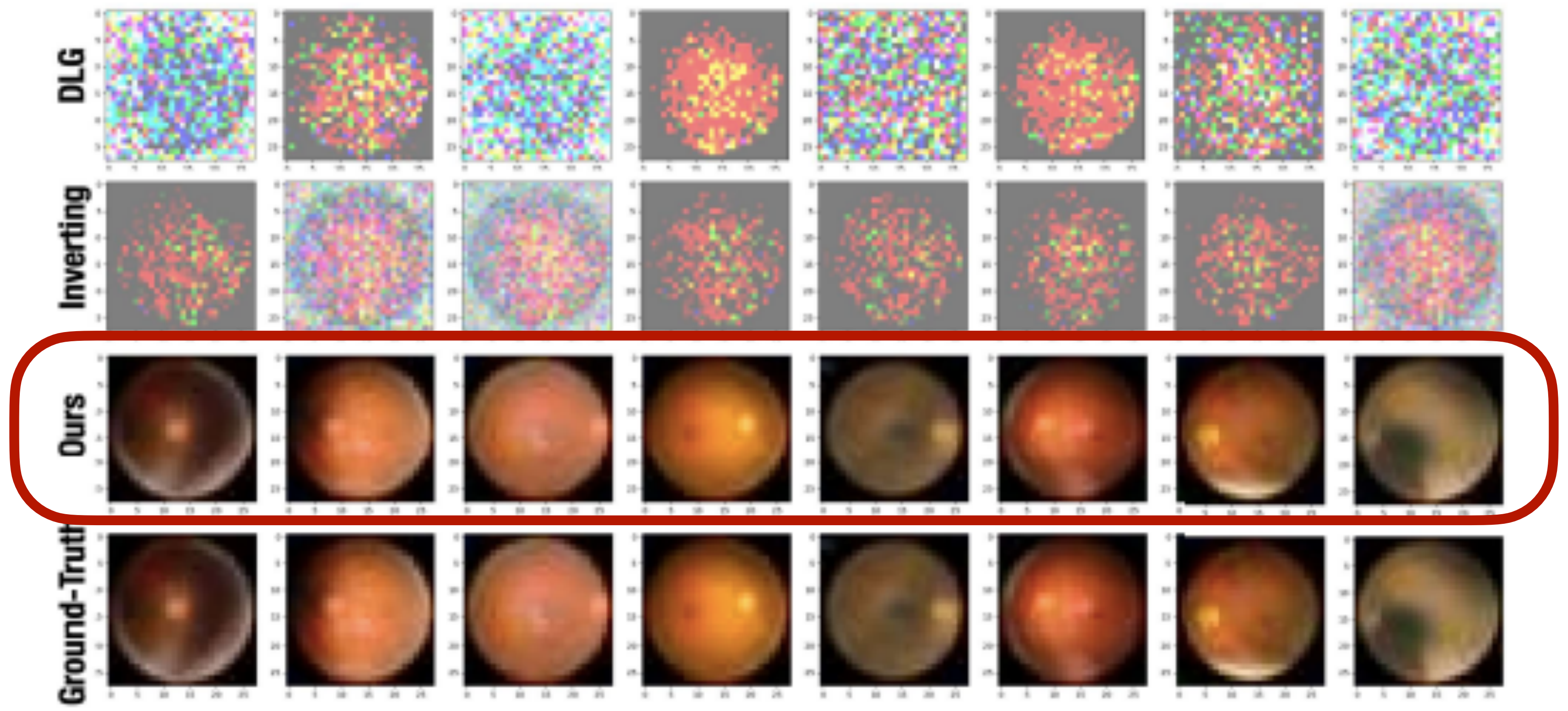
Exploring the Security Boundary of Data Reconstruction via Neuron Exclusivity Analysis

31ST USENIX
SECURITY SYMPOSIUM

Xudong Pan, Mi Zhang*, Yifan Yan, Jiaming Zhu, Min Yang*

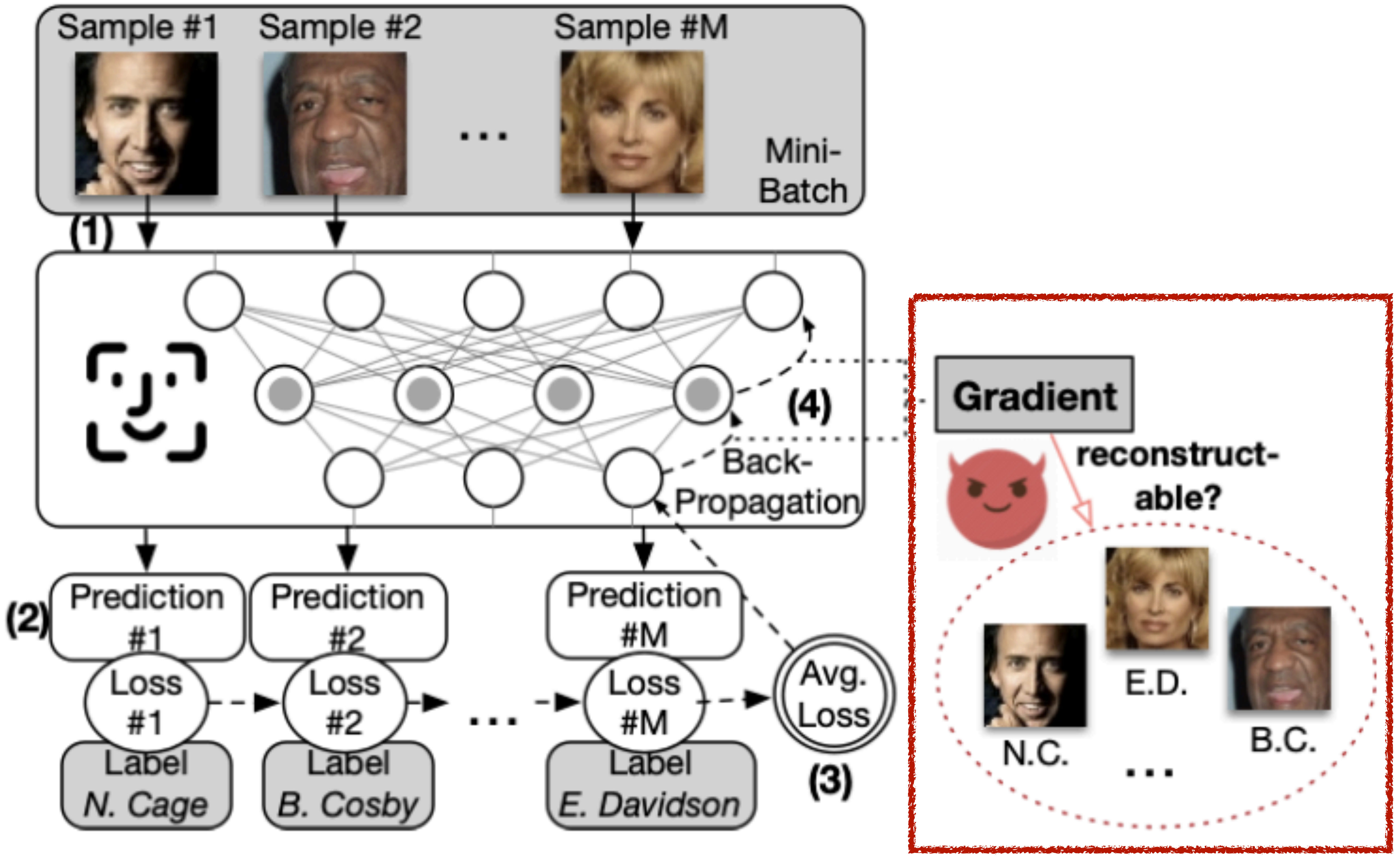
School of Computer Science

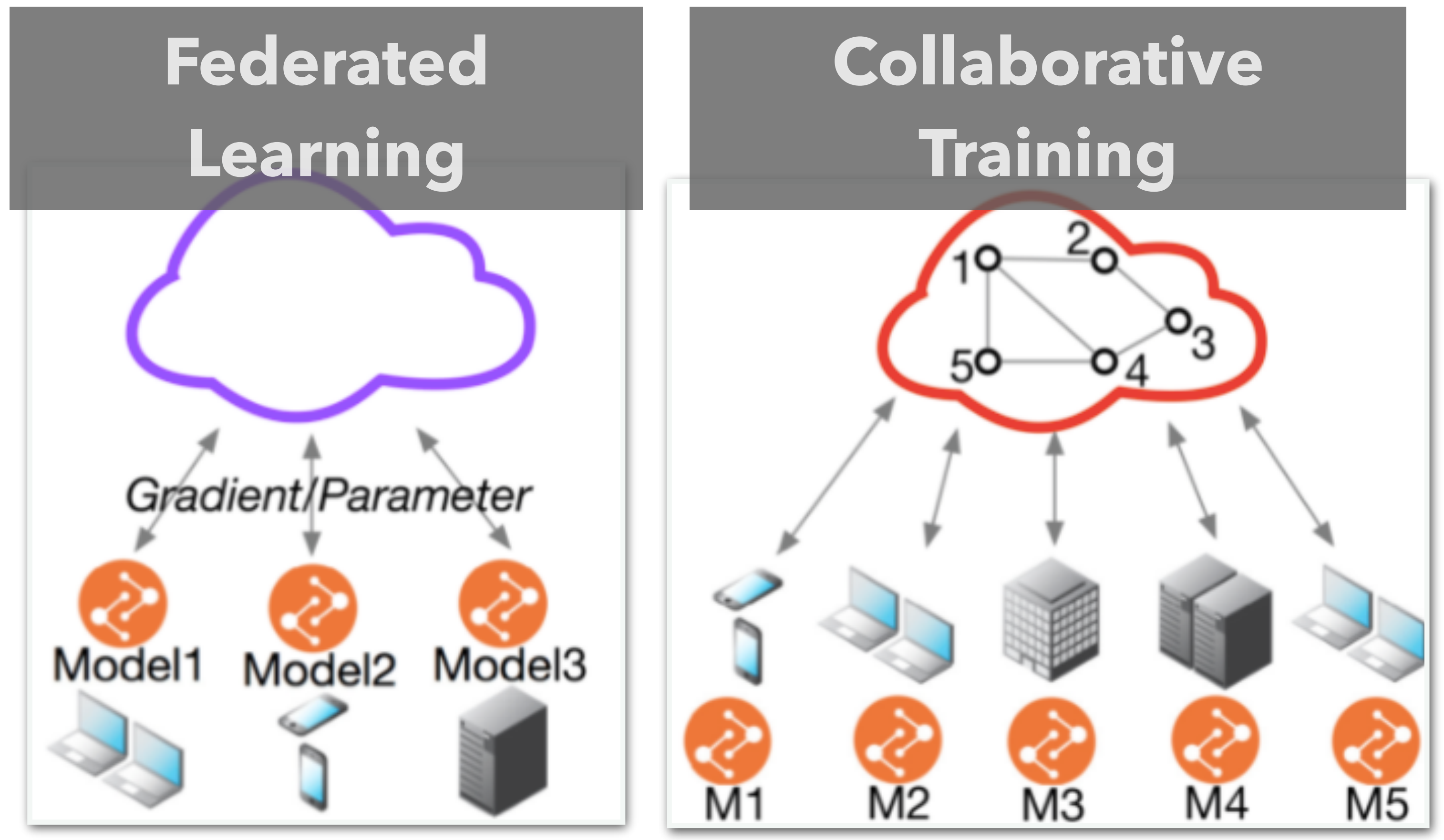
Fudan University



Threats of Data Reconstruction Attacks

- "*The Achilles Heel*" of **Privacy-Preserving Distributed Learning**





- Is such an averaging process **Invertible**?

$$((x_1, y_1), \ldots, (x_M, y_M)) \rightarrow \frac{1}{M} \sum \nabla_\theta l(f_\theta(x_i), y_i)$$

Private Samples **Average Grad**

Existing Attacks Solve the Gradient Matching Problem via Optimization

- **The Gradient Matching Problem**

$$\min_{\{(X_m, Y_m)\}_{m=1}^{M}} d(\frac{1}{M}\nabla_\Theta \ell(f_\Theta(X_m), Y_m), \overline{G})$$

Dummy Inputs/Labels

Distance Metric

Attack Instances

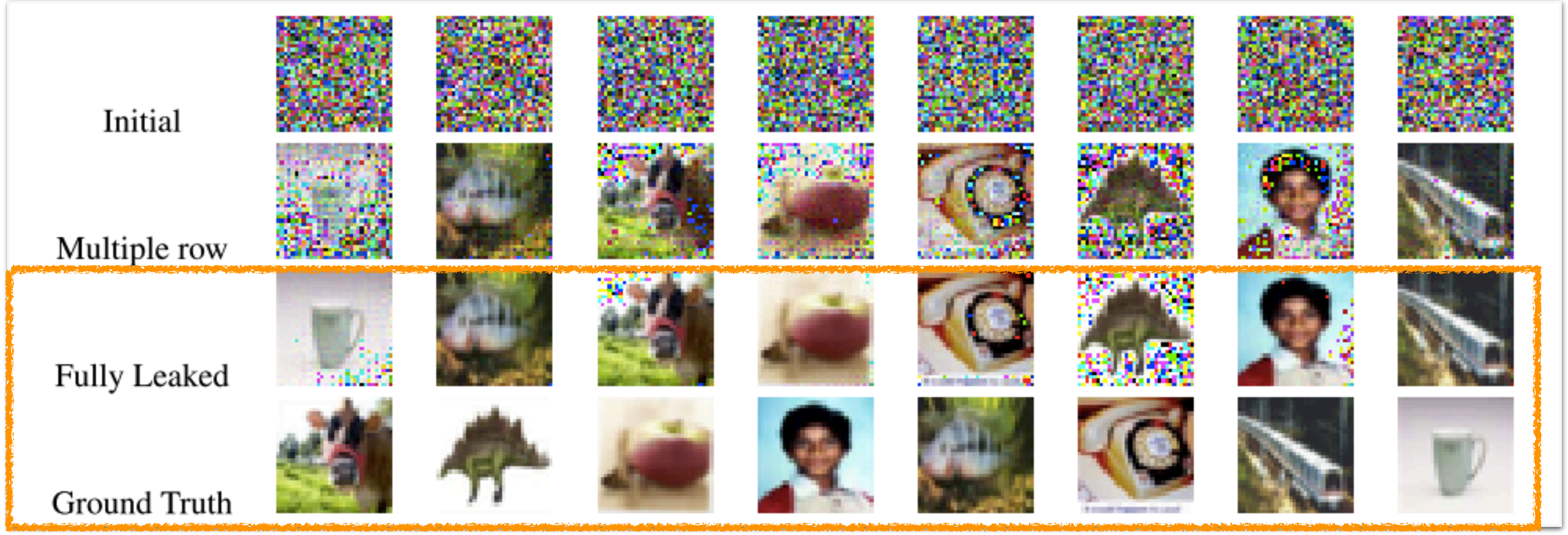
- **DLG** [Zhu et al.; NIPS'19]: $d(G_i, \overline{G}_i) = \|G_i - \overline{G}_i\|_2$, L-BFGS;
- **Inverting** [Geiping et al.; NIPS'20]: $d(G_i, \overline{G}_i) = (1 - \cos(G_i, \overline{G}_i))$, Adam;
- **GradInversion** [Yin et al.;CVPR'21]: *BatchNorm statistics as the* prior;
- ...

Attacker's Knowledge

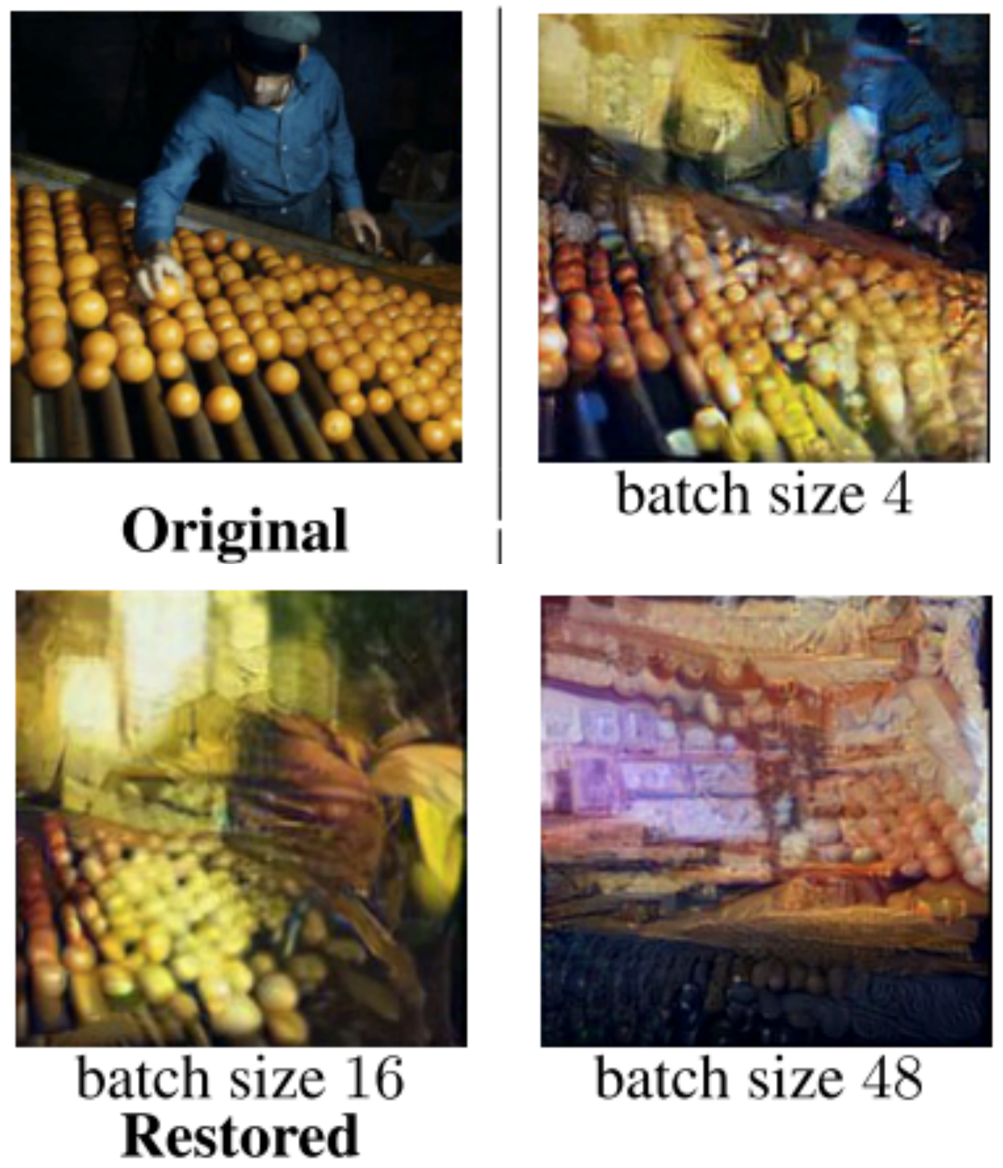
1. The Model Parameter (Θ)
2. The Average Gradient ($\overline{G}$)
3. The NN Architecture (f)
4. ~~The Batch Size (M)~~

Empirical Results and The Mysteries

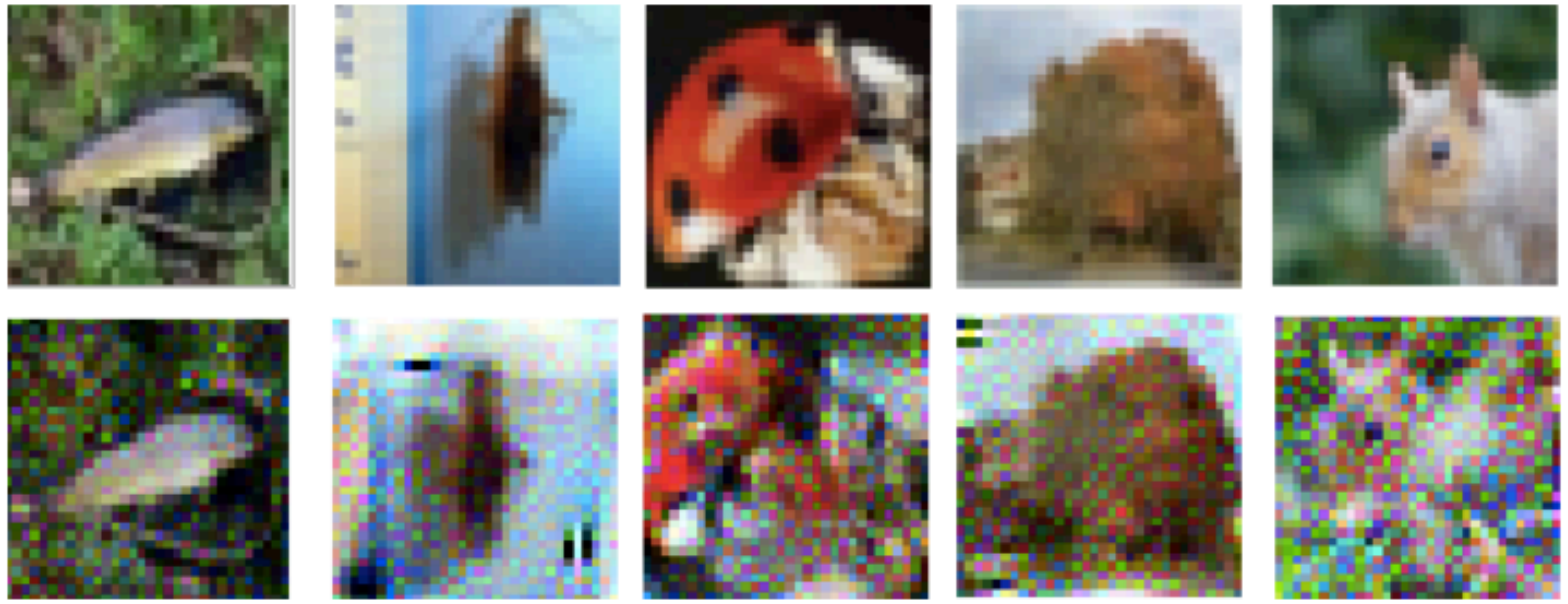
- **DLG Results (Batch Size M = 8, ResNet-56)**



- **GradInversion Results**



- **Inverting Results (Batch Size M = 100, ResNet-32)**



Our Work Answers

1. **How <u>the separation from the average gradient</u> is possible?**
2. **What factors influences data reconstruction attacks?**

Exploiting the ExANs in ReLU Networks

**ExAN* = Exclusively Activated Neurons (dubbed by us)

○ ***Def.*: ExAN at the i-th ReLU layer**

$$\sum_{m=1}^{M} [\underbrace{A_i^m}_{\text{activation pattern}}]_j = 1$$

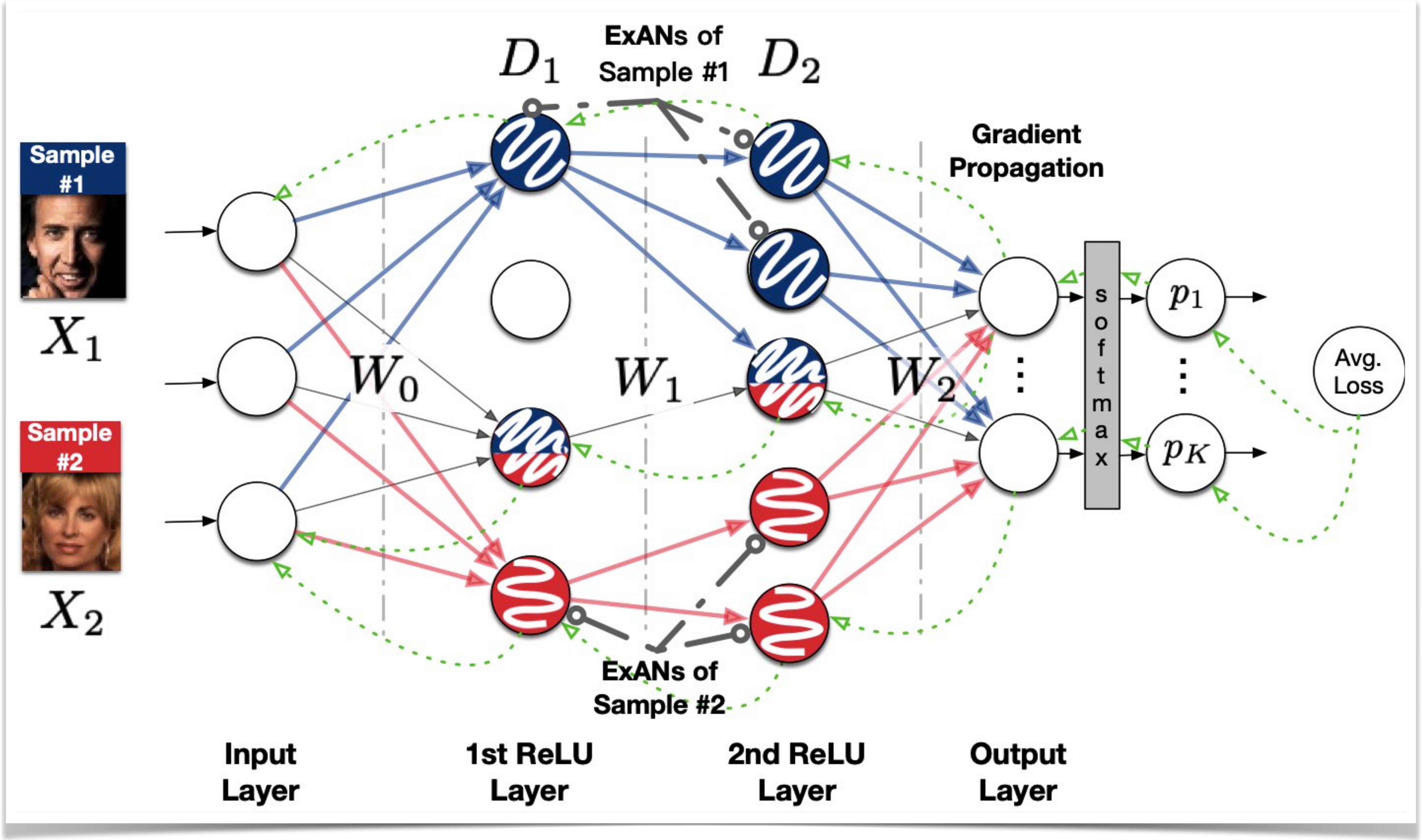
- **First ReLU Layer**

$$A_1^1 = (\boxed{1} \quad 0 \quad 1 \quad 0)$$
$$A_1^2 = (0 \quad 0 \quad 1 \quad \boxed{1})$$

- **Second ReLU Layer**

$$A_2^1 = (\boxed{1} \quad \boxed{1} \quad 1 \quad 0 \quad 0)$$
$$A_2^2 = (0 \quad 0 \quad 1 \quad \boxed{1} \quad \boxed{1})$$

- **Property:** Backward signals (dashed lines) only flow via neurons activated in forward computation (solid lines)



Neuron Exclusivity Analysis on Data Reconstruction

- **Gradient Matching Problem -> Gradient Equation**

$$\min_{\{(X_m, Y_m)\}_{m=1}^{M}} d\left(\frac{1}{M}\nabla_\Theta \ell(f_\Theta(X_m), Y_m), \overline{G}\right) \longleftrightarrow \frac{1}{M}\nabla_\Theta \ell(f_\Theta(X_m), Y_m) = \overline{G}$$

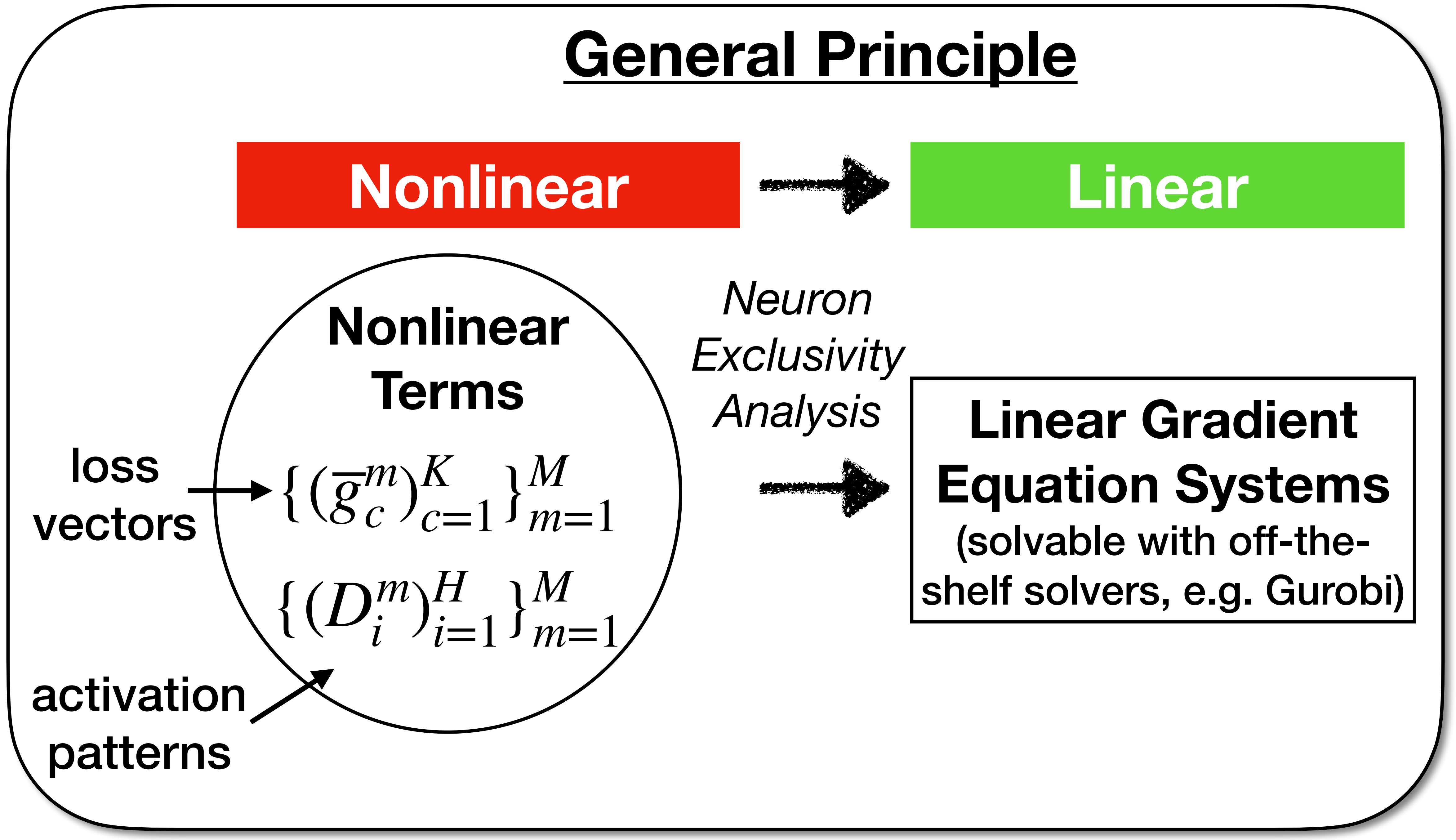
a complex <u>nonlinear</u> matrix equation

- Let's consider the (unbiased) **FCN**: $f(X; W_0, \ldots, W_H) = W_H D_H \ldots W_1 D_1 W_0 X$

$$\overline{G}_i = \sum_{c=1}^{K} \overline{g}_c \frac{\partial f_c}{\partial W_i} \xrightarrow{\text{nonlinearity}} \overline{g}_c = \begin{cases} p_c & \textbf{if } c != Y_m \\ p_c - 1 & \textbf{if } c = Y_m \end{cases}$$

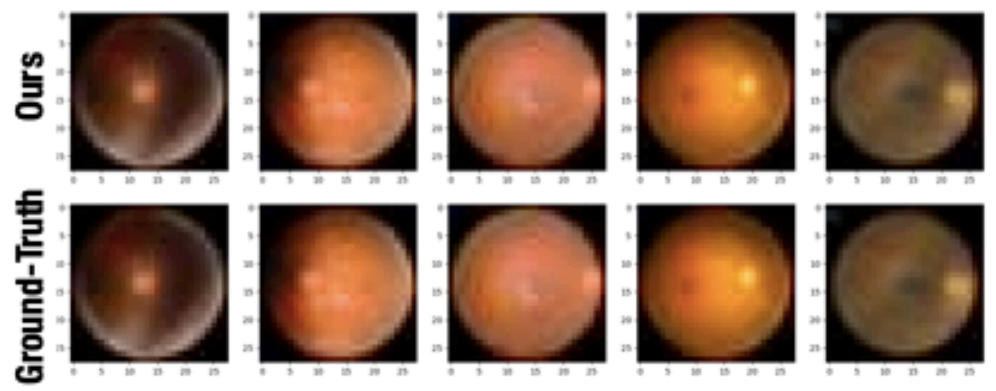
nonlinearity $\downarrow$

$$\frac{\partial f_c(X)}{\partial W_i} = (D_i W_{i-1} \ldots W_0 X)([W_H]_c^T D_H \ldots W_{i+1} D_{i+1})$$

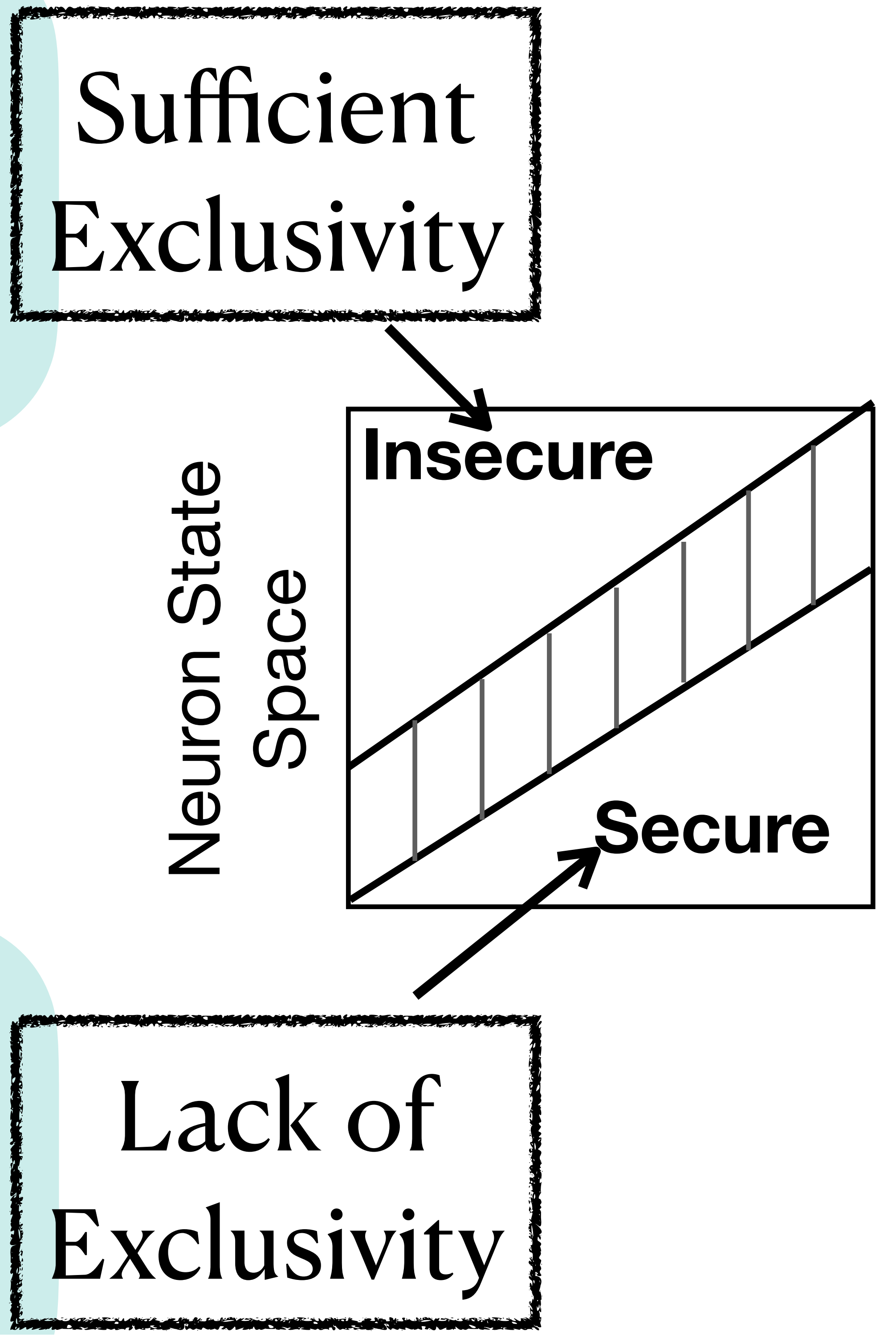


Main Results on Security Boundary

- **(Attack Side)** When a mini-batch satisfies the following ExAN condition, all the samples can **be analytically reconstructed with provably low error**.

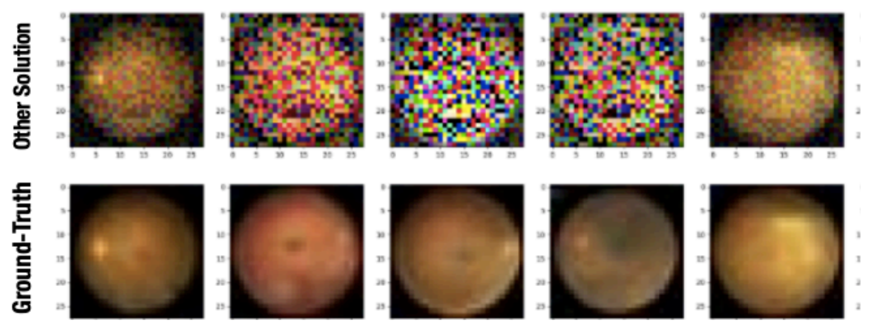


- **In the last ReLU layer, each sample has $\geq$ 2 ExANs** => reconstructing $\bar{g}_c^m$
- **In the remaining ReLU layers, each sample has $\geq$ 1 ExAN** => reconstructing D_i^m



- **(Defense Side)** When a mini-batch satisfies the following condition, there exists infinitely many batches which share the same gradients.

- **In the first ReLU layer, each sample has 0 ExAN** => Impossibility of Reconstruction (due to infinitely many candidate solutions)

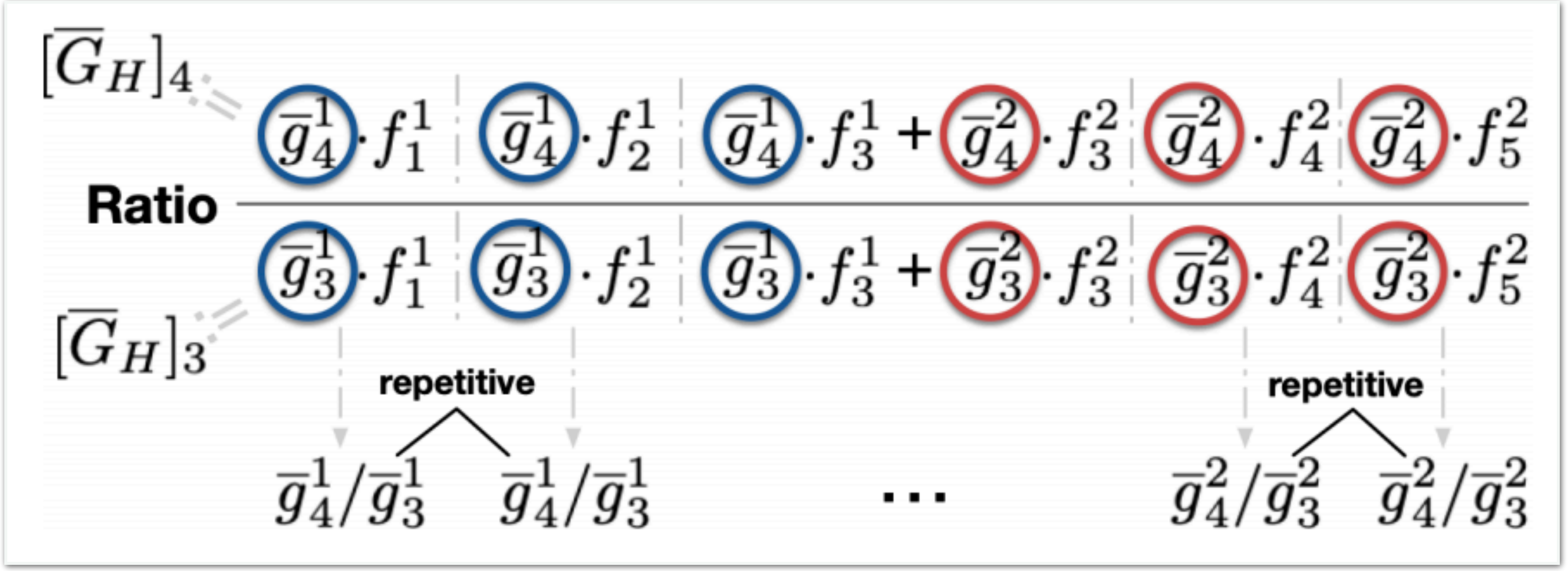


Reconstructing the Loss Vectors

- **Inspecting the gradient equation of the last ReLU layer.**

$$[\overline{G}_H]_c = \frac{1}{M}\sum_{m=1}^{M} \overline{g}_c^m f_{H-1}^m \qquad (8)$$

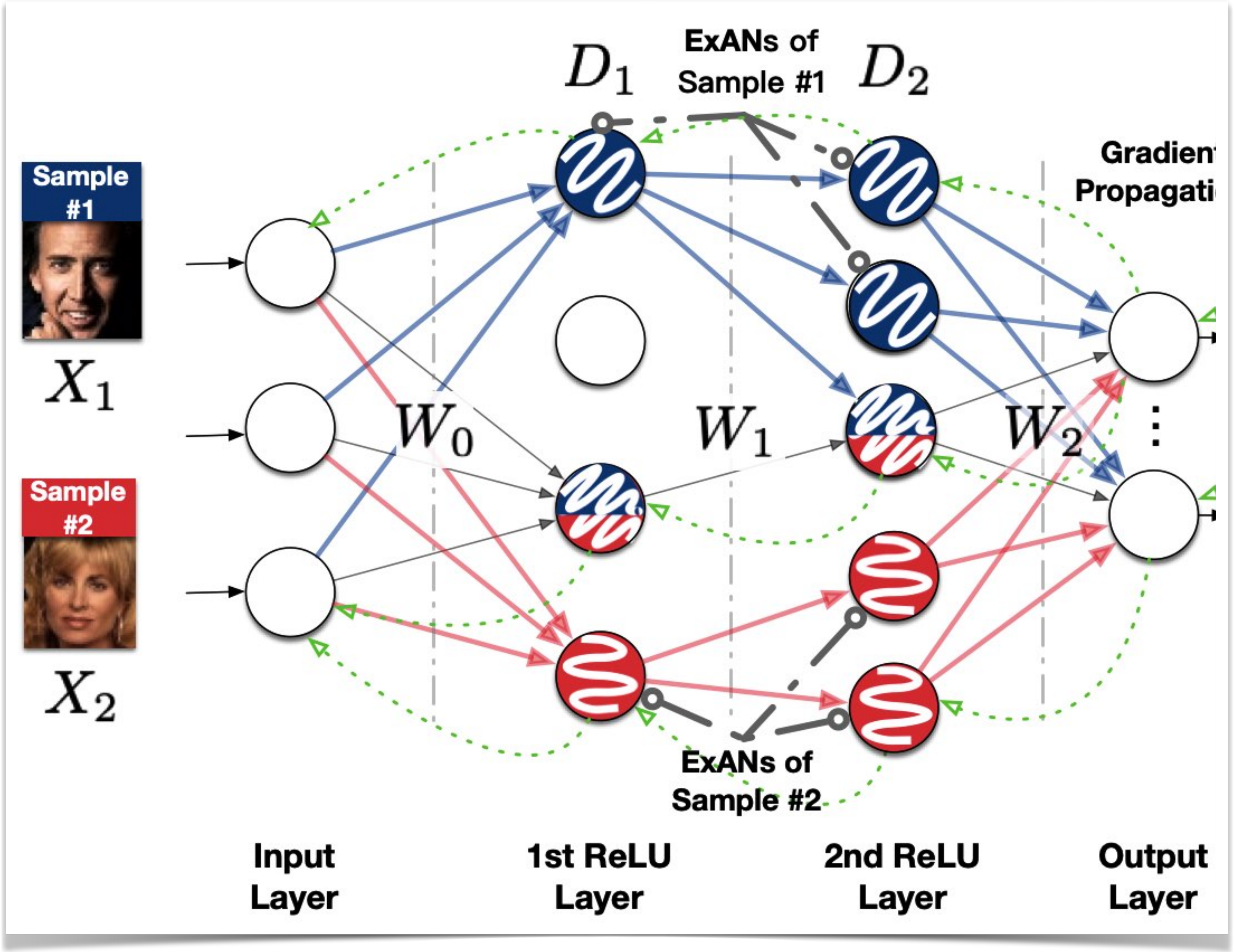
- **Observation I**: If the m-th sample has at least 2 ExANs at the last ReLU layer, there are always **2 more repetitive values** in the ratio vector $[\overline{G}_H]_j/[\overline{G}_H]_k$, which equals to $\overline{g}_j^m/\overline{g}_k^m$ 。



- **With the estimated ratios $\overline{g}_j^m/\overline{g}_k^m$**
  - Determine the labels based on the signs.
  - Determine the range of $\overline{g}_i^1$ based on the constraints $\sum_c p_c = 1$

Reconstructing the Activation Patterns

- **Observation II.** If the m-th sample has 1 ExAN at the i-th layer, then **the non-vanishing gradients to the precedent layer indicate the ExANs at the (i-1)-th layer**, i.e., D_{i-1}^m



- **Recursion:** If the (i-1)-th layer has at least one ExAN, the reconstruction can be done for the (i-2)-th layer ... until the first ReLU layer.

- The 2nd ReLU

$$A_2^1 = (1 \quad 1 \quad 1 \quad 0 \quad 0)$$
$$A_2^2 = (0 \quad 0 \quad 1 \quad 1 \quad 1)$$

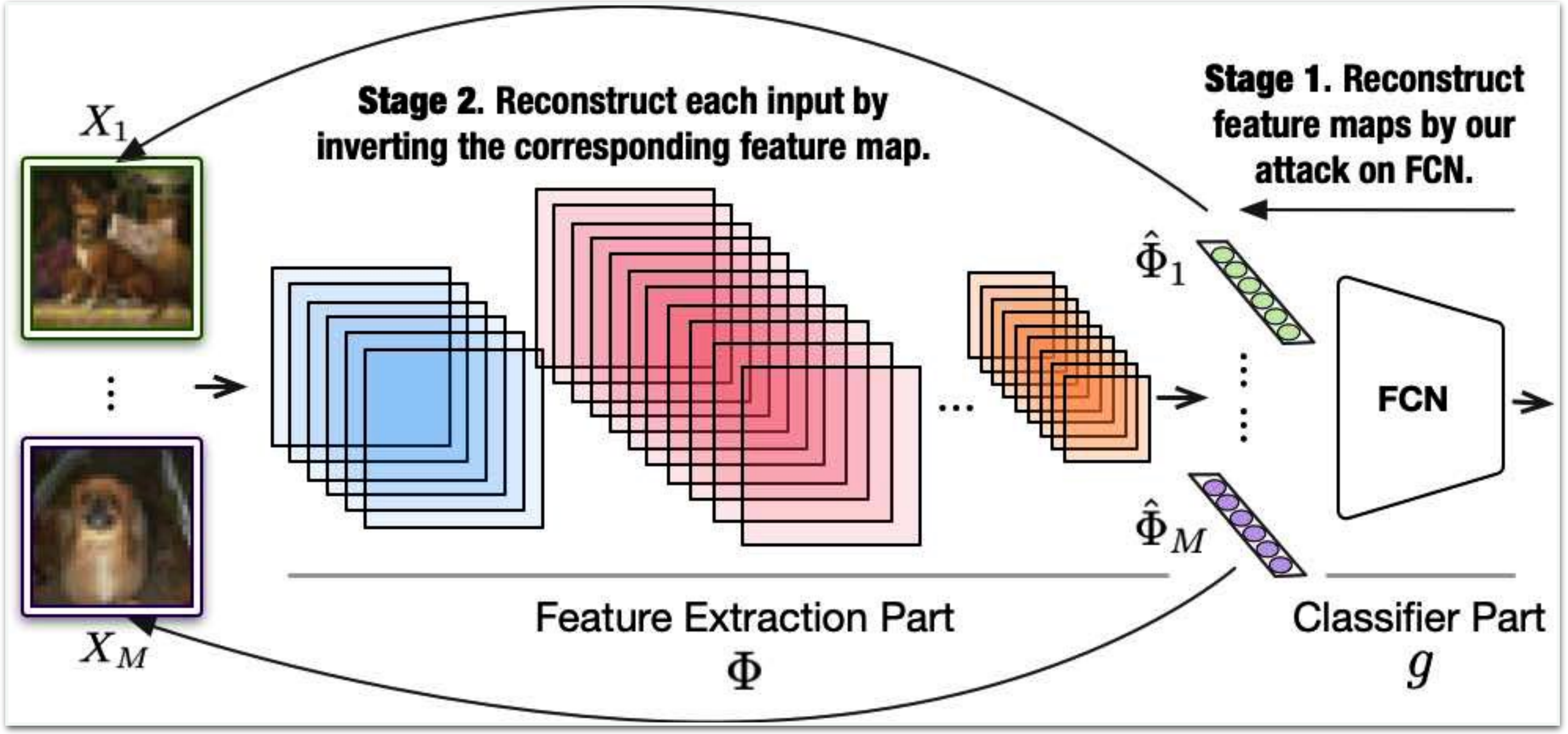
- The 1st ReLU

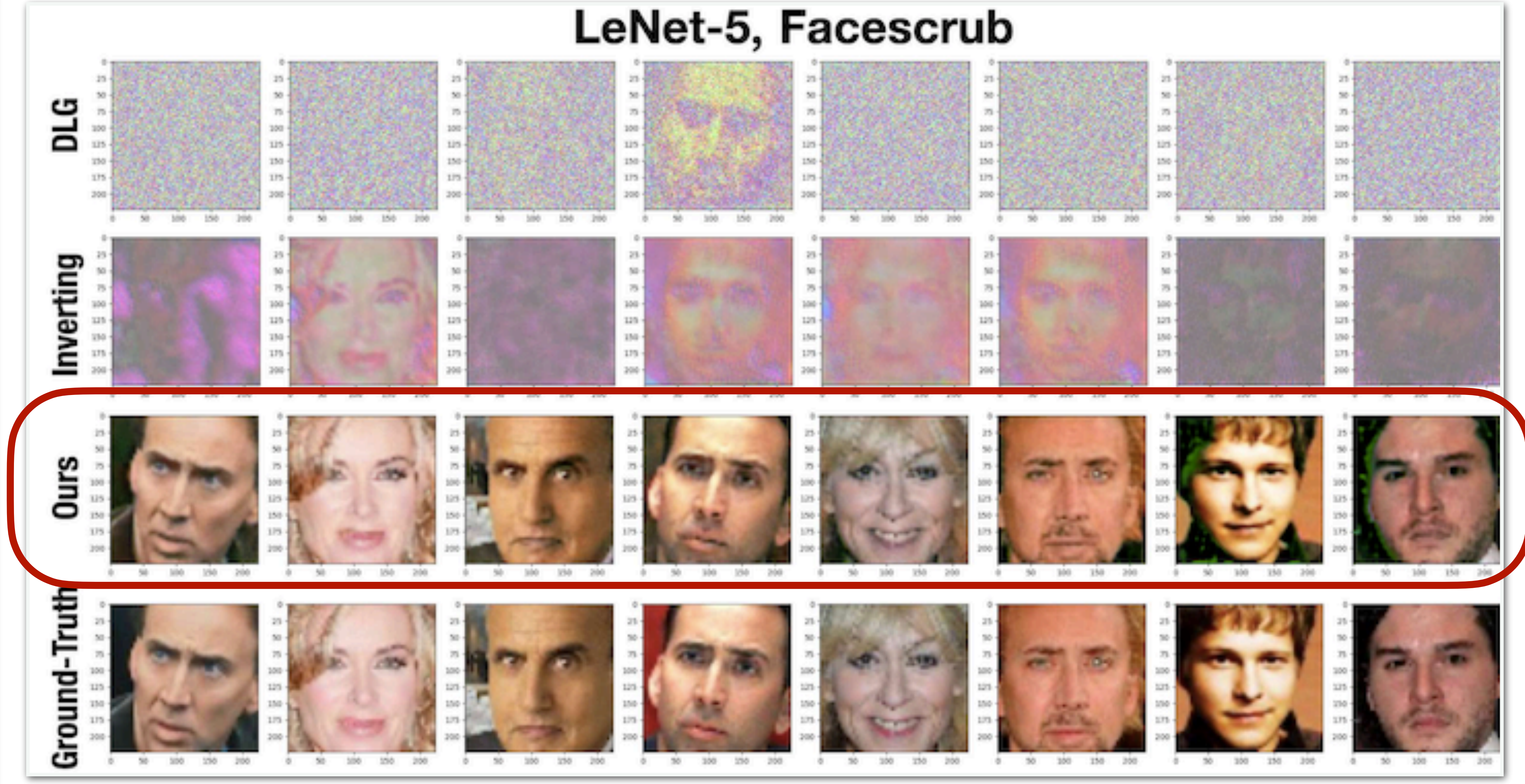
$$A_1^1 = (1 \quad 0 \quad 1 \quad 0)$$
$$A_1^2 = (0 \quad 0 \quad 1 \quad 1)$$

Extension to Deep ConvNets

- 1 × Gradient Matching Problem -> M × **Feature Matching Problem**

$$\mathop{\textbf{arg}\min}_{X_m} \|\Phi(X_m) - \hat{\Phi}_m\|$$



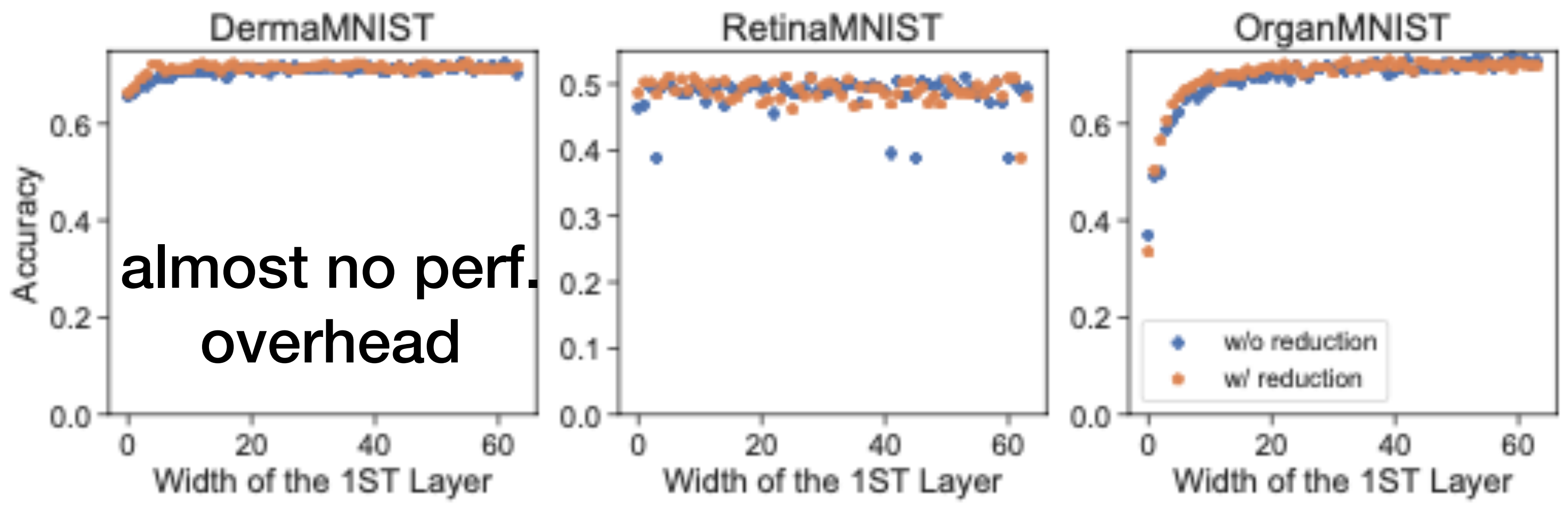


Defense Side Results

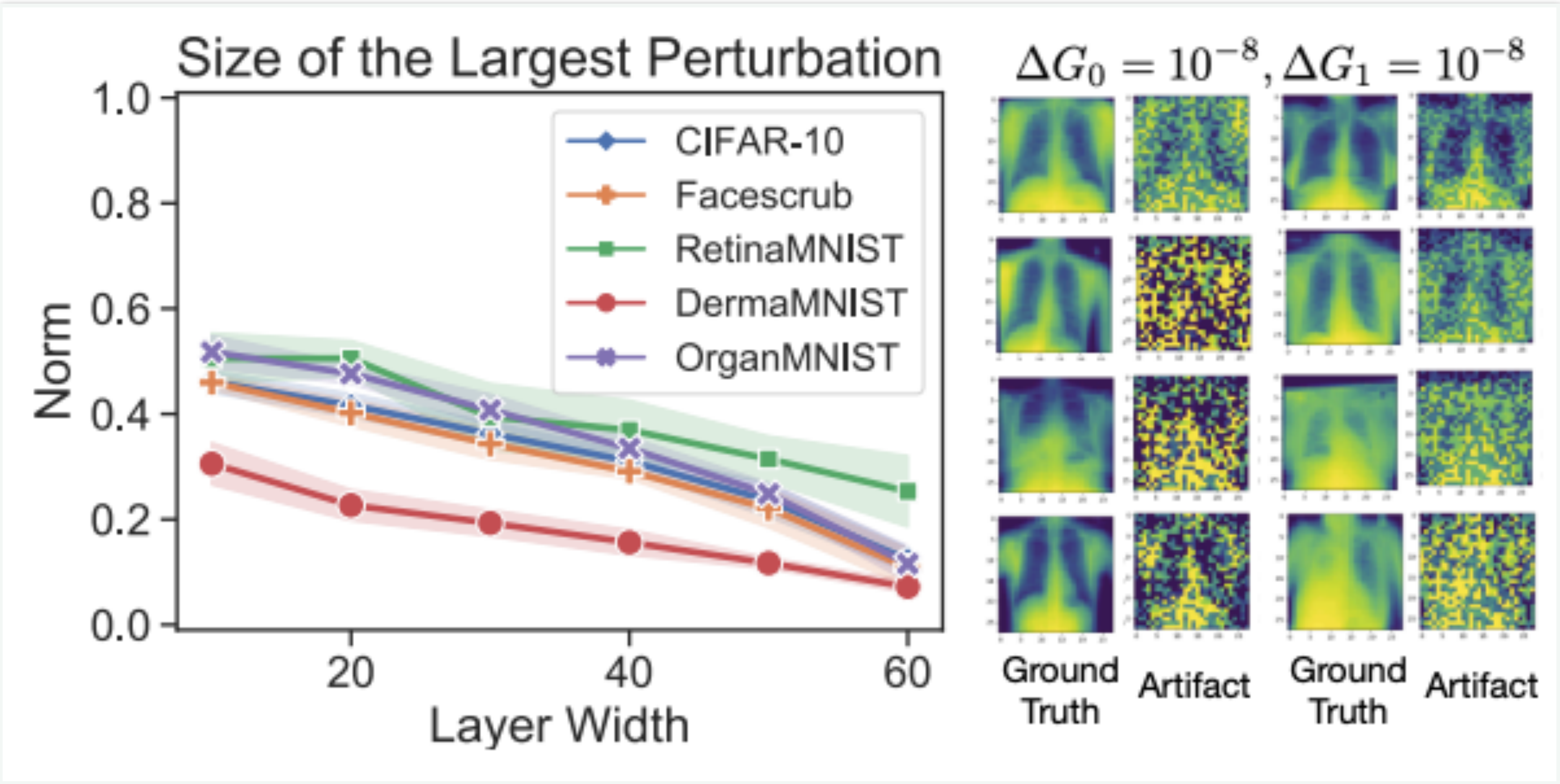
- A Moderate Architectural Change for **Exclusivity Elimination**

$$W_H\sigma(W_{H-1}\dots(W_1\hat{\sigma}(W_0X+b_0)+b_1)\dots+b_{H-1})+b_H$$

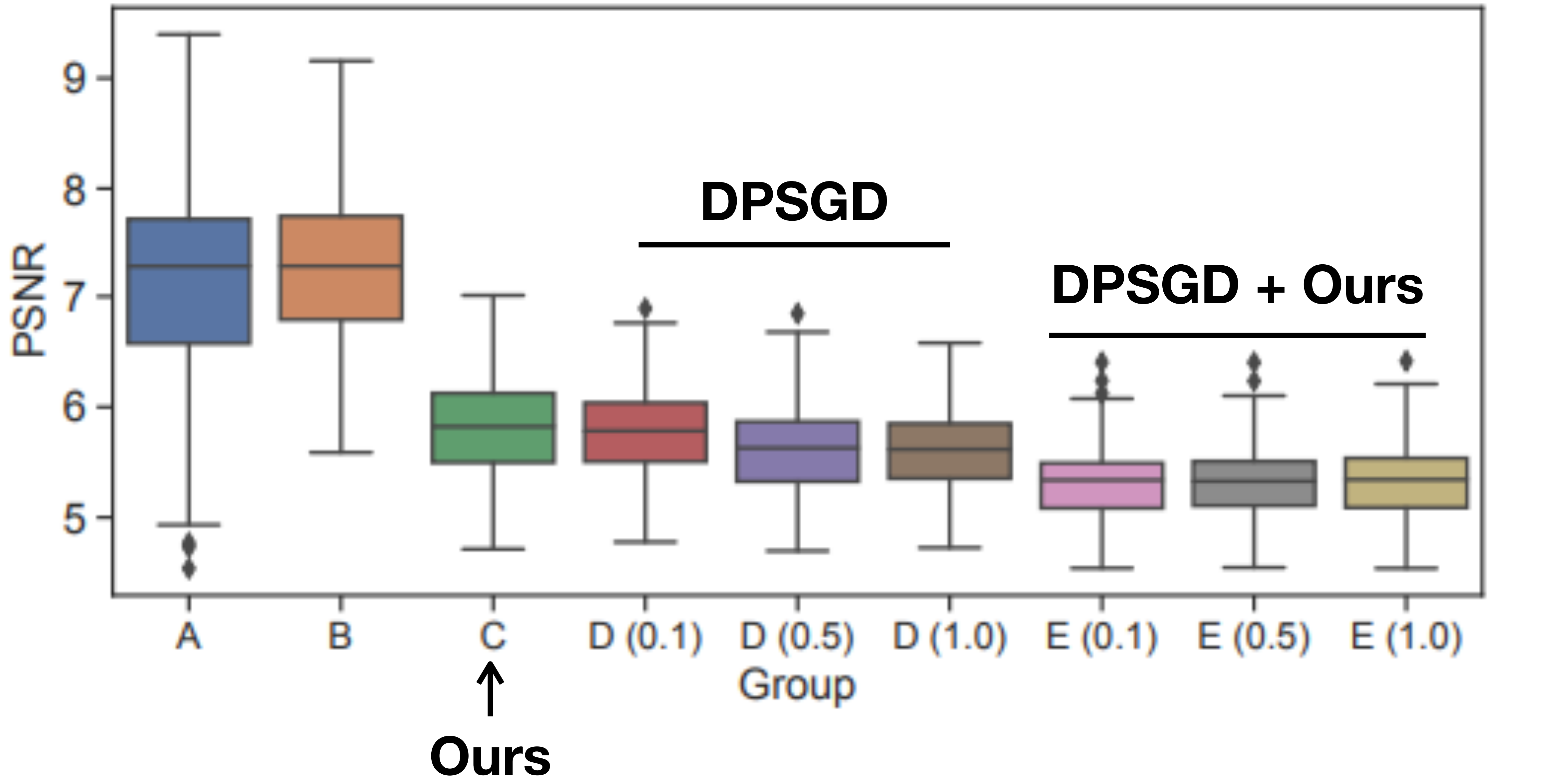
remove the 1st ReLU



- **Infinitely Many Solutions when $M \leq d_1$**

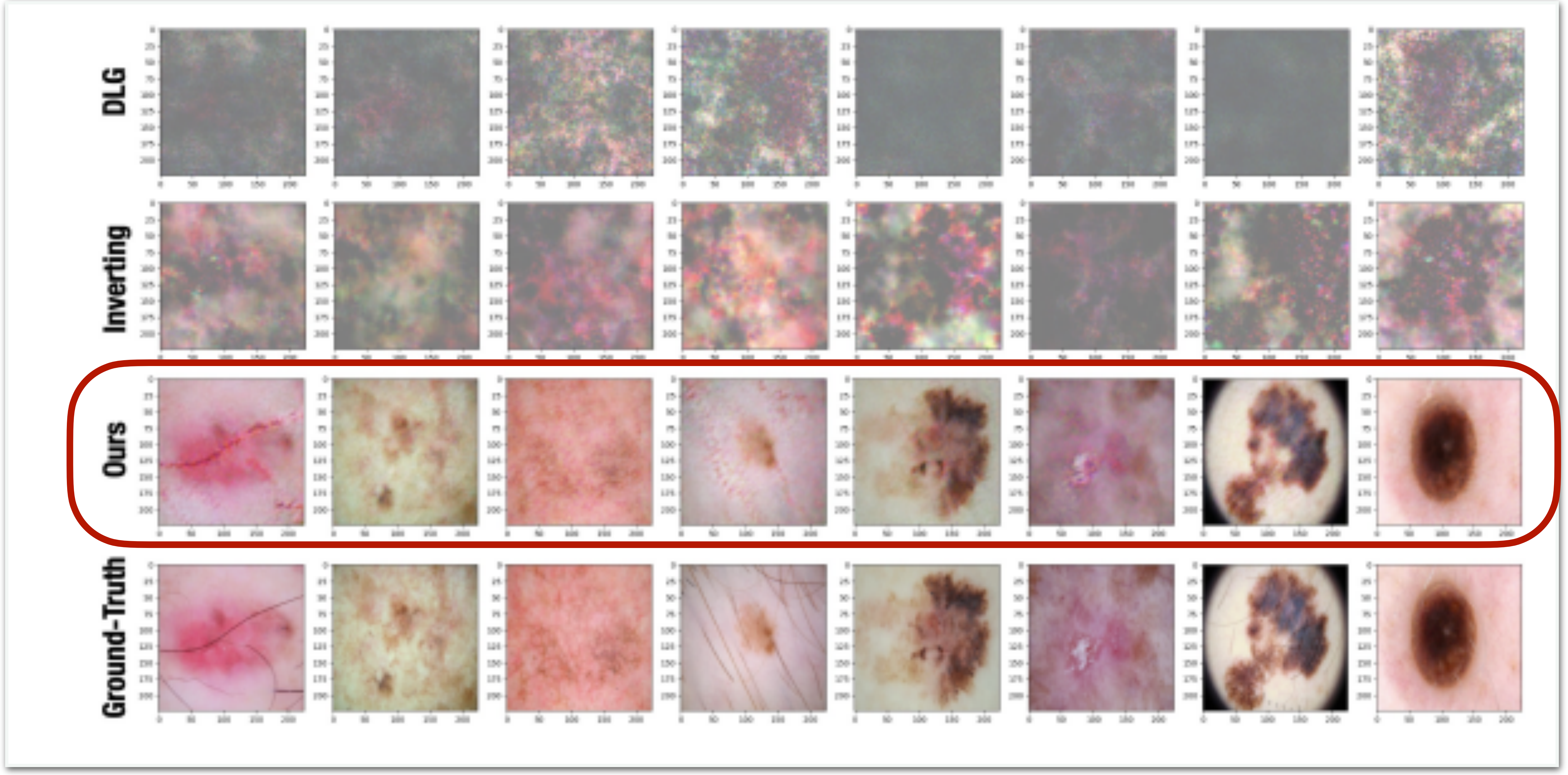


- **Combo with Other Obfuscation (e.g. DPSGD)**

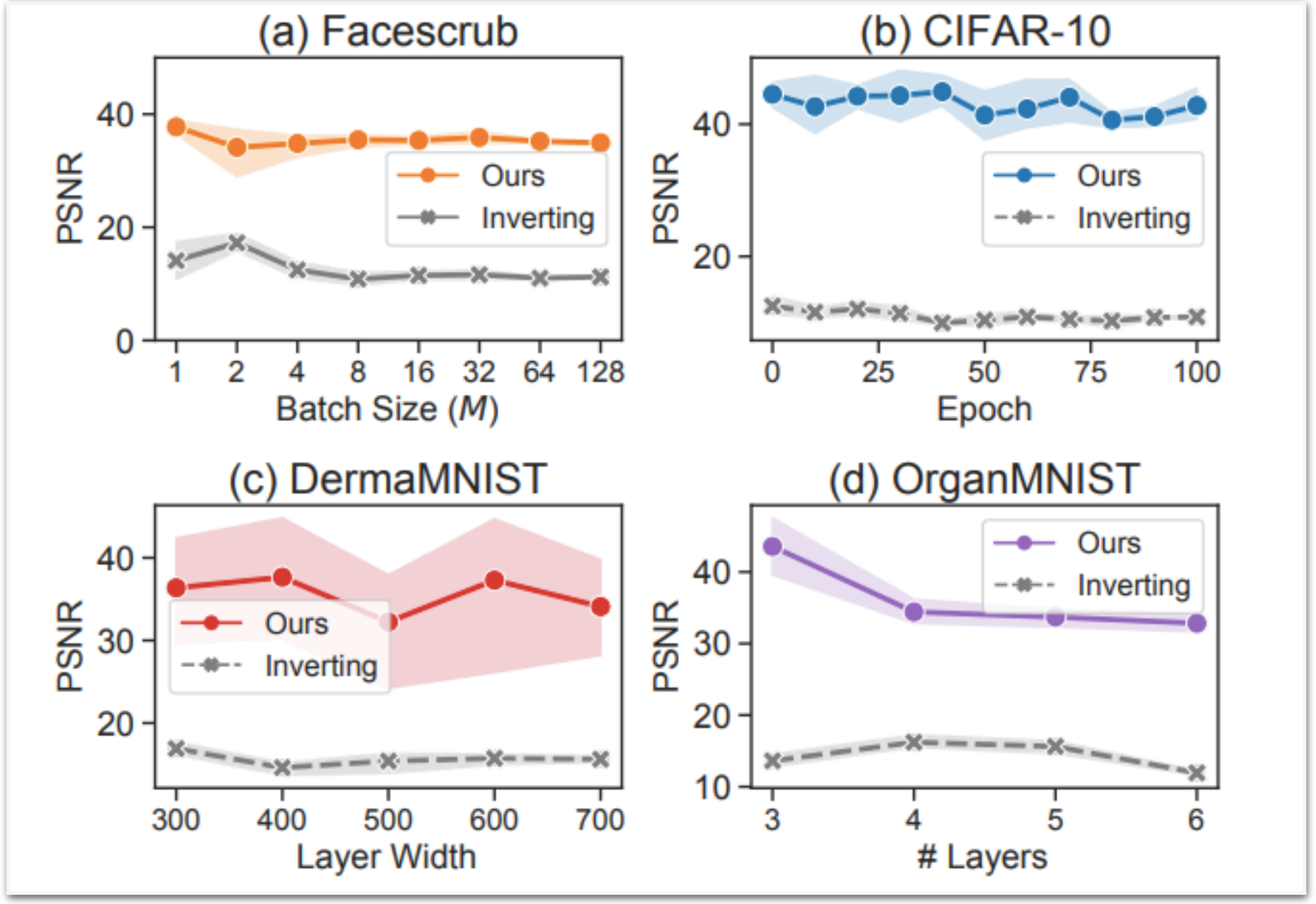


More Evaluation Results

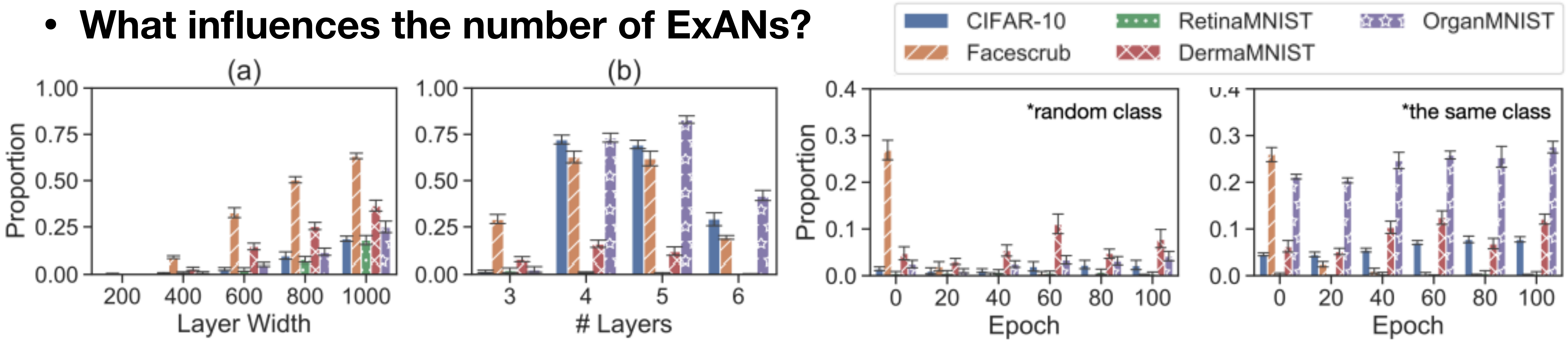
- **Comparison of Reconstruction on VGG13**



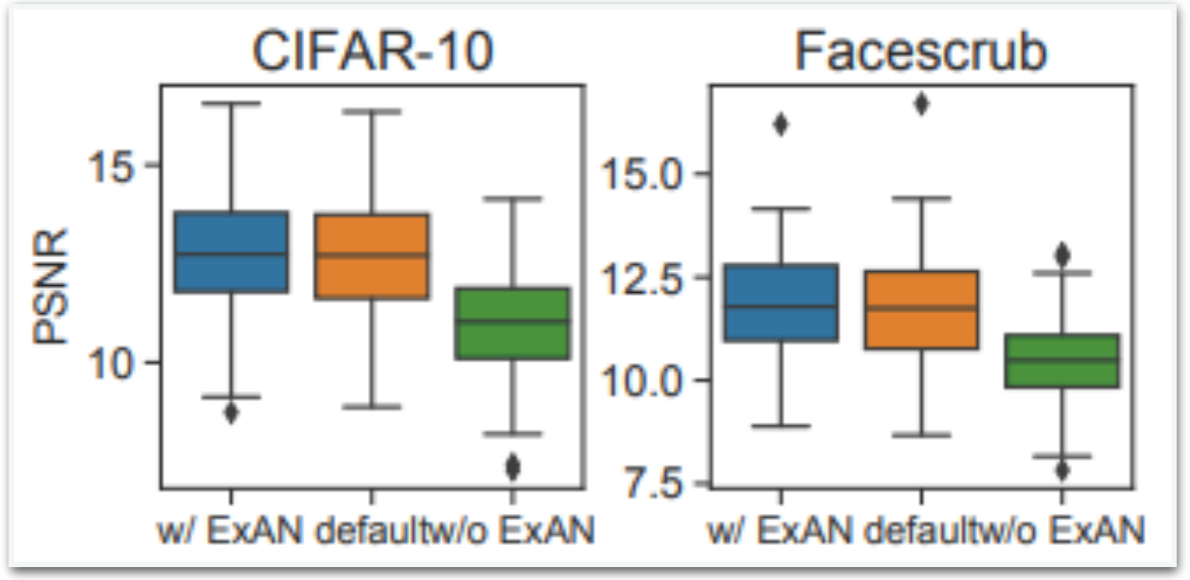
- **Insensitivity to Impact Factors when w/. ExAN)**



- **What influences the number of ExANs?**



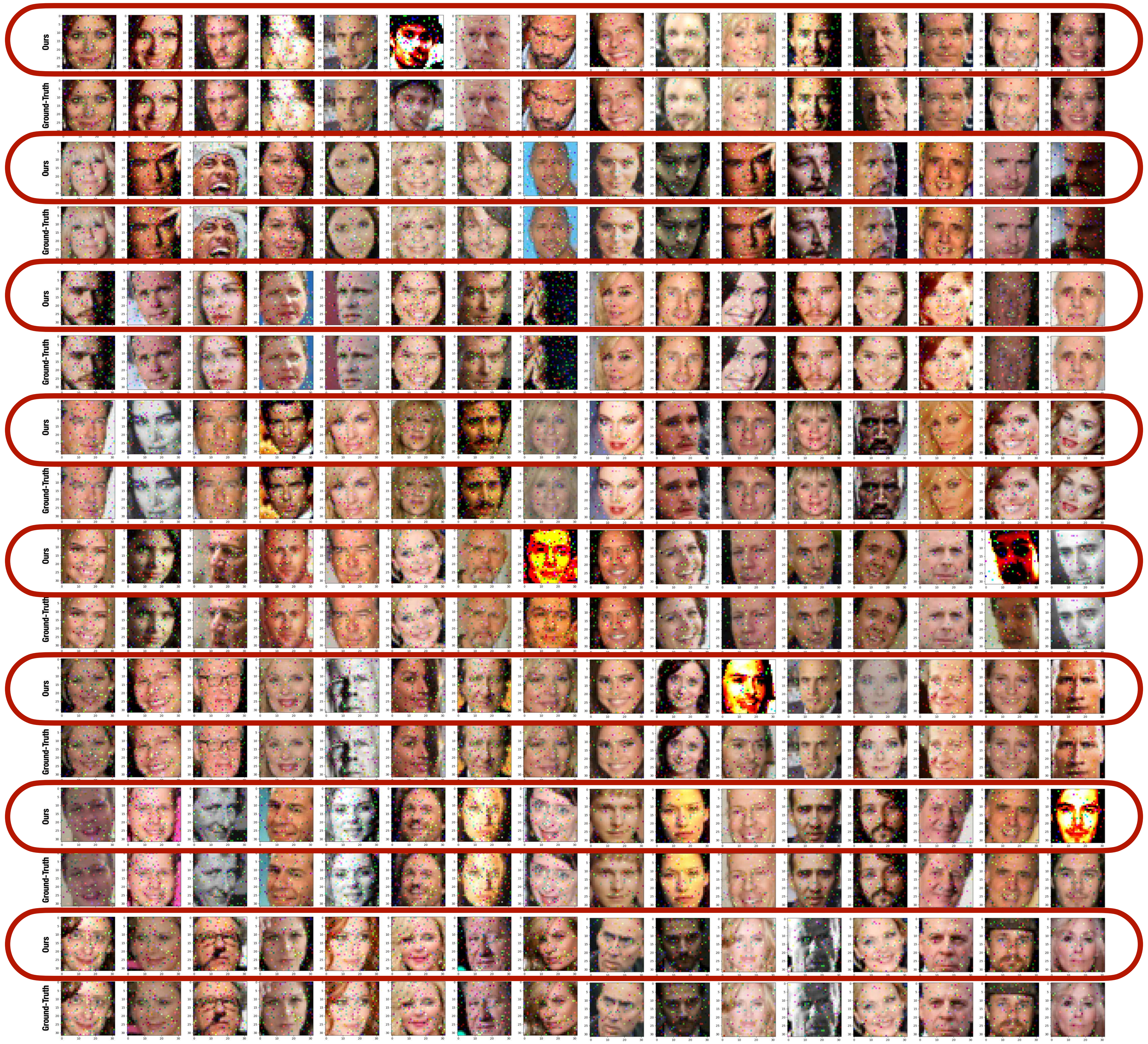
- **ExAN Matters**



More Evaluation Results

- Precise reconstruction of **almost however large** batch when sufficient exclusivity is satisfied

Facescrub
3-Layer FCN
Batch Size 128



Conclusions and Future Directions

Phenomenon

Layer Width | Network Depth | Label Composition | Batch Size | Training Epoch | ...

Common Cause

Neuron Exclusivity

- Last ReLU $\geq$ 2 ExANs
- Other ReLU $\geq$1 ExAN

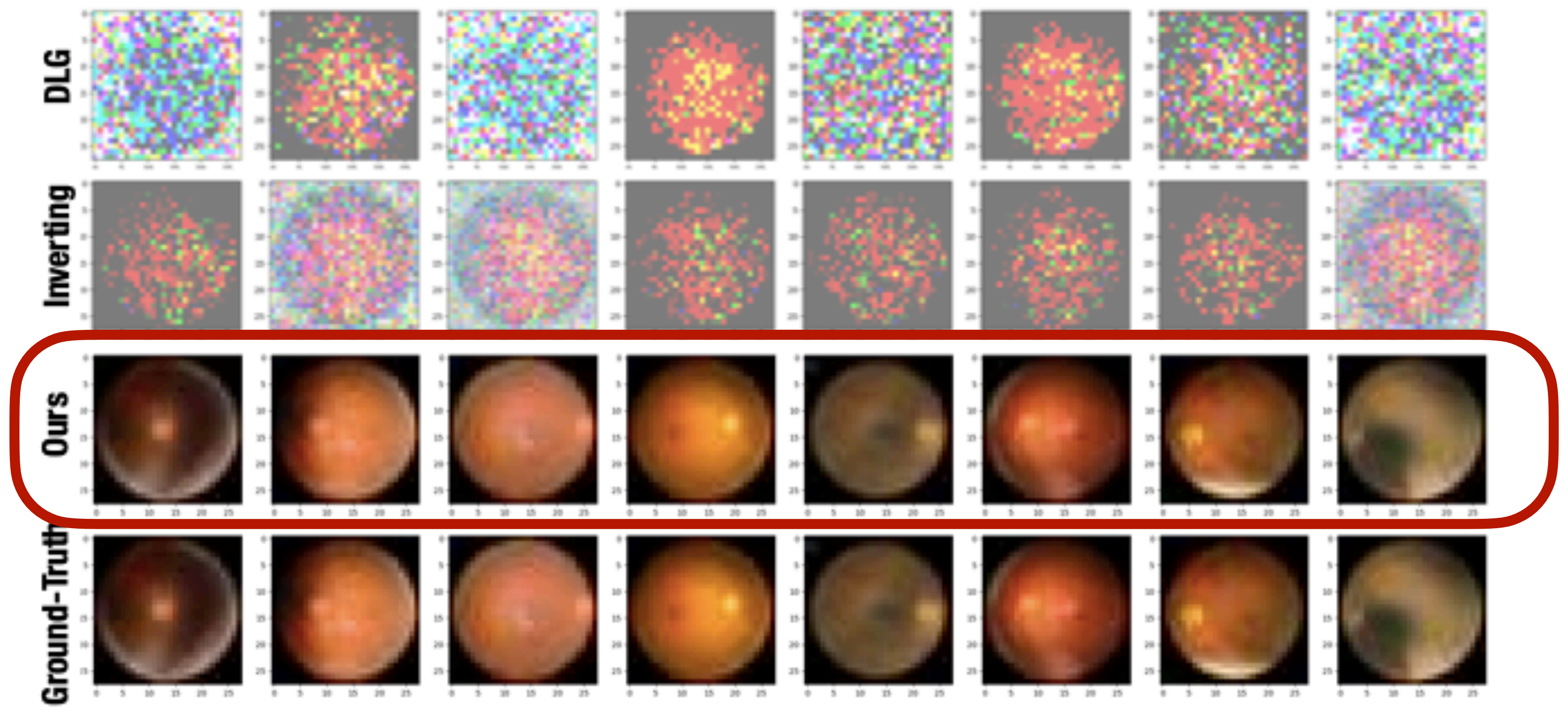
Sufficient

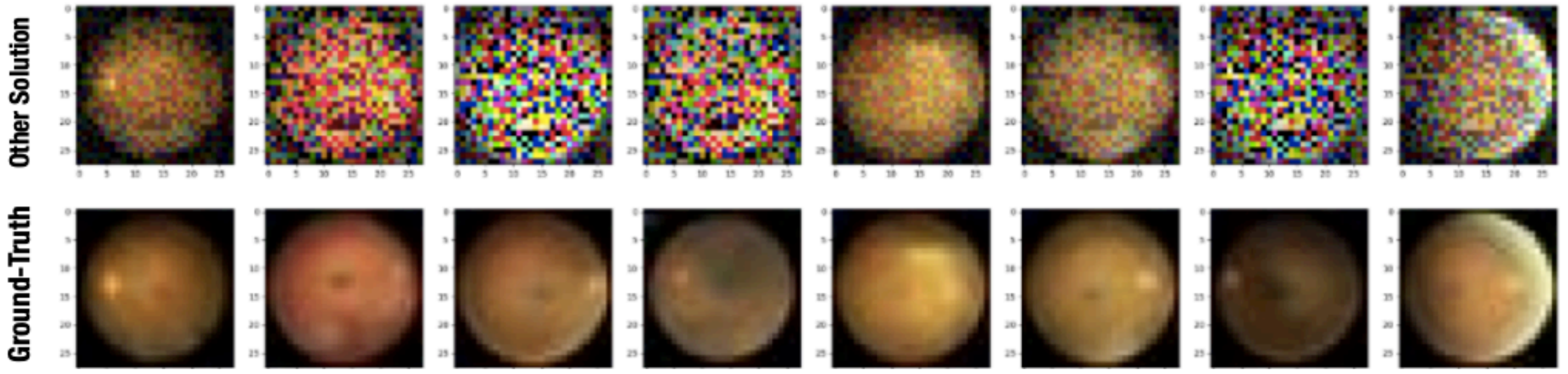
Lack

- The 1st ReLU = 0 ExAN

Data Reconstruction

Guaranteed Reconstruction Attack |— Gap —| Impossibility of Unique Reconstruction





$\Delta G_0 = 4e^{-9}$ **and** $\Delta G_1 = 1e^{-8}$

Thank you for your Audience!

For more details, welcome to follow our paper.

31ST USENIX SECURITY SYMPOSIUM

Exploring the Security Boundary of Data Reconstruction via Neuron Exclusivity Analysis

Xudong Pan, Mi Zhang✉, Yifan Yan, Jiaming Zhu, Min Yang✉
Fudan University, China
{xdpan18, mi_zhang, yanyf20, 19210240146, m_yang}@fudan.edu.cn

Abstract

Among existing privacy attacks on the gradient of neural networks, *data reconstruction attack*, which reverse engineers the training batch from the gradient, poses a severe threat on the private training data. Despite its empirical success on large architectures and small training batches, unstable reconstruction accuracy is also observed when a smaller architecture or a larger batch is under attack. Due to the weak interpretability of existing learning-based attacks, there is little known on