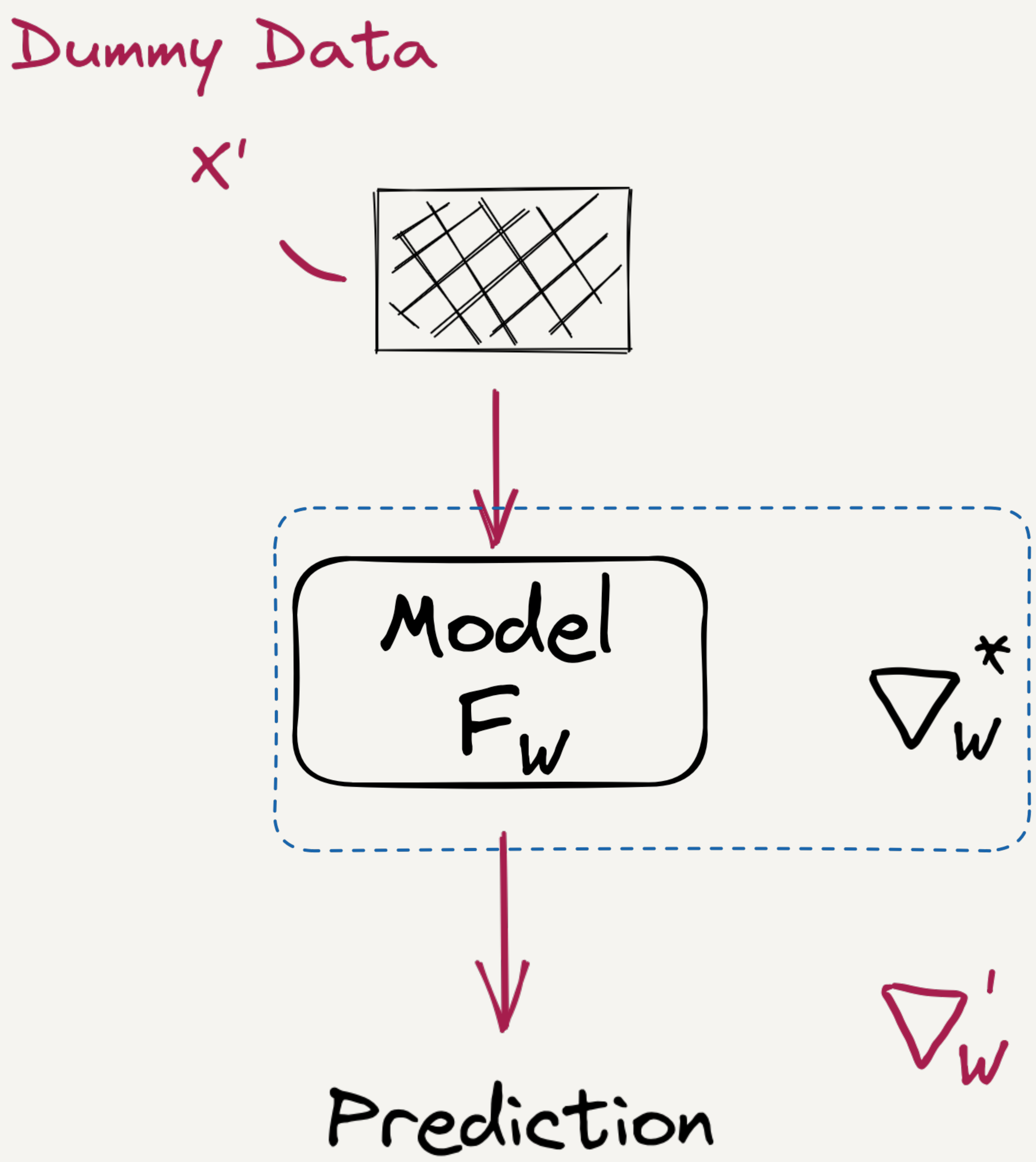
$\nabla_w \longrightarrow$ 

Zhu et al., “Deep Leakage from Gradients,” NeurIPS 2019.

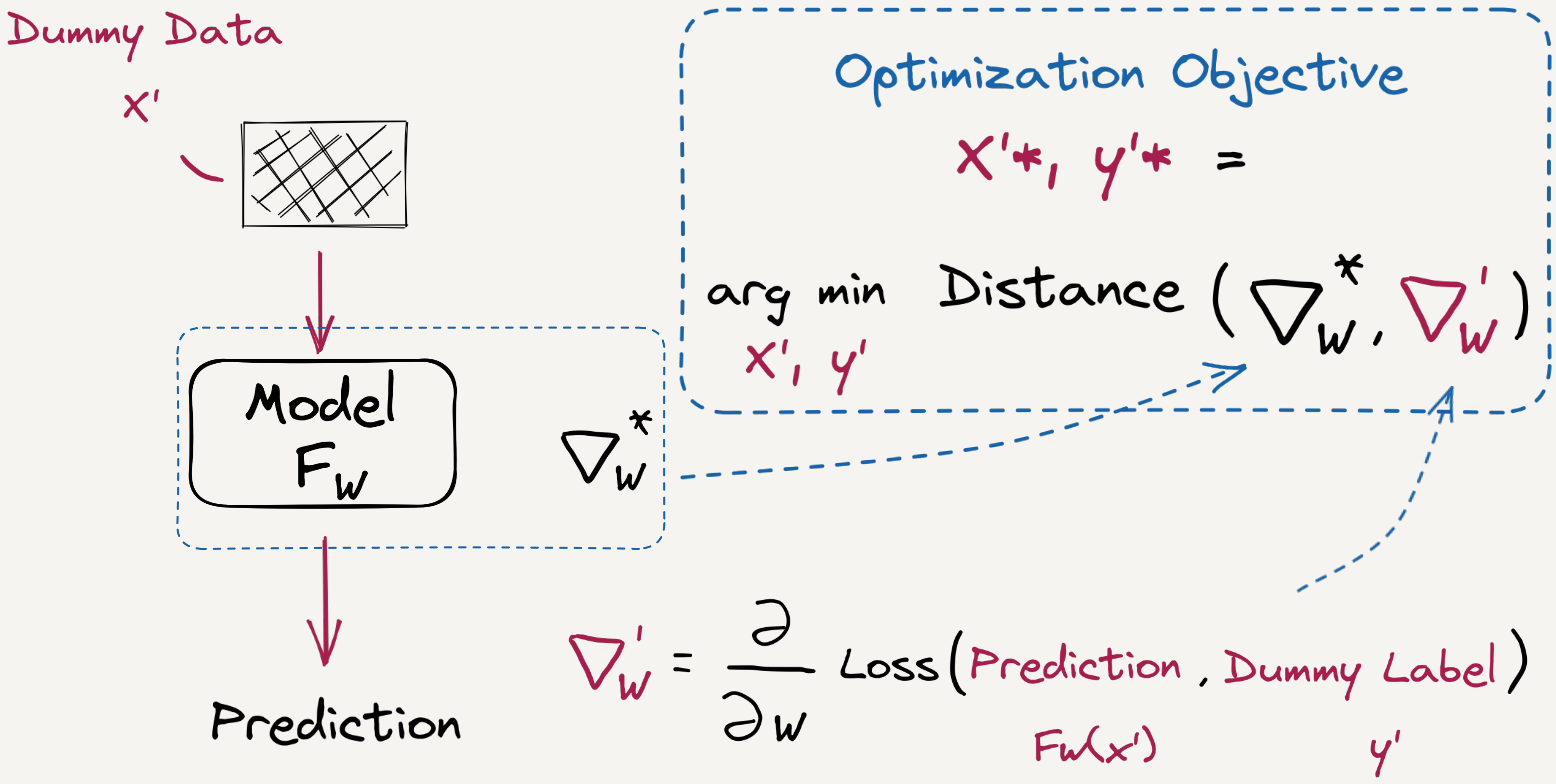

Model
F_W
∇_W^*



Zhu et al., “Deep Leakage from Gradients,” NeurIPS 2019.



$$\nabla'_W = \frac{\partial}{\partial W} \text{Loss}\Big(\underset{F_W(x')}{\text{Prediction}}, \underset{y'}{\text{Dummy Label}}\Big)$$



Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.

Stealing Client's data in Federated Learning

Stealing Client's data in Federated Learning

Client's gradients are shared with the server.

Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.
Zhao et al., "iDLG: Improved Deep Leakage from Gradients," arXiv 2020.
Jeon et al., "Gradient Inversion with Generative Image Prior," NeurIPS 2021.

Client's gradients are shared with the server.

Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.
Zhao et al., "iDLG: Improved Deep Leakage from Gradients," arXiv 2020.
Jeon et al., "Gradient Inversion with Generative Image Prior," NeurIPS 2021.

Client's gradients are ~~shared~~ with the server.

$$\Delta = W' - W$$

Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.
Zhao et al., "iDLG: Improved Deep Leakage from Gradients," arXiv 2020.
Jeon et al., "Gradient Inversion with Generative Image Prior," NeurIPS 2021.

Client's gradients are ~~shared~~ with the server.

$$\Delta = W' - W$$

$$\Delta \neq \nabla!$$

Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.
Zhao et al., "iDLG: Improved Deep Leakage from Gradients," arXiv 2020.
Jeon et al., "Gradient Inversion with Generative Image Prior," NeurIPS 2021.

$$\nabla = \frac{\Delta}{-\eta}$$

Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning,"
Proc. European Symposium on Research in Computer Security 2020.
Wu et al., "Fast-Convergent Federated Learning with Adaptive Weighting,"
IEEE Trans. on Cognitive Communications and Networking 2021.

$$\nabla = \frac{\Delta}{-\eta}$$

Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning,"
Proc. European Symposium on Research in Computer Security 2020.
Wu et al., "Fast-Convergent Federated Learning with Adaptive Weighting,"
IEEE Trans. on Cognitive Communications and Networking 2021.

$$\nabla \cancel{=} \frac{\Delta}{-\eta}$$

There are multiple steps of gradient descent in one round of client's local training.

Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning," Proc. European Symposium on Research in Computer Security 2020.

Wu et al., "Fast-Convergent Federated Learning with Adaptive Weighting," IEEE Trans. on Cognitive Communications and Networking 2021.

$$\nabla \cancel{=} \frac{\Delta}{-\eta}$$

There are multiple steps of gradient descent in one round of client's local training.

Data samples >> 1

Batch size << # Data samples

Epochs > 1

Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning," Proc. European Symposium on Research in Computer Security 2020.
Wu et al., "Fast-Convergent Federated Learning with Adaptive Weighting," IEEE Trans. on Cognitive Communications and Networking 2021.

$$\nabla \cancel{=} -\frac{\Delta}{\eta}$$

There are multiple steps of gradient descent in one round of client's local training.

Data samples >> 1

Batch size << # Data samples

Epochs > 1

More sophisticated gradient descent algorithms are routinely used.

Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning," Proc. European Symposium on Research in Computer Security 2020.
Wu et al., "Fast-Convergent Federated Learning with Adaptive Weighting," IEEE Trans. on Cognitive Communications and Networking 2021.

$$\nabla \cancel{=} \frac{\Delta}{-\eta}$$

There are multiple steps of gradient descent in one round of client's local training.

\# Data samples >> 1

Batch size << # Data samples

\# Epochs > 1

More sophisticated gradient descent algorithms are routinely used.

Momentum

Weight decay

Learning rate scheduler

...

Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning," Proc. European Symposium on Research in Computer Security 2020.
Wu et al., "Fast-Convergent Federated Learning with Adaptive Weighting," IEEE Trans. on Cognitive Communications and Networking 2021.

$$x'^*, y'^* = \underset{x', y'}{\arg\min}\ \text{Distance}\left(\nabla_W^*, \nabla_W'\right)$$

Geiping et al., “Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning,” NeurIPS 2020.

$$x'^*, y'^* = \underset{x', y'}{\arg\min} \; \text{Distance}\left(\cancel{\nabla_w^*, \nabla_w'}\right)$$

$$\Delta_w^*, \Delta_w' \;?$$

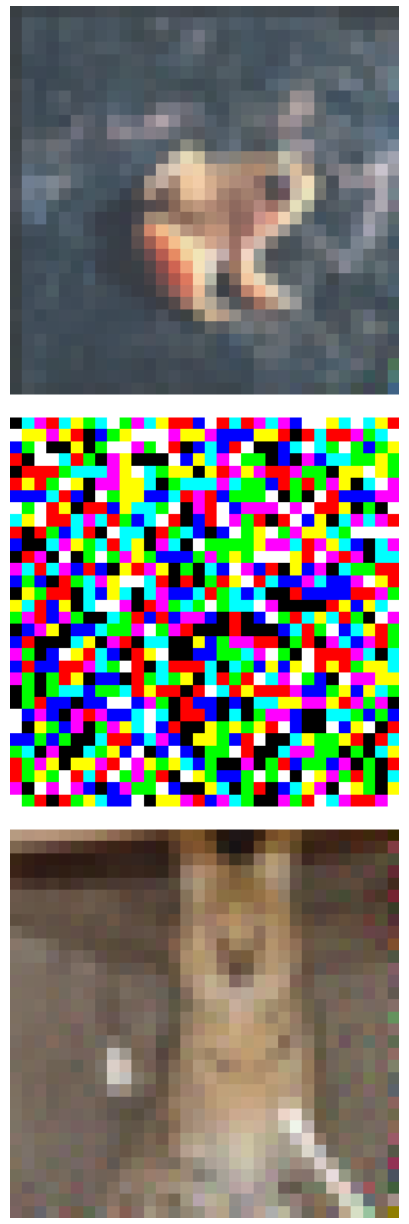
Geiping et al., “Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning,” NeurIPS 2020.

$$x'^*, y'^* = \arg\min_{x', y'} \text{Distance}\left(\cancel{\nabla W^*, \nabla W'}\right)$$

$$\Delta W^*, \Delta W' \; ?$$

To realize the same gradient descent process using the dummy data instead, the server requires a series of prior knowledge.

Geiping et al., "Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning," NeurIPS 2020.

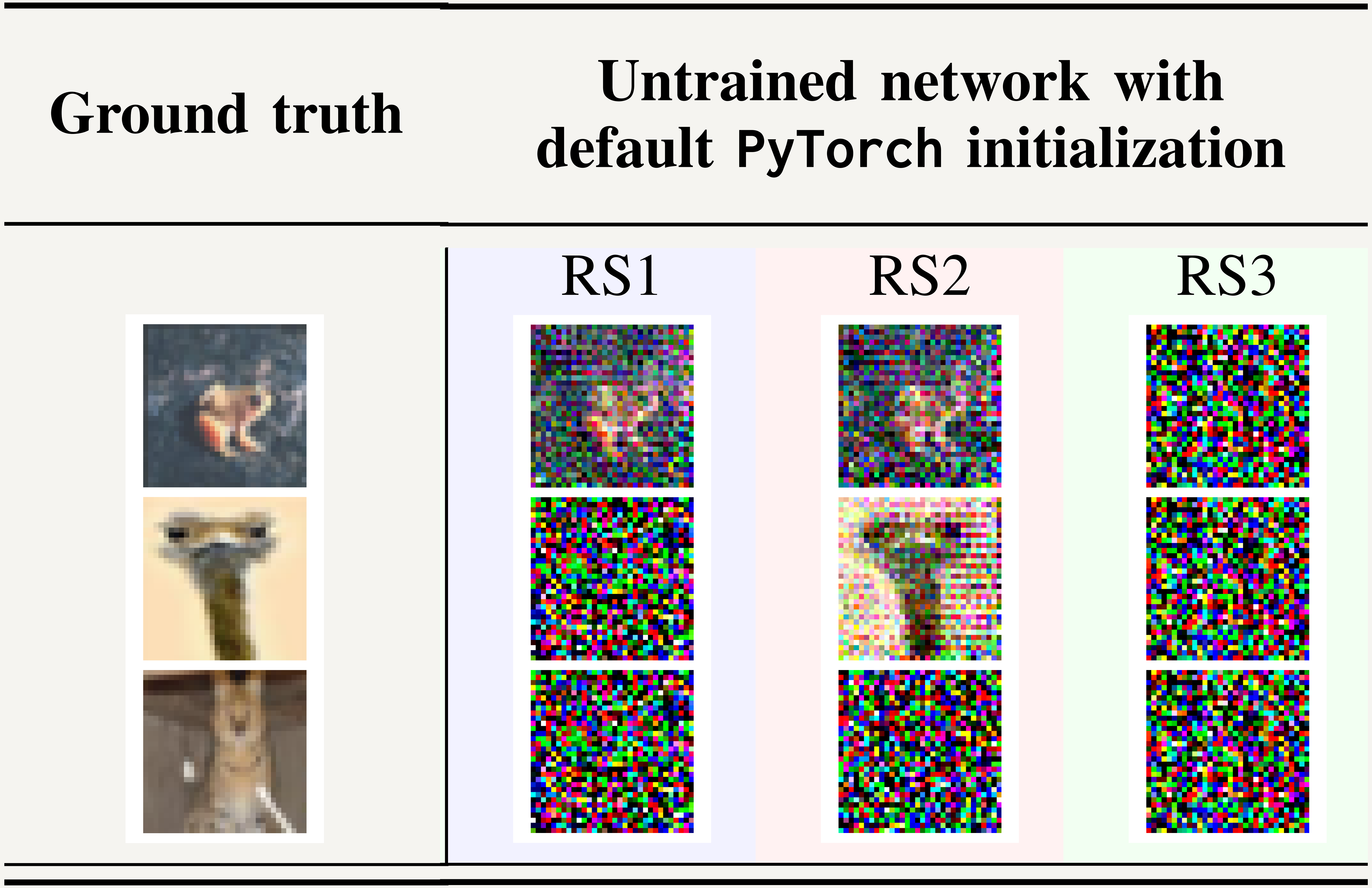
Ground truth	Untrained network with explicit initialization		
	RS1	RS2	RS3
			

Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.

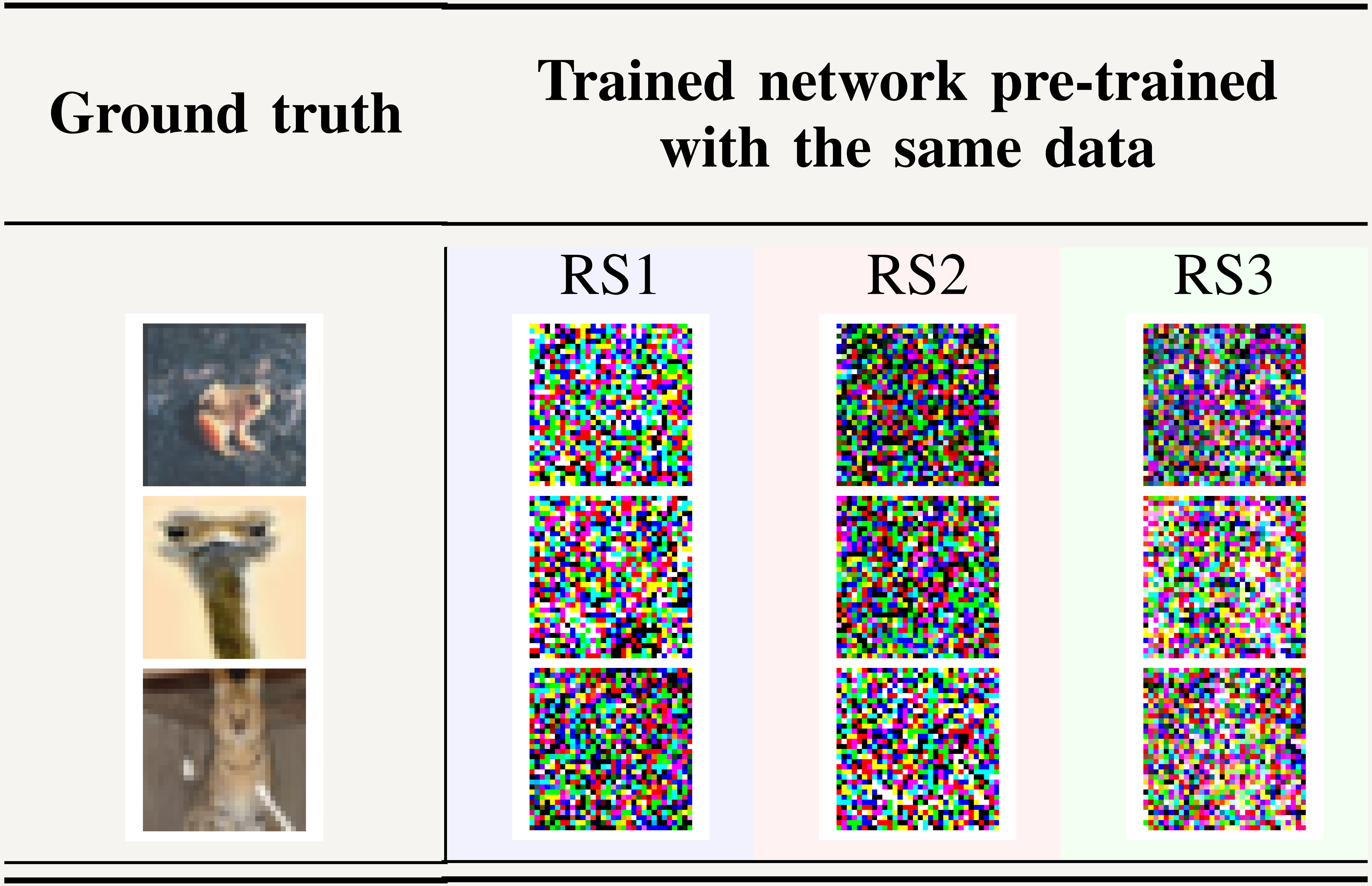
Geiping et al., "Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning," NeurIPS 2020.

Zhao et al., "iDLG: Improved Deep Leakage from Gradients," arXiv 2020.

Jeon et al., "Gradient Inversion with Generative Image Prior," NeurIPS 2021.



Zhu et al., "Deep Leakage from Gradients," NeurIPS 2019.
Geiping et al., "Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning," NeurIPS 2020.
Zhao et al., "iDLG: Improved Deep Leakage from Gradients," arXiv 2020.
Jeon et al., "Gradient Inversion with Generative Image Prior," NeurIPS 2021.



Zhu et al., “Deep Leakage from Gradients,” NeurIPS 2019.

Geiping et al., “Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning,” NeurIPS 2020.

Zhao et al., “iDLG: Improved Deep Leakage from Gradients,” arXiv 2020.

Jeon et al., “Gradient Inversion with Generative Image Prior,” NeurIPS 2021.

Outpost: Our Lightweight Defense

Sufficient and self-adaptive protection

Sufficient and self-adaptive protection

Selective perturbation

Sufficient and self-adaptive protection

The diagonal of the
Fisher information matrix

Selective perturbation

Sufficient and self-adaptive protection

The diagonal of the
Fisher information matrix

Selective perturbation

Sufficient and self-adaptive protection

Different intensity

The diagonal of the
Fisher information matrix

Selective perturbation

Sufficient and self-adaptive protection

Different intensity

The range of weight values

The diagonal of the
Fisher information matrix

Selective perturbation

Sufficient and self-adaptive protection

Different intensity

The range of weight values

The diagonal of the Fisher information matrix

Selective perturbation

Gradually decreased likelihood

Sufficient and self-adaptive protection

Different intensity

The range of weight values

The diagonal of the Fisher information matrix

Selective perturbation

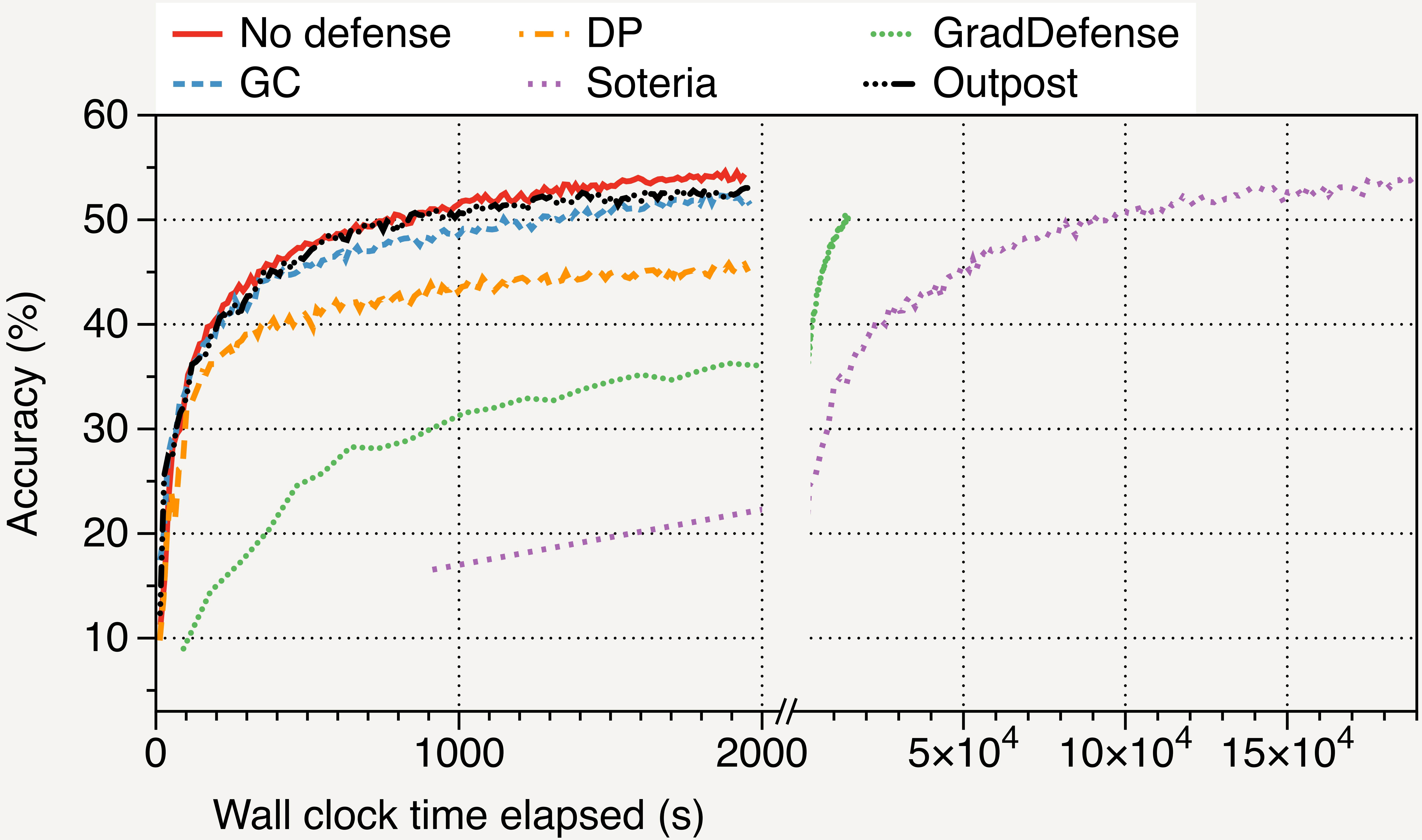
With each local step of gradient descent

Gradually decreased likelihood

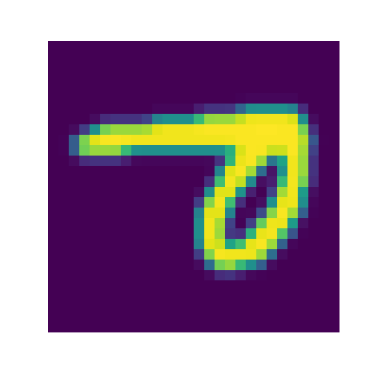
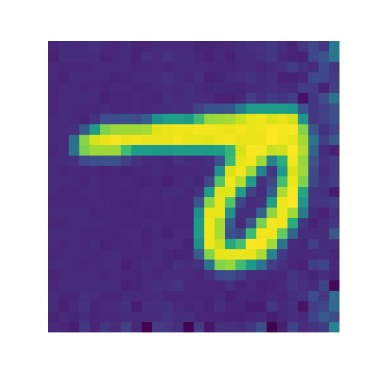
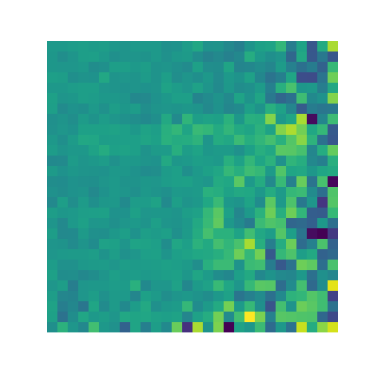
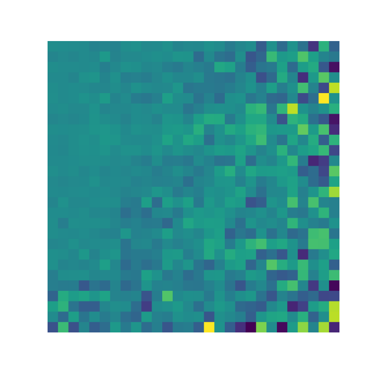
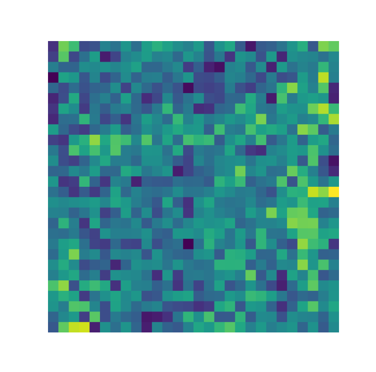
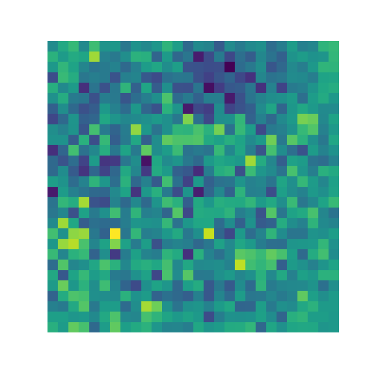
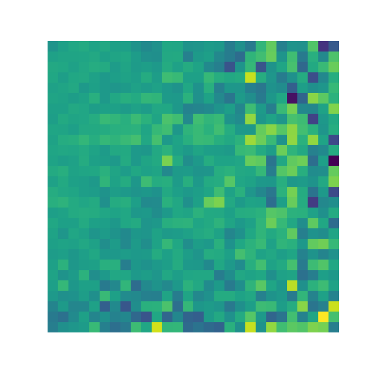
Sufficient and self-adaptive protection

Different intensity

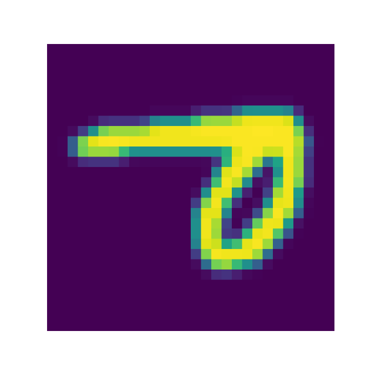
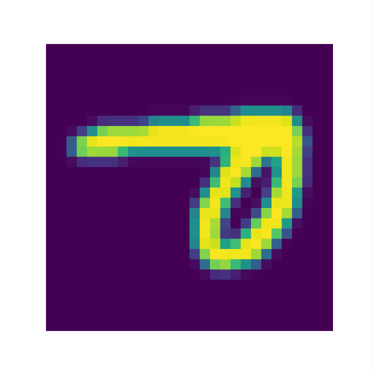
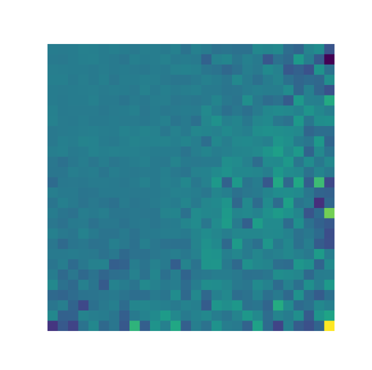
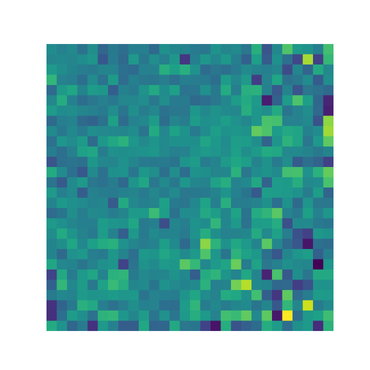
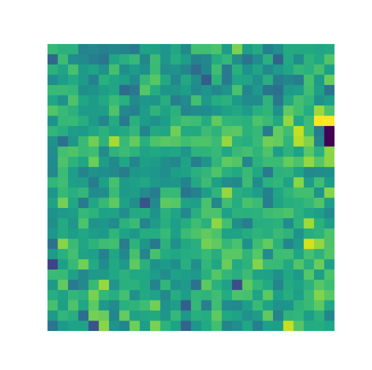
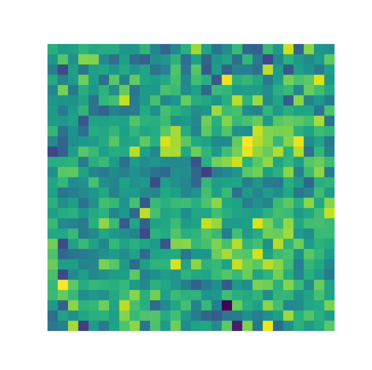
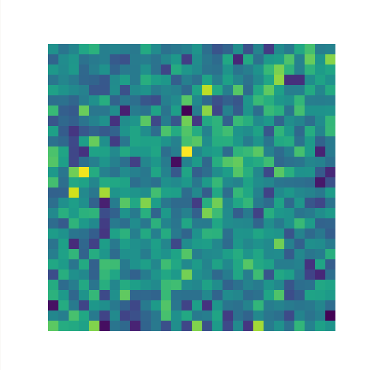
The range of weight values

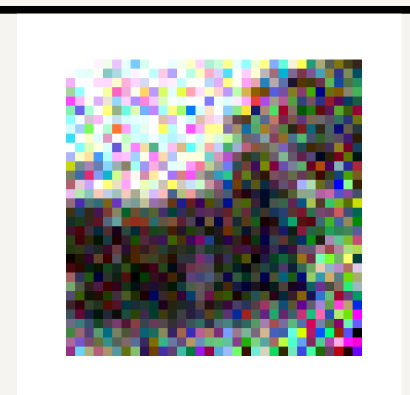
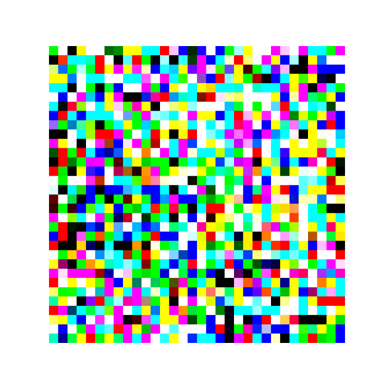
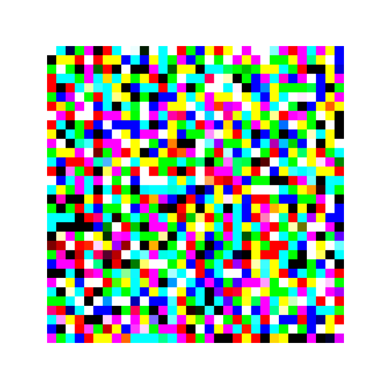
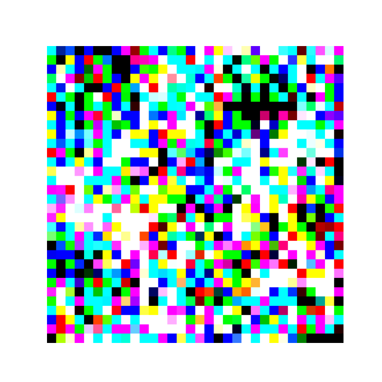
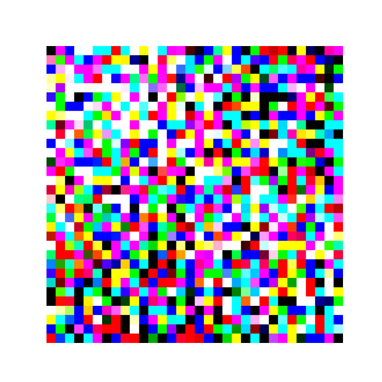


Sun et al., "Soteria: Provable Defense against Privacy Leakage in Federated Learning from Representation Perspective," CVPR 2021.
Wang et al., "Protect Privacy from Gradient Leakage Attack in Federated Learning," INFOCOM 2022.

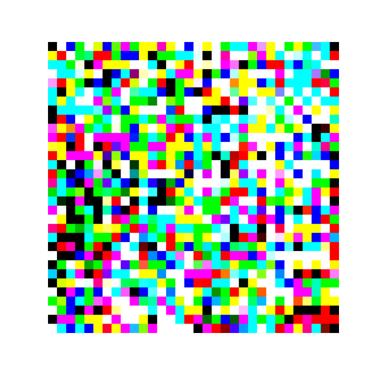
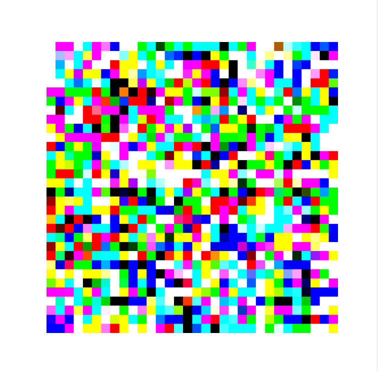
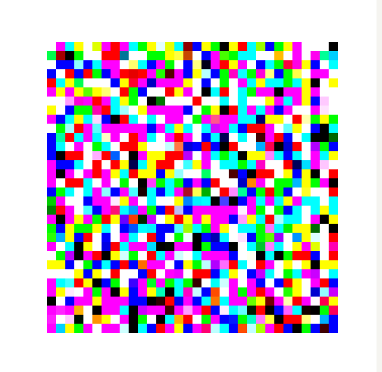
	No defense	GC	DP	Soteria	GD	OUTPOST
MSE ↑	6.6e-3	13.96	113.63	95.19	32.57	77.05
LPIPS ↑	7.1e-2	0.55	0.60	0.63	0.64	0.58
SSIM ↓	0.99	0.30	0.19	3.3e-2	1.0e-2	0.13
						

[Scenario 1] EMNIST: $E = 1, n = 1, B = 1$

	No defense	GC	DP	Soteria	GD	OUTPOST
MSE ↑	2.6e-7	199.08	297.84	296.76	360.98	294.678
LPIPS ↑	5.8e-7	0.60	0.66	0.63	0.64	0.68
SSIM ↓	1.00	0.32	6.5e-2	4.1e-2	1.7e-2	1.6e-2
						

	No defense	GC	DP	Soteria	GD	OUTPOST
MSE ↑	5.1e-2	7.83	34.08	25.91	11.46	13.10
LPIPS ↑	0.53	0.77	0.77	0.76	0.74	0.77
SSIM ↓	0.57	4.4e-2	3.6e-2	5.5e-2	2.2e-2	2.1e-2
						

[Scenario 3] `CIFAR-10`: $E = 1, n = 1, B = 1$

	No defense	GC	DP	Soteria	GD	OUTPOST
MSE ↑	5.9e-5	27.50	34.51	25.91	56.66	35.24
LPIPS ↑	1.8e-3	0.76	0.78	0.76	0.77	0.77
SSIM ↓	0.99	2.6e-2	2.9e-2	5.5e-2	1.7e-2	3.4e-2
						

Plato: A New Framework for Scalable Federated Learning Research

https://github.com/TL-System/plato

A thorough investigation of gradient leakage attacks in production federated learning

A thorough investigation of gradient leakage attacks in production federated learning

Significantly weakened!

A thorough investigation of gradient leakage attacks in production federated learning

Significantly weakened!

Outpost: a defense mechanism

A thorough investigation of gradient leakage attacks in production federated learning

Significantly weakened!

Outpost: a defense mechanism

Sufficient with minimal sacrifice!