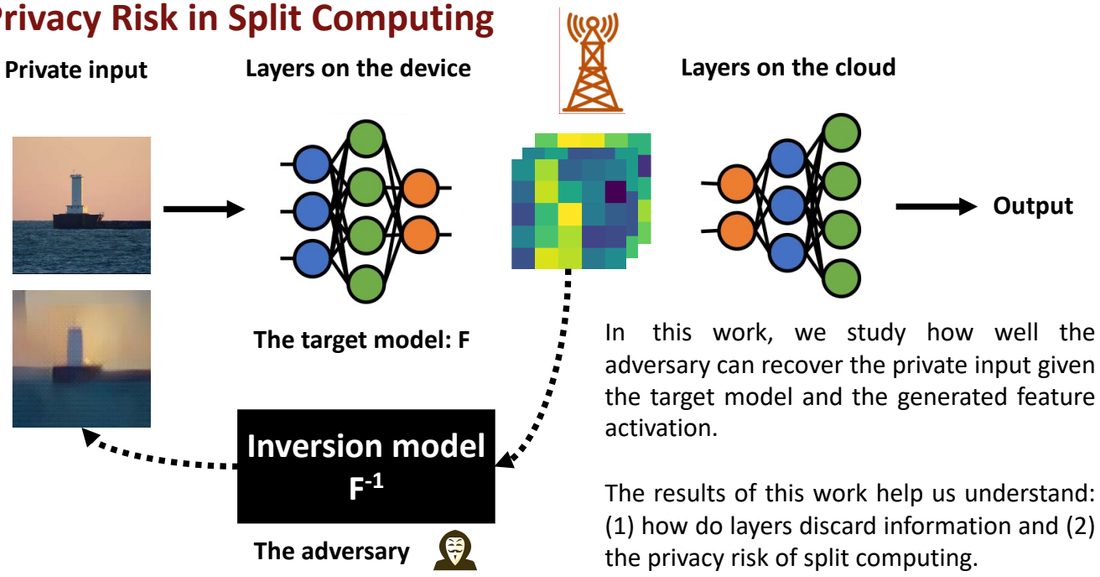
# Auditing Privacy Protection in Split Computing via Data-Free Model Inversion Attack

Xin Dong[1], Hongxu Yin[2], Jose M. Alvarez[2], Jan Kautz[2], Pavlo Molchanov[2]
[1]Harvard University, [2]NVIDIA

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

## Privacy Risk in Split Computing

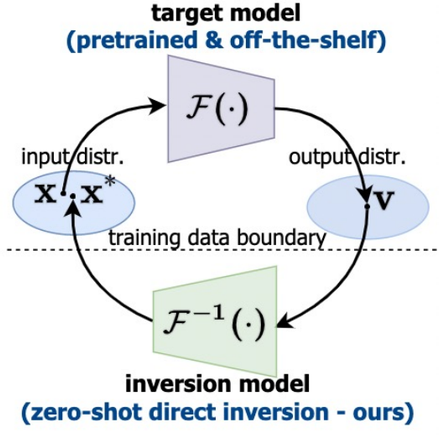



In this work, we study how well the adversary can recover the private input given the target model and the generated feature activation.

The results of this work help us understand: (1) how do layers discard information and (2) the privacy risk of split computing.

## Problem Setting

1. **Generalizability**: off-the-shelf pre-trained networks (e.g., ResNet, RepVGG, ResNet-SelfSup and SNGAN)
2. **Efficiency**: we consider direct model inversion which takes feature activation as input and directly generates the recovered image.
3. **No Real Data**: the adversary does NOT have access to any real data to train the inversion model.
4. **Scale**: deep target model (with 20+ layers) and large-scale datasets (ImageNet and CelebA).

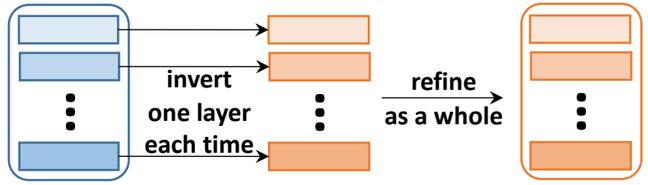



## Method: Divide-and-Conquer Inversion (DCI)

- DCI partitions the overall inversion problem into several block-wise inversion sub-problems before integrating them together for refinement,

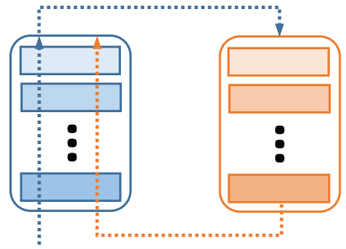
$$\mathcal{L}_{\text{layer}} = \left\| \mathbf{u}_k - \mathcal{F}_k^{-1}\left(\mathcal{F}_k\left(\mathbf{u}_k\right)\right) \right\|_1, \quad \mathbf{u}_k = \mathcal{F}_{1:(k-1)}(\mathbf{x})$$

$$\mathcal{L}_{\text{img}}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_1, \quad \mathbf{x}' = \mathcal{F}_{1:k}^{-1}\left(\mathcal{F}_{1:k}\left(\mathbf{x}\right)\right)$$



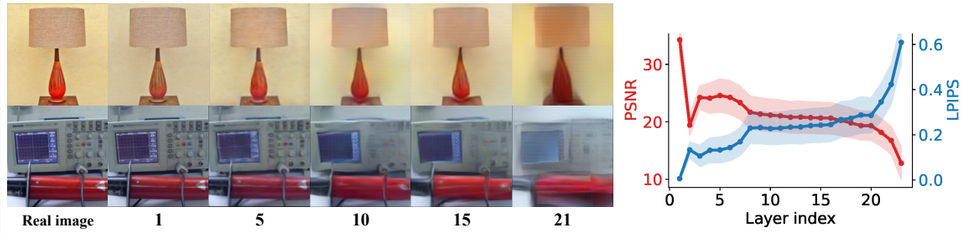

- we explore cycle consistency to measure the quality of recovered inputs by re-checking them with the target model,

$$\mathcal{L}_{\text{cyc}}(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{L} \|\mathcal{F}_{1:l}(\mathbf{x}) - \mathcal{F}_{1:l}(\mathbf{x}')\|_1$$




## Results

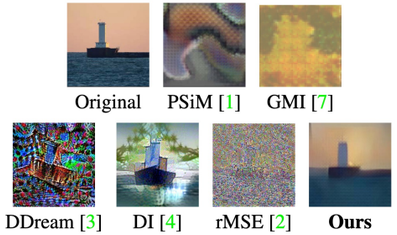
**RepVGG-A0 (Up to 21 Conv Blocks) Inversion - ImageNet**







**Comparison on RepVGG-A0**




**GAN Inversion - CelebA**


**ResNet50 Inversion - ImageNet**