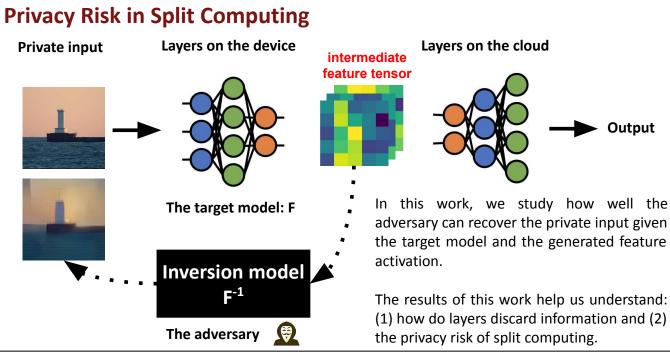


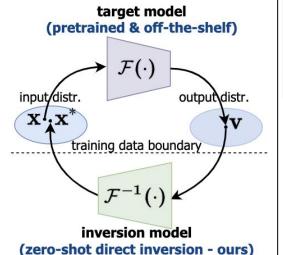
Auditing Privacy Protection in Split Computing via Data-Free Model Inversion Attack

Xin Dong¹, Hongxu Yin², Jose M. Alvarez², Jan Kautz², Pavlo Molchanov², H.T. Kung¹ ¹Harvard University, ²NVIDIA



Problem Setting

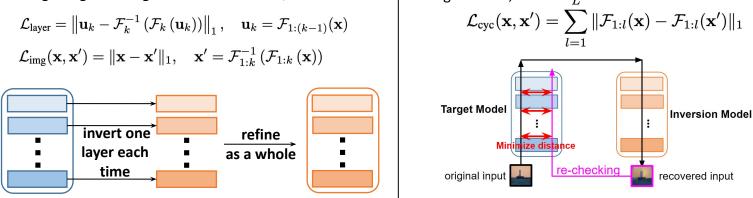
- 1. Generalizability: off-the-shelf pre-trained networks (e.g., ResNet, RepVGG, ResNet-SelfSup and SNGAN)
- Efficiency: we consider direct model inversion which takes 2. feature activation as input and directly generates the recovered image.
- 3. No Real Data: the adversary does NOT have access to any 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)

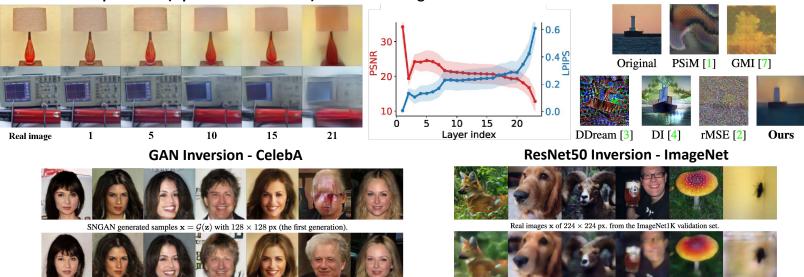
• DCI partitions the overall inversion problem into several block-wise inversion sub-problems before integrating them together for refinement,

$$\begin{split} \mathcal{L}_{\text{layer}} &= \left\| \mathbf{u}_{k} - \mathcal{F}_{k}^{-1} \left(\mathcal{F}_{k} \left(\mathbf{u}_{k} \right) \right) \right\|_{1}, \quad \mathbf{u}_{k} = \mathcal{F}_{1:(k-1)}(\mathbf{x}) \\ \mathcal{L}_{\text{img}}(\mathbf{x}, \mathbf{x}') &= \| \mathbf{x} - \mathbf{x}' \|_{1}, \quad \mathbf{x}' = \mathcal{F}_{1:k}^{-1} \left(\mathcal{F}_{1:k} \left(\mathbf{x} \right) \right) \end{split}$$



Results

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









Recovered images $\mathbf{x}' = \mathcal{G}(\mathcal{G}^{-1}(\mathcal{G}(\mathbf{z})))$ (the second generation)



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

Comparison on RepVGG-A0

Recovered images from ResNet-50 (standard) feature embeddings after 42 convolution layers $\mathcal{F}_{1,d_2}^{-1}(\mathcal{F}_{1,42}(\mathbf{x}))$