

# On the Vulnerability of Skip Connections to Model Inversion Attacks

Koh Jun Hao\*, Sy-Tuyen Ho\*, Ngoc-Bao Nguyen, and Ngai-Man Cheung

Singapore University of Technology and Design (SUTD)  
 {junhao\_koh,ngaiman\_cheung}@sutd.edu.sg

**Abstract.** Skip connections are fundamental architecture designs for modern deep neural networks (DNNs) such as CNNs and ViTs. While they help improve model performance significantly, we identify a vulnerability associated with skip connections to Model Inversion (MI) attacks, a type of privacy attack that aims to reconstruct private training data through abusive exploitation of a model. In this paper, as a pioneer work to understand how DNN architectures affect MI, we study the impact of skip connections on MI. **We make the following discoveries:** 1) Skip connections reinforce MI attacks and compromise data privacy. 2) Skip connections in the *last stage* are the most critical to attack. 3) RepVGG, an approach to remove skip connections in the inference-time architectures, could not mitigate the vulnerability to MI attacks. 4) Based on our findings, we propose *MI-resilient architecture designs* for the first time. Without bells and whistles, we show in extensive experiments that our MI-resilient architectures can outperform state-of-the-art (SOTA) defense methods in MI robustness. Furthermore, our MI-resilient architectures are complementary to existing MI defense methods. **Our project is available at <https://Pillowkoh.github.io/projects/RoLSS/>**

**Keywords:** Model Inversion · Skip Connection · Model Inversion Resilient Architecture

## 1 Introduction

As deep neural networks (DNNs) see growing deployment across various applications like face recognition [8, 11, 13, 20, 25, 30, 34, 41, 52] and healthcare [11, 13, 33, 35, 37, 52], concerns about the privacy implications of DNNs are on the rise. Many DNNs are trained on private and sensitive datasets. There is an increasing concern of potential leakage of information of these private training samples through malicious exploitation of the model.

One particular privacy threat that has garnered growing attention is **Model Inversion (MI) attacks**. In MI attacks, adversaries seek to reconstruct private training samples by exploiting vulnerabilities in the model. For instance, an adversary with access to a face recognition model may abuse it to reconstruct the

---

\* Co-first authors

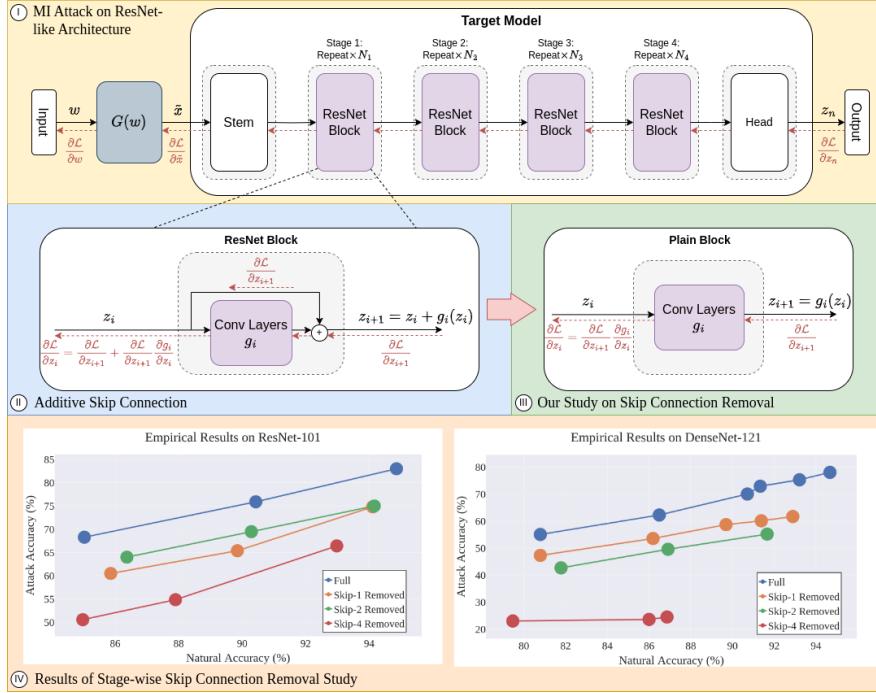
private facial images of individuals from the model’s training dataset. Following previous works [9, 21, 39, 57], we focus on reconstruction of images and use the face recognition as a running example.

**Research gap.** Recently, there is an increasing interest to study MI and to understand the feasibility and extent of reconstructing private training samples from DNNs, from the MI attack and MI defense perspectives. *However, previous studies have overlooked DNN architecture, and there is a lack of study to understand how DNN architecture designs affect MI* (Tab. 1). In particular, MI has been formulated as an optimization problem to seek an image similar to that of an identity in the private training dataset. Commonly, the MI optimization objective is formulated as maximization of likelihood under the model being attacked (target model). Several improved MI objectives have been proposed recently, e.g. logit maximization [39, 56]. Meanwhile, various regularizations on the MI objective have been studied to improve the effectiveness of MI, e.g. prior loss to penalize unrealistic images [57]. In addition, various distributional priors leveraging generative models trained on public datasets have been proposed to guide the inversion (optimization) process during MI attacks [3, 9, 48, 56, 57]. Furthermore, regularizations on the training objective of the target model to reduce the correlation exploited by MI have been studied as methods to defend against MI attacks [40, 49]. However, there is a lack of study to understand the effect of DNN architecture design on MI.

**Table 1:** There is a lack of study to understand how DNN architecture designs affect MI. Previous MI studies are DNN architecture-agnostic, focusing on MI objective, effect of regularizing MI objective, effect of distributional prior based on generative modelling, and regularization on the training objective of the target model. Our work is a pioneer study to understand how DNN architectures affect MI attacks.

	MI objective	Effect of regularizing MI objective	Effect of distributional prior to guide MI	Regularization on the training objective of target model	Effect of DNN architecture design on MI
MI [15]	✓				
GMI [57]		✓	✓		
KEDMI [9]		✓	✓		
VMI [48]			✓		
MIRROR [3]			✓		
PPA [43]	✓				
LOMMA [39]	✓				
PLGMI [56]	✓		✓		
RLBMI [21]	✓				
BREPMI [27]	✓				
MID [49]				✓	
BIDO [40]				✓	
Ours					✓

**In this paper**, we address this research gap and conduct the first study to understand how DNN architecture designs affect MI. We put our focus on skip connections [23, 24], a fundamental network design that facilitates the training of very deep neural networks. Skip connections mitigate the vanishing gradient problem during the training stage [23]. Meanwhile, many state-of-the-art (SOTA) MI attacks [9, 39, 43, 56, 57] require the use of gradients to guide the reconstruction of private training samples during the inversion stage. *We hypothesize that skip connections facilitate flowing of gradient during inversion, reinforcing MI attacks and posing a considerable vulnerability to data privacy in DNNs* (Fig. 11). To validate our hypothesis, we carefully design experiments to single out the effect of skip connections on MI attack performance. Our extensive experiments on SOTA networks (ResNet [23], DenseNet [24], MaxViT [47], EfficientNet [44]) against SOTA MI attacks (PPA [43], PLG-MI [56], LOMMA [39], KEDMI [9])

**Fig. 1: (I) Illustration of MI attack on ResNet-like architecture (Sec. 3.1).**

This figure depicts the MI attack framework for SOTA white-box MI attacks [3, 9, 39, 43, 48, 56, 57], which leverage a generative model  $G(\cdot)$  to exploit the target model via gradient descent and backpropagation. Specifically, for each iteration,  $\tilde{x} = G(w)$  is fed into the target model in the forward pass, and MI loss  $\mathcal{L}$  is computed. In the backward pass, gradients of  $\mathcal{L}$  are computed and back-propagated to obtain  $\partial\mathcal{L}/\partial w$ , which is used to update  $w$  to achieve reconstruction of private training data. **(II) Additive Skip Connection (Sec. 3.1).** During MI attacks, skip connections allow gradients to bypass the residue module and enhance backpropagation. We hypothesize that this reinforces MI attacks. **(III) Our study on skip connection removal (Sec. 3.2 and Sec. 3.3).** To validate our hypothesis that skip connections could reinforce MI, we study the effect of skip connections on MI by removing skip connections within various stages of the target model. We study both additive and concatenative skip connections. **(IV) Results of stage-wise skip connection removal study (Sec. 3.2 and Sec. 3.3).** The sub-figures show that skip connections have a considerable effect on MI. For both additive and concatenative skip connections, we observe that removal of skip connections result in considerable degradation of MI attack accuracy. Furthermore, we observe that *skip connections in the last stage have the most significant effect on MI*. **Best viewed in color with zooming in.**

consistently show that skip connections reinforce MI attacks in DNNs. Furthermore, we find that *skip connections in the last stage have the most significant effect to MI attacks*.

To mitigate the MI vulnerability caused by skip connection, we analyze, RepVGG [10], an established reparameterization technique which converts a multi-branch training-time architecture with skip connections to a plain, VGG-

like inference-time architecture with skip connection removed. However, *our analysis shows that, while RepVGG could enable an inference-time network without any skip connection, the gradients during MI attacks on this inference-time network are the same as that on the multi-branch training-time architecture with skip connections*. Therefore, RepVGG could not mitigate the vulnerability.

To bridge the existing gap, we propose *MI-resilient architecture designs* based on our own findings. Specifically, as we find that the last stage’s skip connections have the most significant effect to MI, we propose Removal of Last Stage Skip-Connection (RoLSS) as MI-resilient architecture designs. As our designs remove only the last stage’s skip connections and keep other stages’ skip connections intact, we could keep the impact on natural accuracy small in many cases. Building on top of RoLSS, we propose Skip-Connection Scaling Factor (SSF) and Two-Stage Training Scheme (TTS) to recover the model’s natural accuracy while maintaining competitive MI robustness. Our contributions are:

- We conduct a pioneer study to understand how skip connections impact MI attacks and MI robustness. We design experiments to carefully single out the effect of skip connections on MI attack performance, accounting for the effect of natural accuracy in our analysis. Through extensive experiments spanning 4 SOTA MI attacks and 10 architectures, we validate that skip connections reinforce MI attacks and pose a considerable vulnerability to data privacy in DNNs (Sec. 3).
- Notably, we discover that *skip connections in the last stage consistently are the most critical to MI attacks* (Sec. 3 and Sec. 3.3).
- We analyze RepVGG, a well-established reparameterization technique that remove skip connections by decoupling training-time and inference-time architectures. Our analysis reveals that this approach could not mitigate vulnerability to MI attacks (Sec. 9.1).
- Based on our findings, we propose MI resilient architecture designs for the first time, including: Removal of Last Stage Skip-Connection (RoLSS), Skip-Connection Scaling Factor (SSF) and Two-Stage Training Scheme (TTS). Extensive experiments show that our MI-resilient architectures can outperform SOTA defense methods in MI robustness (Sec. 5).

## 2 Related Work

**Model Inversion.** The concept of MI was initially studied by Fredrikson et al. [15], who demonstrated that adversaries could employ machine learning to extract genomic and demographic information about patients from a medical imaging model. This work was later extended to facial recognition in [14]. Since then, several MI studies have been conducted to understand the feasibility and extent of reconstructing private training samples from DNNs [21, 27, 39, 40, 43, 48, 49, 54, 56, 57]. These studies encompass both MI attacks and MI defense perspectives. We summarize notable developments in Tab. 1. See Supp. for further discussion of related work. *Despite considerable progress in MI research, there is a lack of study to understand the effect of DNN architecture design on MI.*

**Skip connections and DNNs Attacks.** Skip connections are recognized as an effective approach to alleviate the vanishing gradient problem, allowing us to train very deep neural networks [23, 24]. There are a few works that study the effect of skip connections to adversarial attacks [7, 50] and backdoor attacks [53]. **Differ from existing works, our study is the first to understand how DNN architectures affect model inversion (MI), a growing privacy attack.** Our investigation reveals a distinctive aspect: while previous work observed that skip connections aids adversarial robustness [7], our work instead discovers that skip connections reinforce MI attacks. It is important to emphasize that the nature of MI attacks differs significantly from adversarial or backdoor attacks. Particularly, for adversarial attacks, the goal is to deceive the model into making incorrect predictions. For backdoor attacks, the goal is to implant malicious functionality in the model such that the model produces incorrect outputs when a specific attack trigger is present in the input. Importantly, adversarial attacks/ backdoor attacks are not privacy attacks and they do not extract sensitive training data information from ML models. Rather, adversarial attacks/ backdoor attacks aim to undermine model utility and robustness. *Our work is the first to study the implications of skip connection on data privacy of ML models through the MI attacks.*

### 3 An Investigation on the Skip Connection Vulnerability to Model Inversion Attacks

#### 3.1 Skip connections and MI attacks

We first discuss the potential effect of skip connections on MI in this sub-section. Then, the effect is validated in Sec. 3.2 and Sec. 3.3.

**MI and gradients.** MI attacks are a data privacy threat. For a DNN model  $T$  trained on a private training dataset  $\mathcal{D}_{priv}$ , the adversary tries to exploit sensitive training data  $\mathcal{D}_{priv}$  via the trained model  $T$ . In most works, MI is formulated as the reconstruction of an input  $\tilde{x}$  which is most likely classified into an identity  $y$  by  $T$ . The model  $T$  subject to MI attacks is called *target model*. We focus on *white-box* MI attack, which is the most popular and powerful MI attack [3, 9, 39, 43, 56, 57]. Specifically, we follow previous works and assume attackers have access to the parameters, architectures, and outputs of the models [3, 9, 39, 43, 56, 57].

To reconstruct a high-dimensional image  $\tilde{x}$ , some distributional priors have been proposed in SOTA MI to constrain the search space [9, 57]. The distributional prior is commonly encoded by a generative model  $G(w)$  trained on a public dataset  $\mathcal{D}_{pub}$  which has no class intersection with  $\mathcal{D}_{priv}$ . MI attacks are commonly formulated as the following optimization:

$$w^* = \arg \min_w \mathcal{L}(w; y, T) \quad (1)$$

Here,  $\mathcal{L}(w; y, T)$  is the MI loss which guides reconstruction of  $\tilde{x} = G(w)$  that is most likely to be classified by model  $T$  as identity  $y$ . Commonly, negative log-likelihood is used:  $\mathcal{L}(w; y, T) = -\log \mathbb{P}_T(y|G(w))$ , while other losses have been

proposed, e.g., logit-based [39]. In addition, other regularization can be included in  $\mathcal{L}$ , e.g. prior loss [57].

Importantly, to solve the optimization in Eq. 1, gradient descent and back propagation are used by most SOTA white-box MI attacks [3, 9, 39, 43, 56, 57]: For each iteration,  $G(w)$  is fed into  $T$  in the forward pass, and  $\mathcal{L}$  is computed. In the backward pass, gradients of  $\mathcal{L}$  are computed in  $T$  and back-propagated to obtain  $\partial\mathcal{L}/\partial w$ , which is used to update  $w$  by the attackers.

**Skip connections could reinforce MI.** Following the above discussion, backpropagation of gradients during MI inversion could have a considerable effect on the MI attack performance. Meanwhile, for conventional DNN training, skip connections are a fundamental architecture design that is effective in mitigating gradient vanishing during backpropagation. We hypothesize that skip connections could facilitate gradient backpropagation during MI attacks and reinforce MI, thereby compromising data privacy.

Specifically, in a ResNet-like architecture, there are multiple ResNet blocks. Each ResNet block, with input  $z_i$  and output  $z_{i+1}$ , can be represented as:  $z_{i+1} = z_i + g_i(z_i)$ , including an additive skip connection and a residual module  $g_i$  comprising multiple convolutions (Fig. 11). During MI inversion, the gradient backpropagates across a ResNet block as follows:

$$\frac{\partial \mathcal{L}}{\partial z_i} = \frac{\partial \mathcal{L}}{\partial z_{i+1}} \frac{\partial z_{i+1}}{\partial z_i} = \frac{\partial \mathcal{L}}{\partial z_{i+1}} \left( 1 + \frac{\partial g_i}{\partial z_i} \right) = \frac{\partial \mathcal{L}}{\partial z_{i+1}} + \frac{\partial \mathcal{L}}{\partial z_{i+1}} \frac{\partial g_i}{\partial z_i} \quad (2)$$

Importantly, the first gradient component,  $\frac{\partial \mathcal{L}}{\partial z_{i+1}}$ , enabled by the skip connection, enhances backpropagation. We hypothesize that this reinforces MI attacks.

### 3.2 Stage-wise skip connection removal study

In this section, we validate effect of skip connections on MI.

**MI setup.** We conduct our analysis on ResNet-101 [23] and DenseNet-121 [24] as target models under the attack setup of SOTA MI attack method PPA [43]. We strictly follow PPA MI setups, where we use FaceScrub [38] as private dataset,  $\mathcal{D}_{priv}$  and attack all IDs as per PPA setup. Following previous SOTA MI works [3, 9, 27, 39, 43, 48, 56, 57], we adopt attack accuracy (AttAcc), measured using an evaluation model, as the primary metric for assessing MI performance. Attack accuracy is defined as the percentage of reconstructed images correctly identified by the evaluation model with respect to the target ID. Specific MI attack configuration can be found in the Supp.

We conduct our study by removing skip connections from various stages of the architecture. Each time, skip connections from a specific stage are removed, while those in the remaining stages remain unchanged. Each architecture with removed skip connections is trained using  $\mathcal{D}_{priv}$  in exactly the same settings as the original unaltered architecture. In this study, we focus on two common skip connection mechanisms: additive and concatenative.

**Additive skip connection removal.** We investigate additive skip connections within individual stages of the ResNet-101 architecture. To remove additive skip

**Table 2:** We strictly follow SOTA PPA [43] for the attack setup and evaluation. Here  $\mathcal{D}_{priv}$  = Facescrub [38],  $\mathcal{D}_{pub}$  = FFHQ [28]. Across architectures both additive and concatenative skip connections, we consistently observe that **Skip connections in the last stage are the most critical to MI attacks**, resulting in the most degradation in MI attack accuracy.  $\Delta_{AttAcc}$  represents the reduction in attack accuracy when compared to “Full” setting.

Architecture	Skip Connections	AttAcc	$\Delta_{AttAcc}$	Architecture	Skip Connections	AttAcc	$\Delta_{AttAcc}$
ResNet-34	Full	90.78	-	DenseNet-121	Full	88.11	-
	Skip-1 Removed	83.25	7.53		Skip-1 Removed	61.72	26.39
	Skip-2 Removed	77.92	<b>12.86</b>		Skip-2 Removed	55.21	32.90
	Skip-4 Removed	80.61	10.17		Skip-4 Removed	24.48	<b>63.63</b>
ResNet-50	Full	82.76	-	DenseNet-161	Full	74.86	-
	Skip-1 Removed	73.23	9.53		Skip-1 Removed	55.38	19.48
	Skip-2 Removed	78.77	3.99		Skip-2 Removed	59.25	15.61
	Skip-4 Removed	68.44	<b>14.32</b>		Skip-4 Removed	20.71	<b>54.15</b>
ResNet-101	Full	83.00	-	DenseNet-169	Full	77.15	-
	Skip-1 Removed	74.81	8.19		Skip-1 Removed	60.99	16.16
	Skip-2 Removed	78.75	4.25		Skip-2 Removed	51.86	25.29
	Skip-4 Removed	58.68	<b>24.32</b>		Skip-4 Removed	6.77	<b>70.38</b>
ResNet-152	Full	86.51	-	DenseNet-201	Full	77.62	-
	Skip-1 Removed	80.35	6.16		Skip-1 Removed	57.41	20.21
	Skip-2 Removed	69.04	17.47		Skip-2 Removed	46.65	31.97
	Skip-4 Removed	68.44	<b>18.07</b>		Skip-4 Removed	20.71	<b>56.91</b>

connections in ResNet-like architectures, we ensure that outputs from the previous layers are not added to the subsequent layers during the feedforward process, as shown in Fig. 11-III. For ResNet-101, the resulting ResNet block encompasses a convolutional layer,  $g_i$ , that consists of a single  $1 \times 1$  convolution, followed by a  $3 \times 3$  convolution, and lastly, another  $1 \times 1$  convolution layer, without a shortcut connection linking the input and output of the ResNet block.

**Concatenative skip connection removal.** We investigate concatenative skip connections within individual stages of the DenseNet-121 architecture. We remove the concatenative skip connections in a similar manner as the removal of additive skip connections (The details can be found in the Supp). DenseNet architectures contain DenseBlocks where input features are concatenated with the output features, before being fed into the next DenseBlock. When these features are merged through concatenation, each layer has direct access to the gradients from the loss function and the original input image. To remove these concatenative skip connections, we remove the process of concatenation and only pass the output feature of the current DenseBlock to the subsequent DenseBlock.

**Skip connections reinforce MI.** To benchmark our experiments, we utilize the original unaltered architecture to assess the performance of the modified architectures. For a fair comparison, we consider the strong correlation between natural accuracy and attack accuracy [57]. Thus, we compare these architectures at multiple checkpoints that achieve similar natural accuracy. When presenting our results, we denote the removal of all skip connections in the  $N^{th}$  stage as “Skip- $N$  Removed”. ResNet and DenseNet consist of **four stages**. We examine removal of skip connections from stages 1, 2, and 4. The result of removing skip connections in stage 3 are not included in our study, as this stage contains many parameters, and its removal leads to a severe reduction in model accuracy.

Our findings reveal that architectures with fewer skip connections impede MI attacks, leading to a decrease in MI attack accuracy. As depicted in Fig. 11-IV, both additive and concatenative skip connection studies consistently show that architectures labeled as “Skip-N Removed” exhibit low MI attack accuracy compared to the original architecture.

**Removal of Last Stage Skip-Connection (RoLSS) is the most critical to MI attacks.** Notably, we consistently observe that removing the skip connection in the last stage (i.e., “Skip-4 Removed”) results in the most degradation in MI attack accuracy. We attribute this to the specific position of skip connections removed within the architecture. During gradient backpropagation in MI attack, gradients in earlier stages depend on those in later stages, as illustrated in Eq. 2. When skip connections in stage 4 (last stage) are removed, the degraded gradients in stage 4 permeate throughout the earlier stages of the architecture, resulting in the most degradation in MI attack accuracy. We further validate this observation on various architectures for both additive and concatenative skip connections in Tab. 2.

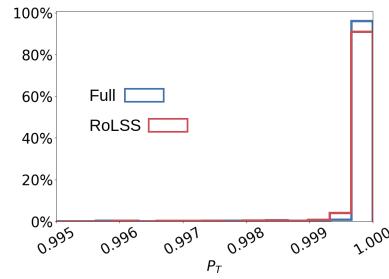
#### Removing the skip connections leads to MI optimization converging with more false positives.

As discussed in Eq. 1, MI adversaries aim to identify optimal  $w^*$  that maximizes the likelihood  $\mathbb{P}_T(y|G(w))$ . We provide this key observation to understand why removing skip connections degrades MI attacks: *When the skip connections are removed, latent variables with high likelihood  $\mathbb{P}_T(y|G(w))$  can still be identified by Eq. 1, but many  $w^*$  are false positives.* Consequently, this leads to a notable decrease in the accuracy of MI attacks.

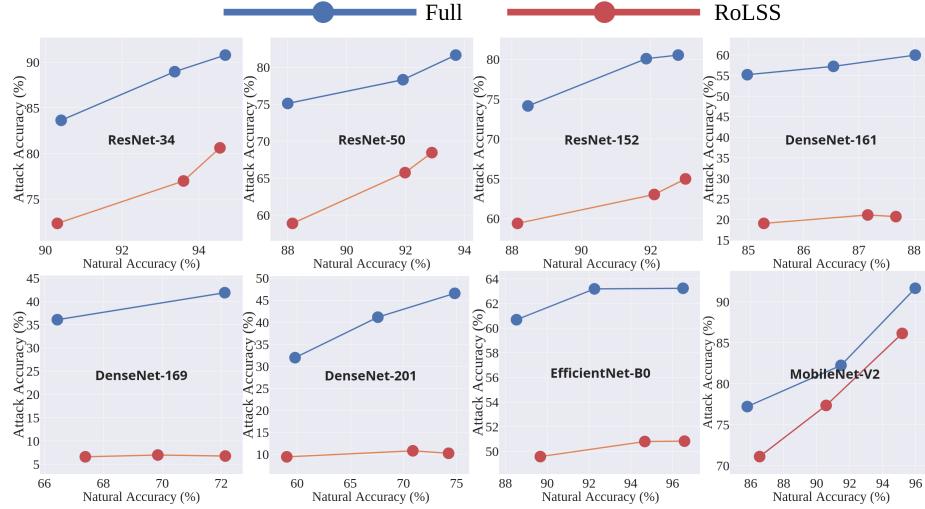
This observation becomes evident when examining the likelihood distribution of “Full” and “Skip-4 removed” settings (see Fig. 2), which are similar and both very close to 1. This results imply that, with “Skip-4 removed” setting, Eq. 1 could still perform well to seek latent variables  $w$  to maximize the likelihood  $\mathbb{P}_T(y|G(w))$ . However, despite the similarity in likelihood distributions, the attack accuracy of the “Skip-4 removed” setting is significantly lower than that of the “Full” setting (i.e., 24.32%). This suggests that, due to the absence of skip connections, the gradients in the “Skip-4 removed” setting lead MI adversaries to exploit many  $w^*$  that do not correspond to images resembling private data.

### 3.3 Extensive validation of the MI vulnerability of skip connections

We conduct extensive experiments to further validate the impact of skip connections to MI attacks. Tab. 3 summarize all MI setups in our validation ranging 10 architectures, 4 SOTA MI attacks, 3 datasets.



**Fig. 2: MI convergence analysis.** We compare histograms of likelihoods for original architecture (Full) and Removal of Last Stage Skip-Connection (RoLSS) architecture for ResNet-101 under PPA attack.



**Fig. 3:** Additional experiments to validate the impact of skip connections on MI attacks across various architecture designs, including networks with additive skip connections (ResNet-34/50/152 [23]), concatenative skip connections (DenseNet-161/169/201 [24]), and compact CNN (EfficientNet-B0 [44]). We strictly follow PPA [43] for MI setups. In all cases, a significant decrease in attack accuracy is consistently observed when skip connections are removed in the *last stage*, demonstrating that skip connections reinforce MI attacks.

**Experimental Setting.** For a fair comparison, we follow the previous MI works [3, 9, 21, 27, 39, 40, 43, 48, 56, 57] to select Evaluation Metrics, Private Dataset, Public Dataset, and Data Preparation Protocol. The details can be found in Tab. 3. Additional details are presented in the Supp.

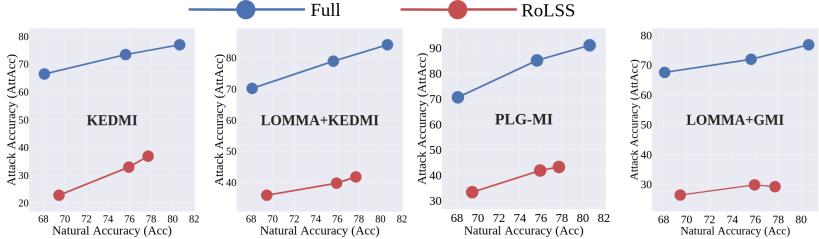
**Skip connections removal.** We apply our finding of Removal of Last Stage Skip-Connection (RoLSS) for various architectures.

**Evaluation Metrics.** Following the previous MI works, we adopt natural accuracy (Acc) and Attack Accuracy (AttAcc) as the main evaluation metrics. The detailed description and additional qualitative results are presented in the Supp.

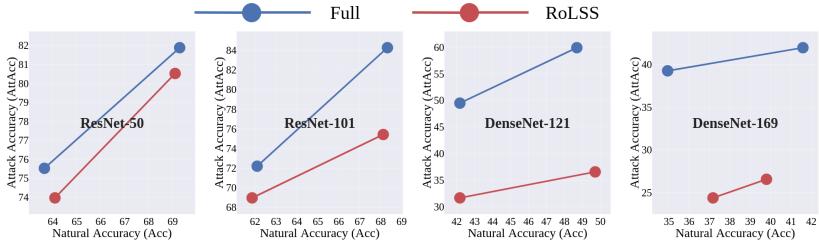
**Experimental results on various architectures.** We note that in Fig. 3, we focus on the range of natural accuracy where the two setups overlap, allowing us to observe changes in attack accuracy at the same natural accuracy level. The results are consistent with our empirical study in Sec. 3, where the RoLSS of additive skip connections (e.g., ResNet-34/50/152/EfficientNet-B0) reduce the attack accuracy by around

**Table 3: The summary of our MI setups.** We follow the exact experiment setups of PPA [43]. For the other MI attacks, we follow the setups in [9, 39] for [9, 39, 57]. In total, our study spans 10 architectures, 4 MI attacks, 3 datasets.

Architectures	MI Attack	Private Dataset
ResNet-34/50/101/152 [23]		
DenseNet-121/161/169/201 [24]		Facescrub [38]
MaxViT-T [47]	PPA [43]	
EfficientNet-B0 [44]		
ResNet-50/101 [23]	KEDMI [9]	
DenseNet-121/169 [24]	LOMMA [39]	Stanford Dogs [29]
IR152 [23]	PLG-MI [56]	CelebA [32]



**Fig. 4: Additional experiments on other SOTA MI attacks** including KEDMI [9], LOMMA [39], and PLG-MI [56]. We follow the standard setup, where  $T = \text{IR152}$ ,  $\mathcal{D}_{priv} = \text{CelebA}$ ,  $\mathcal{D}_{pub} = \text{CelebA/FFHQ}$ . Across all SOTA MI attacks, a consistent and notable reduction in attack accuracy is observed when skip connections are removed in the last stage, demonstrating that skip connections reinforce MI attacks.



**Fig. 5: Additional experiments on the Stanford Dogs** [29] dataset as  $\mathcal{D}_{priv}$ . The experiments are conducted under PPA [43] attacks across various architectures, including ResNet-50/101 and DenseNet-121/169. We strictly follow MI setups in PPA.

10% to 15% while the RoLSS of concatenative skip connections (e.g., DenseNet-161/169/201) reduce the attack accuracy by around 30% to 35%. Overall, in all cases including addictive connections (ResNet-34/50/152), concatenative skip connections (DenseNet-161/169/201), compact CNN (EfficientNet-B0), we consistently observe the significant drops in MI attack accuracy, ranging from 10% to 35%. This results further validate our findings in Sec. 3 that skip connections reinforce MI attacks. More results can be found in the Supp.

**Experimental results on other SOTA MI attacks.** Beside PPA attack [43], we validate our findings of RoLSS on other SOTA MI attacks including KEDMI [9], PLG-MI [56] and LOMMA [39]. The details for these MI attack can be found in the Tab. 3. Following previous setups, we use IR152 [23] as the architecture of the target classifier. The experimental results are presented in Fig. 4. In Fig. 4, we focus on the range of natural accuracy where the two setups overlap, allowing us to observe changes in attack accuracy at the same natural accuracy level. Across all these MI attacks, the results are consistent to those under PPA attack in Sec. 3.2, where the attack accuracy reduces significantly when the skip connections are removed regardless the effect from the natural accuracy.

**Experimental results on other private datasets.** In addition to Facescrub dataset [38] in the main study, we further validate our findings of RoLSS on other datasets, including CelebA [32] and Stanford Dogs [29]. For CelebA [32], we conduct our study on KEDMI [9]/LOMMA [39]/PLG-MI [56]. We use the

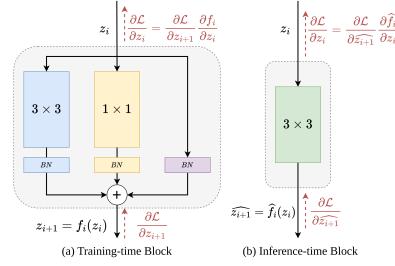
standard setup of IR152 [23] as the architecture of target classifier. For Stanford Dogs, we conduct our study on PPA [43]. We use the setup of ResNet-50/101 [23] and DenseNet-121/169 [24] as the architecture of the target classifier. The experimental results are presented in Fig. 4 and Fig. 5 for CelebA and Stanford Dogs, respectively. The results in both datasets are consistent with the results for Facescrub, where the attack accuracy experiences a significantly reduction when the skip connection is removed regardless the effect from the natural accuracy.

#### 4 Removing skip connection in inference-time architecture via RepVGG could not help mitigate vulnerability to MI

In this section, we analyze RepVGG [10], an established method to decouple the training-time and inference-time architectures through structural re-parameterization. RepVGG converts a training-time multi-branch block (with skip connection) into an inference-time plain convolutional layer (without skip connection) to accelerate inference speed, as shown in Fig. 10. As RepVGG removes skip connections in inference-time architecture, we seek to explore: *Can RepVGG inference-time architecture mitigate vulnerability to MI attacks?*

We denote the training-time multi-branch architecture by  $T_{RepVGG}$ , and the inference-time plain architecture by  $\widehat{T}_{RepVGG}$ . Through our analytical and empirical analysis, we show that the gradients in  $\widehat{T}_{RepVGG}$  under MI attacks and that in  $T_{RepVGG}$  are the same. Therefore, removing skip connection in the inference-time architecture  $T_{RepVGG}$  could not help mitigate vulnerability to MI. Our detailed analytical analysis can be found in the Supp.

To empirically validate this, we assess the vulnerability of RepVGG inference-time architecture to MI attacks, comparing it to the training-time architecture. We apply SOTA MI attack, PPA [43], on both the training-time and inference-time RepVGG-A0/B3/D2 architectures. The results for other RepVGG architectures can be found in Supp. We strictly follow the training and inference conversion implementation from the original RepVGG source code for the target classifiers trained on the Facescrub dataset [38]. For the PPA attack [43], we follow the attack setup provided in the original PPA source code. Our results in Tab. 11 clearly show that, **despite the removal of skip connections, the**



**Fig. 6:** RepVGG [10] converts a training-time multi-branch block (with skip connection) into an inference-time plain convolutional layer (without skip connection). Through our analytical and empirical analysis, we show that despite the differences in architectures, the gradients across the blocks are similar (See our analysis in the Supp.). Consequently, even with the removal of skip connections, RepVGG cannot mitigate vulnerability to MI attacks.

**inference-time architecture obtained via RepVGG remains as vulnerable to MI attacks as the training-time architecture.**

The skip connection removed architectures in Sec. 3 and those in the study of RepVGG in this section are fundamentally different. During training, RepVGG training-time architecture still receives gradients via skip connection branches. We show that the gradients in RepVGG inference-time architecture remain the same, including the skip connection branch’s gradient components even though skip connections are absent in the inference-time architecture. In contrast, our study in Sec. 3 removes skip connections during training, eliminating gradients on skip connections. The study demonstrates the causal effect of skip connections on MI attack accuracy.

## 5 Model Inversion Resilient Architecture Design

Our findings so far reveal that skip-connections reinforce MI attacks while existing reparameterization technique, RepVGG [10], to remove skip connection in the inference-time architecture cannot mitigate vulnerability to MI attack. To bridge this gap, we propose MI-resilient architecture designs for the first time, including: Removal of Last Stage Skip-Connection (RoLSS), Skip-Connection Scaling Factor (SSF), and Two-stage Training Scheme (TTS). Our RoLSS, SSF, and TTS are remarkably simple, maintaining the same training procedure as the original model. Applying RoLSS involves no additional hyperparameters, while SSF/TTS requires only one extra hyperparameter, making them easily applicable to various architectures. In contrast, the SOTA MI defense BiDO requires an extensive grid search for each architecture [40].

**Table 4:** Experimental results of RepVGG [10] training-time and inference-time architectures. We strictly follow PPA [43] for the attack setup and evaluation. Here  $\mathcal{D}_{priv}$  = Facescrub,  $\mathcal{D}_{pub}$  = FFHQ. Despite the removal of skip connections, RepVGG inference-time architecture remains as vulnerable to MI attacks as the training-time architecture.

Architecture	Acc↑	AttAcc↓
RepVGG-A0	Training-time	94.90
	Inference-time	85.19
RepVGG-B3	Training-time	94.55
	Inference-time	80.69
RepVGG-D2	Training-time	94.55
	Inference-time	80.61
RepVGG-D2	Training-time	92.05
	Inference-time	67.08
	Training-time	92.05
	Inference-time	66.44

**Table 5:** Our simple RoLSS outperforms SOTA MI defense BiDO [40]. Our further proposed SSF and TTS help recover Acc while offer competitive MI robustness.  $\Delta$  represents the ratio of attack accuracy drop to natural accuracy drop. We could not compare with unsupported BiDO architectures (i.e., DenseNet), as BiDO requires extensive hyperparameter grid search.

Architecture	Defense	Acc↑	AttAcc↓	$\Delta \uparrow$
ResNet-34	No Def.	94.69	90.78	-
	BiDO	91.66	81.98	2.90
	RoLSS (Ours)	91.38	71.86	5.72
	SSF (Ours)	94.21	79.79	22.90
	TTS (Ours)	94.40	81.65	<b>31.48</b>
ResNet-50	No Def.	94.58	82.76	-
	BiDO	91.12	58.41	7.04
	RoLSS (Ours)	92.89	68.44	<b>8.47</b>
	SSF (Ours)	93.05	74.79	6.87
	TTS (Ours)	93.56	77.21	5.44
ResNet-101	No Def.	94.86	83.00	-
	BiDO	90.31	67.07	3.50
	RoLSS (Ours)	92.40	58.68	9.89
	SSF (Ours)	93.79	71.06	<b>11.16</b>
	TTS (Ours)	94.16	77.26	8.20
ResNet-152	No Def.	95.43	86.51	-
	BiDO	91.80	58.14	7.82
	RoLSS (Ours)	93.00	64.98	8.86
	SSF (Ours)	93.79	70.71	<b>9.63</b>
	TTS (Ours)	93.97	73.59	8.85

### 5.1 Removal of Last Stage Skip-Connection (RoLSS)

In Sec. 3, our investigation reveals that removing skip connections in the last stage yields significant degradation in MI attack accuracy, suggesting a promising approach to improve MI robustness from architectural perspective. The MI defense results are presented in Tab. 6 and Tab. 5. Across architectures and skip connection mechanism, the results consistently show that **removing the skip connections in last stage (i.e., RoLSS) can improve the MI robustness**. Notably, our simple RoLSS achieves highly competitive MI robustness compared to the SOTA MI defense BiDO [40]. For instance, with ResNet-101, our RoLSS improves model accuracy by 2.09%, while the MI attack accuracy degrades by 8.49%, resulting in superior MI robustness compared to BiDO.

### 5.2 Skip-Connection Scaling Factor (SSF)

We further propose SSF on top of RoLSS to improve natural accuracy of the model while maintaining competitive MI robustness. For additive skip connections, we introduce a scale factor  $0 \leq k \leq 1$  for the signal on the skip connection of the last stage:

$$z_{i+1} = g_i(z_i) + k \cdot z_i \quad (3)$$

Details of SSF for concatenative skip connections can be found in Supp. Our SSF generalizes the skip connection, where  $k = 1$  corresponds to the original skip connection, while  $k = 0$  is similar to our skip connection removal study. With  $k < 1$ , gradients can be limited during MI attack, and this could degrade MI.

The results are presented in Tab. 6 and Tab. 5. We apply SSF over RoLSS and set  $k = 0.01$  for all concatenative skip connection architectures and  $k = 0.2$  for all additive skip connection architectures in our experimental setups. Overall, our SSF further improves MI robustness beyond RoLSS and outperforms the SOTA MI defense BiDO [40] across various architectures. Notably, for DenseNet, SSF significantly aids in recovering model performance while still mitigating MI attacks, resulting in much improved MI robustness. For example, with DenseNet-201, SSF only incurs a  $\sim 1\%$  drop in model accuracy while degrading MI attack accuracy by  $\sim 12\%$ .

**Table 6: Our simple RoLSS outperforms SOTA MI defense BiDO [40]. Our further proposed SSF and TTS help recover Acc while offer competitive MI robustness.**  $\Delta$  represents the ratio of attack accuracy drop to natural accuracy drop. We could not compare with unsupported BiDO architectures (i.e., DenseNet), as BiDO requires extensive hyperparameter grid search.

Architecture	Defense	Acc $\uparrow$	AttAcc $\downarrow$	$\Delta$ $\uparrow$
DenseNet-121	No Def.	94.67	78.09	-
	RoLSS (Ours)	86.86	24.48	6.86
	SSF (Ours)	91.73	56.32	<b>7.35</b>
DenseNet-161	No Def.	93.93	74.86	-
	RoLSS (Ours)	87.67	20.71	<b>8.65</b>
	SSF (Ours)	93.77	74.27	3.69
DenseNet-169	No Def.	94.28	77.15	-
	RoLSS (Ours)	72.14	6.77	3.18
	SSF (Ours)	92.95	60.99	<b>12.15</b>
DenseNet-201	No Def.	94.32	77.62	-
	RoLSS (Ours)	74.25	10.24	3.36
	SSF (Ours)	93.09	65.21	<b>10.09</b>

### 5.3 Two-stage Training Scheme (TTS)

We further introduce a Two-stage Training Scheme (TTS) on top of RoLSS to improve model accuracy while still maintaining competitive MI robustness. Inspired by Transfer Learning literature [55], TTS consists of two training stages:

*Stage 1:* We train model  $T$  with *full skip-connections architecture* over  $M$  epochs. This stage ensures the reasonable convergence of  $\theta_T$  with well-backpropagated gradients through the full skip connections architecture. Note that model parameters are far from optimum initially. With full skip connections in this stage, large gradients can be backpropagated in making larger parameter updates.

*Stage 2:* We remove skip connections in the last stage, i.e. RoLSS, to create *skip connection-removed architecture*  $T_p$ . Then, we continue to train  $\theta_{T_p}$  over  $N$  epochs. The pre-trained parameters in Stage 1 serves as initialization for  $\theta_{T_p}$ , thereby aiding the enhanced convergence of  $T_p$ .

We build TTS on top of RoLSS. For a fair comparison, the total training epochs for both stages (i.e.,  $M + N$ ) match the total training epochs of the original model. Across all setups, we set  $M = 5$  and  $N = 95$ . The results in Tab. 5 demonstrate that TTS outperforms the SOTA MI defense BiDO [40]. Furthermore, TTS improves model accuracy while maintaining competitive MI robustness when compared to our RoLSS. For instance, in the ResNet-34 setup, TTS achieves similar MI attack accuracy as BiDO but maintains comparable model accuracy with No Def. model, achieving very competitive MI robustness.

## 6 Conclusion

We conducted a pioneering study to examine the impact of DNN architecture on SOTA MI attacks. Our findings reveal that skip connections reinforce MI attacks, thereby jeopardizing data privacy. Through extensive MI setups, we find that the skip connections in the last stage is the most critical to MI attacks. Furthermore, our analytical and empirical analysis on RepVGG reveal that the removal of skip connections in the inference-time architecture could not help mitigate the MI vulnerability. Based on our own findings, we propose MI-resilient architecture designs for the first time, including: Removal of Last Stage Skip-Connection (RoLSS), Skip-Connection Scaling Factor (SSF), and Two-stage Training Scheme (TTS). Our MI-resilient architecture designs are remarkably simple to apply and achieve very competitive MI robustness compared to SOTA MI defense.

## 7 Acknowledgement

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); The Agency for Science, Technology and Research (A\*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This material is based on the research/work support in part by the Changi General Hospital and Singapore University of Technology and Design, under the HealthTech Innovation Fund (HTIF Award No. CGH-SUTD-2021-004).

## References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018)
3. An, S., Tao, G., Xu, Q., Liu, Y., Shen, G., Yao, Y., Xu, J., Zhang, X.: Mirror: Model inversion for deep learning network with high fidelity. In: Proceedings of the 29th Network and Distributed System Security Symposium (2022)
4. Arpit, D., Campos, V., Bengio, Y.: How to initialize your network? robust initialization for weightnorm & resnets. Advances in Neural Information Processing Systems **32** (2019)
5. Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K.W.D., McWilliams, B.: The shattered gradients problem: If resnets are the answer, then what is the question? In: International Conference on Machine Learning. pp. 342–350. PMLR (2017)
6. Beardon, A.: The klein, hilbert and poincaré metrics of a domain. Journal of computational and applied mathematics **105**(1-2), 155–162 (1999)
7. Cazenavette, G., Murdock, C., Lucey, S.: Architectural adversarial robustness: The case for deep pursuit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7150–7158 (2021)
8. Chang, X., Zhang, W., Qian, Y., Le Roux, J., Watanabe, S.: End-to-end multi-speaker speech recognition with transformer. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6134–6138. IEEE (2020)
9. Chen, S., Kahla, M., Jia, R., Qi, G.J.: Knowledge-enriched distributional model inversion attacks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16178–16187 (2021)
10. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13733–13742 (2021)
11. Dippel, J., Vogler, S., Höhne, J.: Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. arXiv preprint arXiv:2104.04323 (2021)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., McDonald, C., Piguet, C.M., et al.: Contrastive learning with continuous proxy meta-data for 3d mri classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 58–68. Springer (2021)
14. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1322–1333 (2015)

15. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: 23rd USENIX Security Symposium (USENIX Security 14). pp. 17–32 (2014)
16. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y.: Generative adversarial networks, 1–9. arXiv preprint arXiv:1406.2661 (2014)
18. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16. pp. 63–77. Springer (2005)
19. Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B., et al.: Kernel methods for measuring independence (2005)
20. Guo, J., Zhu, X., Zhao, C., Cao, D., Lei, Z., Li, S.Z.: Learning meta face recognition in unseen domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6172 (2020)
21. Han, G., Choi, J., Lee, H., Kim, J.: Reinforcement learning-based black-box model inversion attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20504–20513 (2023)
22. Hanin, B.: Which neural net architectures give rise to exploding and vanishing gradients? Advances in neural information processing systems **31** (2018)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
24. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
25. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricular-face: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
26. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
27. Kahla, M., Chen, S., Just, H.A., Jia, R.: Label-only model inversion attacks via boundary repulsion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15045–15053 (2022)
28. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
29. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO (June 2011)
30. Krishna, G., Tran, C., Yu, J., Tewfik, A.H.: Speech recognition with no speech or with noisy speech. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1090–1094. IEEE (2019)

31. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
32. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
33. Luo, L., Xu, D., Chen, H., Wong, T.T., Heng, P.A.: Pseudo bias-balanced learning for debiased chest x-ray classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 621–631. Springer (2022)
34. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14234 (2021)
35. Mishra, S., Zhang, Y., Zhang, L., Zhang, T., Hu, X.S., Chen, D.Z.: Data-driven deep supervision for skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 721–731. Springer (2022)
36. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
37. Müller, P., Kaassis, G., Zou, C., Rueckert, D.: Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 647–657. Springer (2022)
38. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP). pp. 343–347. IEEE (2014)
39. Nguyen, N.B., Chandrasegaran, K., Abdollahzadeh, M., Cheung, N.M.: Rethinking model inversion attacks against deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
40. Peng, X., Liu, F., Zhang, J., Lan, L., Ye, J., Liu, T., Han, B.: Bilateral dependency optimization: Defending against model-inversion attacks. In: KDD (2022)
41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
42. Sriramanan, G., Addepalli, S., Baburaj, A., et al.: Guided adversarial attack for evaluating and enhancing adversarial defenses. Advances in Neural Information Processing Systems **33**, 20297–20308 (2020)
43. Struppek, L., Hintersdorf, D., Correia, A.D.A., Adler, A., Kersting, K.: Plug & play attacks: Towards robust and flexible model inversion attacks. arXiv preprint arXiv:2201.12179 (2022)
44. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
45. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. pp. 10096–10106. PMLR (2021)
46. Tang, Y., Han, K., Xu, C., Xiao, A., Deng, Y., Xu, C., Wang, Y.: Augmented shortcuts for vision transformers. Advances in Neural Information Processing Systems **34**, 15316–15327 (2021)
47. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: European conference on computer vision. pp. 459–479. Springer (2022)

48. Wang, K.C., Fu, Y., Li, K., Khisti, A., Zemel, R., Makhzani, A.: Variational model inversion attacks. *Advances in Neural Information Processing Systems* **34**, 9706–9719 (2021)
49. Wang, T., Zhang, Y., Jia, R.: Improving robustness to model inversion attacks via mutual information regularization. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 11666–11673 (2021)
50. Wu, D., Wang, Y., Xia, S.T., Bailey, J., Ma, X.: Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990* (2020)
51. Xu, J., Pan, Y., Pan, X., Hoi, S., Yi, Z., Xu, Z.: Regnet: self-regulated network for image classification. *IEEE Transactions on Neural Networks and Learning Systems* (2022)
52. Yang, J., Chen, H., Yan, J., Chen, X., Yao, J.: Towards better understanding and better generalization of few-shot classification in histology images with contrastive learning (2022)
53. Yang, S., Li, Y., Jiang, Y., Xia, S.T.: Backdoor defense via suppressing model shortcuts. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2023)
54. Yang, Z., Zhang, J., Chang, E.C., Liang, Z.: Neural network inversion in adversarial setting via background knowledge alignment. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. pp. 225–240 (2019)
55. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? *Advances in neural information processing systems* **27** (2014)
56. Yuan, X., Chen, K., Zhang, J., Zhang, W., Yu, N., Zhang, Y.: Pseudo label-guided model inversion attack via conditional generative adversarial network. *Thirty Seventh AAAI Conference on Artificial Intelligence (AAAI 23)* (2023)
57. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D.: The secret revealer: Generative model-inversion attacks against deep neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 253–261 (2020)

## Overview

We provide additional results and analysis in this Supp, including:

- Our skip connection removal study on Vision Transformers (Sec. 8).
- Detailed analysis and additional empirical results for RepVGG [10] under MI study (Sec. 9).
- Additional discussion and results for our MI-resilient architectures (Sec. 10).
- User study (Sec. 11).
- Discussion on architectures without skip connection (Sec. 12)
- The detailed experimental setting for skip connection removal (Sec. 13) and MI attack (Sec. 14).
- Further discussion on related works (Sec. 15).
- The limitation (Sec. 16) and ethical consideration (Sec. 17) of our work.

## 8 Skip Connection Removal Study on Vision Transformer

Similar to the study on CNNs architectures in the main manuscript, we conduct the skip connection removal study on Vision Transformer (ViT) architectures. Specifically, we put our focus on vanilla ViT [12] and MaxViT [47]. **Our observations are consistent with those in CNNs architectures, where skip connections reinforce MI attacks.**

*MI Attacks.* To assess MI vulnerability, we employ the SOTA MI attack, PPA [43], utilizing StyleGAN [28] as the prior distribution. In PPA, the attack is performed on StyleGAN  $\mathcal{W}$  space, which is previously optimized from StyleGAN  $\mathcal{Z}$  space during MI initialization stage.

*Skip connections removal.* Due to feature collapse phenomenon in ViTs: Removing skip connections in the later stages [46] results in very poor model performance for ViTs, we are only able to remove the skip connections in the first stage.

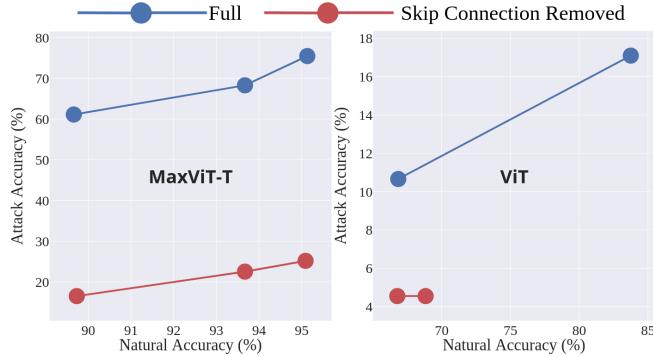
*Target classifier  $T$ .* We conduct our study on vanilla ViT [12] and SOTA Vision Transformer architecture, MaxViT [47].

*Evaluation Metrics.* Following the previous MI works, we adopt natural accuracy (Acc) and Attack Accuracy (AttAcc) as the main evaluation metrics.

*Dataset  $\mathcal{D}_{priv}$ .* As facial recognition are commonly used in real-world scenarios, following the existing MI works, we focus on the study of Facescrub [38].

*Data Preparation Protocol.* Following previous MI works, the private dataset  $\mathcal{D}_{priv}$  is exclusively used for training the target classifier  $T$ , while the public dataset  $\mathcal{D}_{pub}$  is utilized to extract prior information. There is no class intersection between  $\mathcal{D}_{priv}$  and  $\mathcal{D}_{pub}$  to ensure that the adversary only access to  $\mathcal{D}_{pub}$  to extract general features, and does not access to the information about  $\mathcal{D}_{priv}$  used for training target model.

**Experimental results.** The results in Fig. 7 are consistent with our study conducted on CNN architectures, where the skip connection reinforce MI attack. For example, we observe a significant drop in MI attack accuracy  $\sim 30\%$  in MaViT



**Fig. 7: Our Skip Connection Removal Study on Vision Transformer Architectures.** We follow the MI setup from PPA [43],  $\mathcal{D}_{priv}$  = FaceScrub,  $\mathcal{D}_{pub}$  = FFHQ. Consistent with our observations in the CNNs, we note a significant decrease in attack accuracy when skip connections are removed, indicating that skip connections reinforce MI attacks.

when skip connections are removed regardless the effect of natural accuracy. This results further validate our findings in the main manuscript.

Notably, in the MaxViT, the decrease in natural accuracy is minimal when skip connections are removed in a single stage, but the decrease in attack accuracy is significant. This leads to the improvement in MI robustness. As depicted in Fig. 8, the removal of skip connections in the MaxViT setup significantly influences the quality of reconstructed images, resulting in a better MI robustness.

	Real											Natural Acc	Attack Acc
Real												96.94%	80.78%
Full												96.94%	80.78%
Skip-1 Removed												95.09%	25.17%

**Fig. 8:** Qualitative results. Here  $T$ =MaxViT,  $\mathcal{D}_{priv}$ =FaceScrub,  $\mathcal{D}_{pub}$ =FFHQ. The selection of identities and images is entirely random, without any cherry-picking, aiming to provide an unbiased comparison. The results show that the skip connections (i.e., Full) reinforce the MI attack, resulting in the better attack accuracy and reconstructed images that exhibit more visual characteristics of the target identities.

## 9 Detailed RepVGG Study

In this section, we provide detailed analysis and additional results for the study of RepVGG [10] under MI attacks.

### 9.1 Removing skip connection in inference-time architecture via RepVGG could not help

As discussed in the main manuscript, we seek to explore: *Can RepVGG inference-time architecture (without skip connections) mitigate vulnerability to MI attacks?*

We denote the training-time multi-branch architecture by  $T_{RepVGG}$ , and the inference-time plain architecture by  $\widehat{T_{RepVGG}}$ . In what follows, we provide analytical and empirical analysis to show that **the gradients in  $T_{RepVGG}$  under MI attacks and that in  $\widehat{T_{RepVGG}}$  are the same. Therefore, removing skip connection in the inference-time architecture  $\widehat{T_{RepVGG}}$  could not help mitigate vulnerability to MI.**

Specifically, in the training-time multi-branch architecture  $T_{RepVGG}$ , there are several blocks. Each block includes three branches: a  $3 \times 3$  conv kernel, a  $1 \times 1$  conv kernel, and a skip connection. The skip connection helps mitigate gradient vanishing. We denote the  $i^{th}$   $T_{RepVGG}$  block as  $z_{i+1} = f_i(z_i)$ , which is shown in Eq. 4, where  $z_i$  represents the input of the  $i^{th}$  block,  $W^{(k)}$  represents the weight of the  $k \times k$  conv kernel (where  $k = 0$  indicates additive skip connections), and  $\mu^{(k)}, \sigma^{(k)}, \gamma^{(k)}, \beta^{(k)}$  represent the accumulated mean, standard deviation, learned scaling factor, and bias of the Batch Normalization (BN) [26] layer following the  $k \times k$  conv kernel. The  $i^{th}$   $T_{RepVGG}$  block is [10]:

$$\begin{aligned} z_{i+1} = f_i(z_i) &= BN(z_i * W^{(3)}, \mu^{(3)}, \sigma^{(3)}, \gamma^{(3)}, \beta^{(3)}) \\ &\quad + BN(z_i * W^{(1)}, \mu^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(1)}) \\ &\quad + BN(z_i, \mu^{(0)}, \sigma^{(0)}, \gamma^{(0)}, \beta^{(0)}) \end{aligned} \quad (4)$$

After the training phase, RepVGG uses reparameterization to convert a  $T_{RepVGG}$  block into a  $\widehat{T_{RepVGG}}$  block  $\widehat{z_{i+1}} = \widehat{f}_i(z_i)$ , which is a plain conv layer [10]:

$$\widehat{z_{i+1}} = \widehat{f}_i(z_i) = z_i * \widehat{W} + \widehat{b} \quad (5)$$

We show that **the outputs of a  $T_{RepVGG}$  block and a  $\widehat{T_{RepVGG}}$  block are equal despite their differences in architectures**, i.e.,  $f_i(z_i) = \widehat{f}_i(z_i)$ . Therefore, the gradients in a  $T_{RepVGG}$  block,  $\partial f_i / \partial z_i$ , are the same as that in a  $\widehat{T_{RepVGG}}$  block,  $\partial \widehat{f}_i / \partial z_i$ .

To show  $f_i(z_i) = \widehat{f}_i(z_i)$ , we note that  $\widehat{W}$  and  $\widehat{b}$  in Eq. 5 are obtained in [10] with the following reparameterization procedure: After training, the conv and BN in each branch of  $f_i(z_i)$ , with kernel  $W^{(k)}$  and BN parameters  $\{\mu^{(k)}, \sigma^{(k)}, \gamma^{(k)}, \beta^{(k)}\}$  resp., are replaced by another conv layer with parameters  $\{\widehat{W}^{(k)}, \widehat{b}^{(k)}\}$ , where [10]:

$$\begin{aligned} \widehat{W}^{(k)} &= \frac{\gamma^{(k)}}{\sigma^{(k)}} W^{(k)} \\ \widehat{b}^{(k)} &= \beta^{(k)} - \frac{\mu^{(k)} \gamma^{(k)}}{\sigma^{(k)}} \end{aligned} \quad (6)$$

We remark that  $\frac{\gamma^{(k)}}{\sigma^{(k)}} W^{(k)}$  in Eq. 6 is channel-wise multiplication, as  $\sigma^{(k)}, \gamma^{(k)}$  are BN parameters, see [10]. The same is applied to the skip connection branch, as an identity can be viewed as a  $1 \times 1$  conv with an identity matrix as the kernel. In [10], the two  $1 \times 1$  kernels are then zero-padded to  $3 \times 3$  kernels. Then, the three kernels are summed together to obtain  $\widehat{W}$  and  $\widehat{b}$  in Eq. 5 (See [10]). Therefore, Eq. 5 can be re-written as:

$$\begin{aligned}
\widehat{f}_i(z_i) &= z_i * \widehat{W} + \widehat{b} \\
&\stackrel{(a)}{=} z_i * \widehat{W^{(3)}} + \widehat{b^{(3)}} + z_i * \widehat{W^{(1)}} + \widehat{b^{(1)}} + z_i * \widehat{W^{(0)}} + \widehat{b^{(0)}} \\
&\stackrel{(b)}{=} z_i * \frac{\gamma^{(3)}}{\sigma^{(3)}} W^{(3)} + \beta^{(3)} - \frac{\mu^{(3)} \gamma^{(3)}}{\sigma^{(3)}} \\
&\quad + z_i * \frac{\gamma^{(1)}}{\sigma^{(1)}} W^{(1)} + \beta^{(1)} - \frac{\mu^{(1)} \gamma^{(1)}}{\sigma^{(1)}} \\
&\quad + z_i * \frac{\gamma^{(0)}}{\sigma^{(0)}} W^{(0)} + \beta^{(0)} - \frac{\mu^{(0)} \gamma^{(0)}}{\sigma^{(0)}} \\
&\stackrel{(c)}{=} (z_i * W^{(3)} - \mu^{(3)}) \frac{\gamma^{(3)}}{\sigma^{(3)}} + \beta^{(3)} \\
&\quad + (z_i * W^{(1)} - \mu^{(1)}) \frac{\gamma^{(1)}}{\sigma^{(1)}} + \beta^{(1)} \\
&\quad + (z_i - \mu^{(0)}) \frac{\gamma^{(0)}}{\sigma^{(0)}} + \beta^{(0)} \\
&\stackrel{(d)}{=} f_i(z_i)
\end{aligned} \tag{7}$$

In (a), we rewrite  $\widehat{W}$  as the sum of three kernels, and remove the zero-padded coefficients to obtain the two  $1 \times 1$  kernels:  $\widehat{W^{(1)}}$  and  $\widehat{W^{(0)}}$ . In (b), we use Eq. 6. In (c), we re-arrange the terms. In (d), we use the definition of BN:  $BN(z_i, \mu^{(k)}, \sigma^{(k)}, \gamma^{(k)}, \beta^{(k)}) = (z_i - \mu^{(k)}) \frac{\gamma^{(k)}}{\sigma^{(k)}} + \beta^{(k)}$  and follow Eq. 4. Overall, Eq. 7 shows that  $\widehat{f}_i(z_i) = f_i(z_i)$ . As a result, the gradients in a  $T_{RepVGG}$  block,  $\partial \widehat{f}_i / \partial z_i$  are the same as that in a  $T_{RepVGG}$  block,  $\partial f_i / \partial z_i$ . Consequently, during MI attacks, the gradients in inference-time architecture,  $\widehat{T_{RepVGG}}$ , are the same as that in training-time architecture  $T_{RepVGG}$ .

## 9.2 Additional results on other RepVGG architectures

In addition to the empirical results on RepVGG-A0/B3/D2 in the main manuscript, we provide additional empirical results on RepVGG-A1/A2/B0/B1 in Tab. 7. Overall, the results are consistent with those in the main manuscript, where the inference-time architecture with skip connections removed via RepVGG remains as vulnerable to MI attacks as the training-time architecture.

**Table 7: Additional experimental results of other RepVGG architectures** [10]. We strictly follow PPA [43] for the attack setup and evaluation. Here  $\mathcal{D}_{priv} =$  Facescrub,  $\mathcal{D}_{pub} =$  FFHQ. We report the natural accuracy (Acc), attack accuracy (AttAcc) given in % and the distance between the reconstructed features and private training features computed using Evaluation Model  $\delta_{eval}$  and FaceNet Model [41]  $\delta_{face}$ . Across all RepVGG architectures, we find that despite the removal of skip connections, RepVGG inference-time architecture remains as vulnerable to MI attacks as the training-time architecture

	Architecture	Acc	AttAcc	$\delta_{eval}$	$\delta_{face}$
RepVGG-A1	Training-time	95.34	87.85	122.37	0.7600
	Inference-time	95.34	88.04	122.36	0.7614
RepVGG-A2	Training-time	95.25	87.50	121.90	0.7697
	Inference-time	95.25	87.29	121.75	0.7693
RepVGG-B0	Training-time	95.50	90.24	120.32	0.7455
	Inference-time	95.50	90.09	120.17	0.7445
RepVGG-B1	Training-time	95.65	84.29	124.98	0.7650
	Inference-time	95.65	84.76	124.77	0.7652

## 10 Additional MI-resilient Architectures

### 10.1 Skip Connection Scaling Factor (SSF) for concatenative skip connection for DenseNets

We further propose SSF on top of RoLSS to improve natural accuracy of the model while maintaining competitive MI robustness. In the main manuscript, we discuss the SSF for additive skip connections as in ResNets, yet SSF is equally applicable to concatenative skip connections as in DenseNets. Concatenative skip connection architectures contain DenseBlocks where input features are concatenated with the output features, before being fed into the next DenseBlock. The scale factor  $0 \leq k \leq 1$  adjusts the signal on the skip connection of the last stage as shown below:

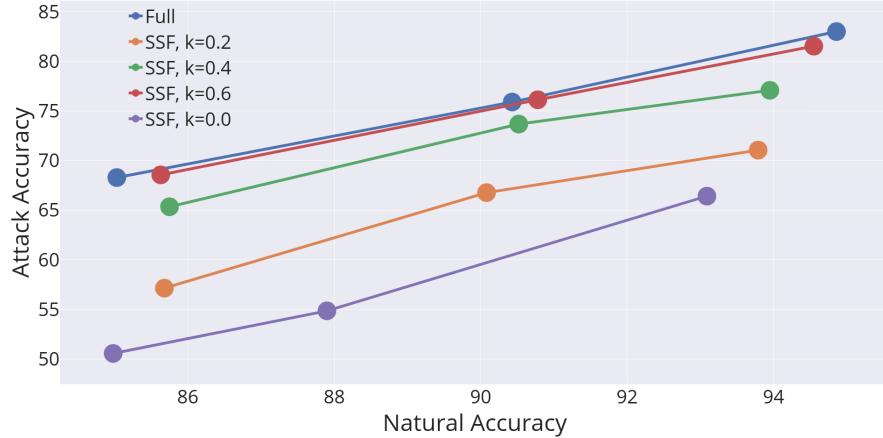
$$z_{i+1} = [z_i^{scale}, g_i(z_i)] \quad (8)$$

Here  $z_i^{scale}$  is a subset  $z_i$  including  $k \cdot n$  features from  $z_i$ , where,  $n$  is total number of features of  $z_i$ .

Similar to our discussion on SSF for additive skip connections, our SSF generalizes the skip connections, where  $k = 1$  corresponds to the original skip connection, while  $k = 0$  is out skip connection removal study. With  $k < 1$ , gradients can be limited during MI attack, and this could degrade MI.

### 10.2 MI-resilient architectures are complementary to existing MI defense

We are the first to explore MI defense from architectural perspective. Therefore, our MI-resilient architectures are complementary to existing MI defense.



**Fig. 9: Our ablation study on the effect of  $k$  on SSF.** We follow PPA MI setup, where  $T=\text{ResNet-101}$ ,  $\mathcal{D}_{priv}=\text{FaceScrub}$ ,  $\mathcal{D}_{pub}=\text{FFHQ}$ .

In this section, we combine the SOTA MI defense, BiDO, with our MI-resilient architectures to further improve MI robustness.

**MI setup.** We follow PPA [43] for the MI setup on Facescrub private dataset.

**Implementation.** When combining our MI-resilient architecture and BiDO, we strictly follow BiDO. The only difference is that we conduct BiDO on top of our RoLSS architectures.

**Experimental result.** The results in Tab. 8 show that the trade-off between utility and robustness is much improved with the incorporation of our RoLSS architecture into BiDO. Particularly, the reduction in MI attack accuracy by 25.63% compared to BiDO alone while only experiencing a marginal 1% decrease in natural accuracy. As pioneering exploration of MI robustness from an architectural perspective, our MI-resilient architecture is complementary to existing regularization-based SOTA MI denfense, such as BiDO.

**Table 8: MI-resilient architectures are complementary to existing MI defense.**  $\Delta$  represents the ratio of attack accuracy drop to natural accuracy drop.

Defense	Acc $\uparrow$	Att Acc $\downarrow$	$\Delta \uparrow$
No Def.	94.86	83.00	-
BiDO	90.31	67.07	3.50
BiDO+RoLSS	89.13	41.44	<b>7.25</b>

### 10.3 An ablation study on Skip Connection Scaling Factor (SSF)

In SSF, a scale factor  $0 \leq k \leq 1$  is employed to adjust the signal of the skip connection. We conduct an ablation study to examine the impact of  $k$  on SSF. We follow the setup of ResNet-101 under PPA attack in the main manuscript with varying  $k$ . The results in Fig. 9 show that as  $k$  increase, natural accuracy is more effectively restored. However, larger values of  $k$  also lead to a stronger reinforcement of MI attacks.

#### 10.4 MI-resilient architectures offer flexible control

**Our method can have flexible control and our defense performance can be easily improved even further.** Specifically, our proposed RoLSS focuses on removing skip connections in the last stage only, which is the most critical to MI attacks based on our discovery. This can be easily extended to the other stages to offer greater flexibility and control over privacy utility tradeoff. In Tab. 9, our results show that RoLSS+ can achieve better privacy utility trade-off than RoLSS.

#### 10.5 Additional Comparison Against Other MI Defenses

We provide additional baseline comparison with MID and DP, in Tab. 9. As shown in the main manuscript and Tab. 9, **our proposed method achieves the best tradeoff compared to previous SOTA MI defense.**

### 11 User Study

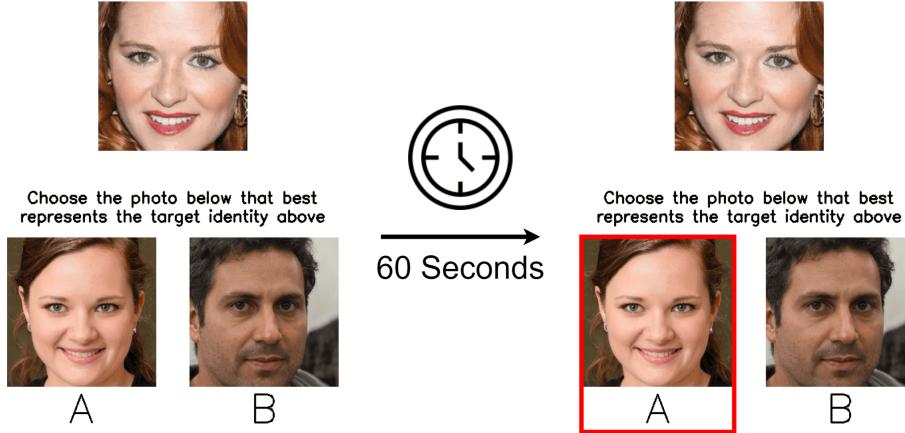
We utilize Amazon MTurk<sup>1</sup> for our user study, where participants are presented with an image of the target class and tasked with choosing the inverted image that closely resembles the target. Survey questions are randomized, and each image pair is displayed for 60 seconds. Our study covers all 530 identities in the FaceScrub Dataset, with each pair assigned to 10 unique individuals. In this user study, images are generated through the PPA attack under the MaxViT configuration (see Sec. 8 in this Supp.). Each

**Table 9:** (a) RoLSS+ builds on RoLSS by further removing 10% of skip connections from the second last stage. We show that RoLSS+ can degrade MI attack accuracy more aggressively, which demonstrates that our method offers flexibility and control over privacy utility tradeoff. (b) Our comparison with SOTA MI defenses including MID, DP, and BiDO. The results show that our methods achieve the best MI robustness tradeoff compared to existing MI defenses.

Architecture	Defense	Acc↑	AttAcc↓	Δ ↑
ResNet-34	No Def.	94.69	90.78	-
	MID [49]	91.12	46.25	12.47
	DP [1]	89.66	72.19	3.70
	BiDO [40]	91.66	81.98	2.90
	RoLSS (Ours)	91.38	71.86	5.72
	RoLSS+ (Ours)	93.49	65.78	20.83
	SSF (Ours)	94.21	79.79	22.90
ResNet-50	TTS (Ours)	94.40	81.65	<b>31.48</b>
	No Def.	94.58	82.76	-
	MID [49]	89.62	66.82	3.21
	DP [1]	89.97	68.89	3.01
	BiDO [40]	91.12	58.41	7.04
	RoLSS (Ours)	92.89	68.44	8.47
	SSF (Ours)	93.05	74.79	5.21
ResNet-101	RoLSS+ (Ours)	92.51	64.50	<b>8.82</b>
	TTS (Ours)	93.56	77.21	5.44
	No Def.	94.86	83.00	-
	MID [49]	90.85	52.61	7.58
	DP [1]	91.36	74.88	2.32
	BiDO [40]	90.31	67.07	3.50
	RoLSS (Ours)	92.40	58.68	9.89
ResNet-152	RoLSS+ (Ours)	91.05	52.74	7.94
	SSF (Ours)	93.79	71.06	<b>11.16</b>
	TTS (Ours)	94.16	77.26	8.20
	No Def.	95.43	86.51	-
	MID [49]	91.56	66.18	5.25
	DP [1]	91.61	75.33	2.93
	BiDO [40]	91.80	58.14	7.82
RoLSS (Ours)	RoLSS (Ours)	93.00	64.98	8.86
	RoLSS+ (Ours)	92.19	54.79	<b>9.79</b>
	SSF (Ours)	93.79	70.71	9.63
	TTS (Ours)	93.97	73.59	8.85

<sup>1</sup> <https://www.mturk.com>

image pair comprises one MI reconstructed from full skip connection architecture and the other from skip connection removed architecture. Results indicate that when skip connections are removed, the reconstructed images tends to be less similar to the target class, with 69.51% of users identifying images inverted by the full skip connection architecture as more similar to the target. This reinforces our hypothesis that architectures with fewer skip connections consistently reduce MI attack accuracy.



**Fig. 10: Example of user study survey interface.** The users are asked to choose one image between two options (Full and Skip Connection Removed) that best represents the target identity. For each assignment, users are given 60 seconds to complete the task.

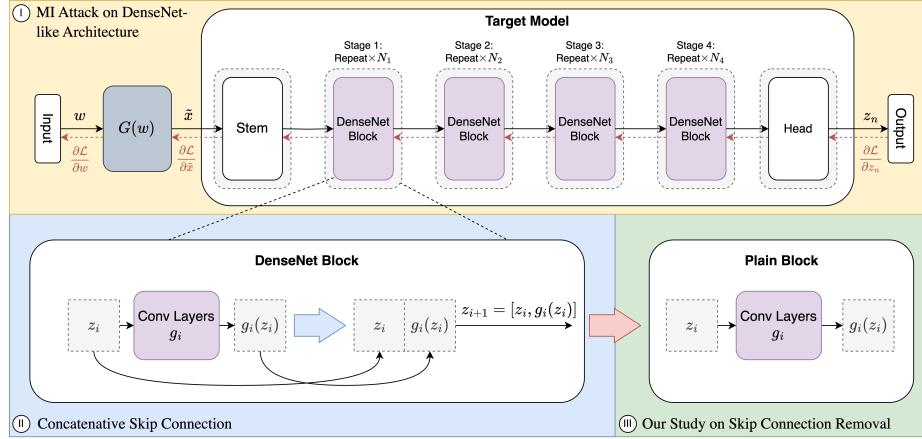
## 12 Discussion on architectures without skip connections

In this section, we discuss the model inversion to the architectures without skip connections. For a fair comparison with our study, we conduct high-resolution MI attack experiments on VGG in rebuttal Tab. 11. We observe that attack accuracy for VGG (49.39% to 55.57%) is significantly lower than for architectures with skip connections (82.76% to 90.78%, see No. Def results in rebuttal Table 9). This results are consistent with our observation that skip connections reinforce MI attack.

**Table 10: User study results.** using PPA attack on MaxViT architectures (Full and Skip Connection Removed). The user study results are consistent with Attack Accuracy, which shows that the skip connections reinforce the MI attack.

Architecture	Acc ↑	AttAcc ↓	User Preference ↓
No Def.	96.94%	80.78%	69.51%
Ours	95.09%	25.17%	30.49%

### 13 Details of Skip Connection Study Setting



**Fig. 11: (I) Illustration of MI attack on DenseNet-like architecture.** This figure depicts the MI attack framework for SOTA white-box MI attacks [3, 9, 39, 43, 48, 56, 57], which leverage a generative model  $G(\cdot)$  to exploit the target model via gradient descent and backpropagation. Specifically, for each iteration,  $\tilde{x} = G(w)$  is fed into the target model in the forward pass, and MI loss  $\mathcal{L}$  is computed. In the backward pass, gradients of  $\mathcal{L}$  are computed and back-propagated to obtain  $\partial\mathcal{L}/\partial w$ , which is used to update  $w$  to achieve reconstruction of private training data. **(II) Concatenative Skip Connection.** During MI attacks, skip connections enhance backpropagation. We hypothesize that this reinforces MI attacks. In concatenative skip connections, input signals are concatenated with the outputs of the current DenseBlock during the feed-forward process. **(III) Our study on skip connection removal.** To validate our hypothesis that skip connections could reinforce MI, we study the effect of skip connections on MI by removing skip connections within various stages of the target model. We study both additive skip connections as discussed and concatenative skip connections as shown in this sub-figure. **Best viewed in color with zooming in.**

In this section, we provide additional details to the Stage-wise Skip Connection Removal Study as mentioned in the main paper, where we specifically discuss about DenseNet-like architectures that utilizes concatenative skip connections. DenseNet architectures contain DenseBlocks where input features are concatenated with the output features, before being fed into the next DenseBlock as shown in Fig. 11-II, where  $z_{i+1} = [z_i, g_i(z_i)]$ .

**Table 11:** Experimental results on high-resolution MI attacks against VGG. We follow PPA for the attack setup and evaluation. Here,  $\mathcal{D}_{priv}$ =Facescrub and  $\mathcal{D}_{pub}$ =FFHQ

Architecture	Acc↑	AttAcc↓
VGG-16	93.70	49.39
VGG-19	93.51	55.57

### 13.1 Removal of Concatenative Skip Connections

To remove concatenative skip connections from DenseNet-like architectures, we remove the concatenation process during the feed forward process within DenseBlocks of these architectures. After removal of these concatenative skip connections, the new latent from subsequent DenseBlocks can be represented as  $z_{i+1} = g_i(z_i)$ , similar to our study when we remove additive skip connections from ResNet-like architectures. This process is illustrated in Fig. 11-III.

### 13.2 Reproducibility

**Hyper-parameters.** We strictly follow the implementations from official source codes [9, 39, 43]. The details for these hyper-parameter selection are presented in Tab. 12 for training  $T$ . For a fair comparison, we ensure that the similar training conditions for both architecture with full skip connections and architecture with skip connections removed.

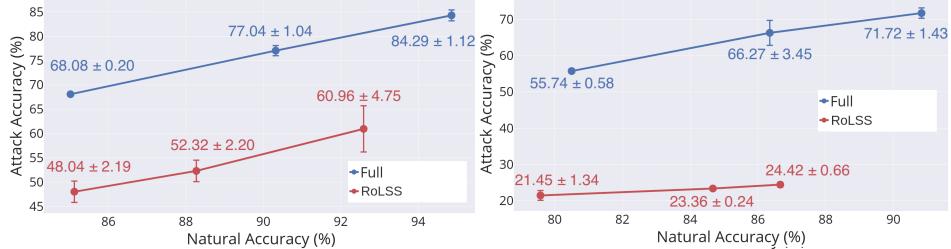
**Table 12: Hyper-parameter selection for training  $T$ .** We follow the hyper-parameter selection from previous works [9, 39, 43]. We remark that in our skip connection study, the training conditions for architecture with full skip connections and architecture with skip connections removed are similar.

Architecture	Input Size	Transformation	Optimizer	LR	LR scheduler	#Epoch	Batch Size
ResNet-34/50/101/152		RandomResizedCrop					
DenseNet-121/161/169/201		ColorJitter					
EfficientNet-B0	224 × 224	RandomHorizontalFlip	Adam	0.001	MultiStepLR	100	128
RepVGG-A0/A1/A2/B0/B1/B3/D2							
IR152	64 × 64	RandomHorizontalFlip	SGD	0.01	-	100	64

**Error bar.** To ensure the reproducibility of our findings, we repeat the main experiments reported in the original paper. As MI attacks is very time-consuming, we select key data points from the original paper and evaluate the variations in the results obtained. Specifically, we repeat the stage-wise skip connection removal on ResNet-101 and DenseNet-121 (full and skip-4 removed configurations) for 3 times and report the mean natural accuracy and attack accuracy as well as the standard deviation of the attack accuracy obtained. The results are reported in Fig. 12 and Fig. 13. For each experiment, we follow the setup as the previous works.

## 14 Details of MI Attack Setup

**Methods for Model Inversion Attacks.** Model Inversion attacks seek to generate synthetic images that capture class-wise characteristics inherent in the private dataset used for training the target classifier. Recent advancements leverage

**Fig. 12:** Error bar of ResNet-101 Setup**Fig. 13:** Error bar of DenseNet-121 Setup

generative adversarial networks (GANs) to enhance attack accuracy, surpassing traditional methodologies. This study meticulously explores generative MI attacks due to their substantial implications for data privacy, with a particular focus on understanding their interaction with skip connections across five different MI attacks.

*KEDMI* [9] utilizes a MI-specific GAN tailored for MI attacks, integrating knowledge from the target classifier during GAN training. Introducing a new head, the discriminator assumes a dual role by not only discerning between real and fake samples but also predicting the class-wise label of the input. Additionally, the authors advocate for latent distribution modeling to streamline inversion time and enhance the quality of generated samples.

*LOMMA* [39] improve prior MI attacks by introducing new logit loss for MI loss and model augmentation concept to avoid MI overfitting.

*PLG-MI* [56] employs conditional GANs for MI attacks, effectively segregating the search space for various image classes. Furthermore, the authors incorporate Max-Margin Loss to optimize MI, addressing the vanishing gradient problem inherent in widely used cross-entropy.

*PPA* [43] concentrates on MI attacks tailored for high-resolution images, employing StyleGAN for the inversion task. The proposed SOTA framework highlights its modular nature, allowing for minimal adjustments to the attack setup across diverse architectures and datasets.

**Metrics for MI Attack Evaluation.** In alignment with prior research works [9, 39, 43, 56], we utilize Attack Accuracy (AttAcc), K-Nearest-Neighbors Distance (KNN Dist), and distance metrics  $\delta_{EvalNet}$  and  $\delta_{FaceNet}$  (introduced in PPA) to assess the effectiveness of MI attacks.

**Attack Accuracy (AttAcc):** We utilize a pre-trained evaluation classifier to predict the identities of inverted images, with attack accuracy serving as a metric to gauge the similarity between these inverted images and the target images. To ensure reliability, we employ existing evaluation models from prior studies known for achieving high accuracy scores.

**K-Nearest Neighbors Distance (KNN Dist):** Quantifies the shortest feature distance between the inverted image and the target image, utilizing  $l_2$  distance within the penultimate layer of the evaluation model as a measure of feature

distance. Consequently, KNN Dist acts as a metric to assess the feature similarity between the reconstructed images and the actual images belonging to the same class.

*$\delta$  Distance:* This metric, introduced in PPA attack, quantifies the similarity between reconstructed images and private training images. It is determined by the  $l_2$  distance, measuring the difference in activations between the penultimate layers. Variations of this metric arise based on the model employed to extract these penultimate layer activations. Specifically,  $\delta_{EvalNet}$  is calculated using the Evaluation Model, whereas  $\delta_{FaceNet}$  is computed utilizing the pre-trained FaceNet [41].

## 15 Related Work

**Model Inversion.** The concept of MI was initially studied by Fredrikson et al. [15], who demonstrated that adversaries could employ machine learning to extract genomic and demographic information about patients from a medical imaging model. This work was later extended to facial recognition in [14]. An adversarial model inversion approach was introduced by Yang et al. in [54]. This approach utilizes the target classifier as an encoder to generate a prediction vector, which is used as input to a second network for reconstructing the original data.

Since then, several MI studies have been conducted to understand the feasibility and extent of reconstructing private training samples from DNNs. These studies encompass both MI attacks and MI defense perspectives. Firstly, recent works analyzed the limitations of conventional MI objectives and proposed enhancements to MI attacks, where PLGMI [56], LOMMA [39], PPA [43] utilize logit maximization loss, Max-Margin loss [42, 52] or Point Care loss [6]. Other works modified the MI objective to facilitate MI attacks in black-box [21] and label-only [27] scenarios. Secondly, regularization techniques in MI were explored to improve the realism of reconstructed images [57]. Thirdly, advanced MI attacks for high-dimensional data, such as images, examined the effect of distributional priors in guiding MI. Specifically, GMI [57] used a pretrained GAN [17] to learn the image structure of an auxiliary dataset with a similar structure to the target image space. Inversion images are then found through the latent vector of the generator. VMI [48] offers a probabilistic interpretation of MI, which leads to a variational objective for the attack. KEDMI [9] proposed to use a MI specific GAN trained on knowledge from the target model. PLGMI [56] proposed to use conditional GAN [36] to decouple the search space for different classes of images. For high-resolution MI attacks, MIRROR [3] and PPA [43] leverage the power StyleGAN [28] and perform MI on  $\mathcal{W}$  space. Finally, regularizations on the training objective of the target model as methods to defend against MI attacks have been studied in [40, 49]. More concretely, MID [49] limits the input-output dependency through a mutual information penalization, while BiDO [40] aims to minimize the dependency (via COCO [19] or HSIC [18] measurements) between latent representations and inputs while maximizing the dependency be-

tween latent representations and outputs. *Despite considerable progress in MI research, there is a lack of study to understand the effect of DNN architecture design on MI.*

**Skip connections.** A notorious problem of training very deep networks is that gradients could vanish when they reach initial layers of the networks [5, 16, 22]. Various efforts have been employed to address this issue, including the utilization of Rectified Linear Units (ReLU) [2], the implementation of Batch Normalization [26], and the application of specialized weight initialization methods [4]. From the DNNs architecture perspective, adding shortcut connections has been recognized as an effective approach to alleviate the vanishing gradient problem.

The implementation of skip connections in DNNs commonly adopts additive skip connections [23], where the output of a previous layer is added to the output of the current layer. This implementation is known for its simplicity and effectiveness. Another prevalent implementation is the concatenative skip connection [24], wherein each layer receives concatenated feature maps from all preceding layers. Through the concatenation operation along the channel dimension, this method ensures a more comprehensive set of features for subsequent layers to process. Noteworthy advanced deep neural networks, such as DenseNet [24], ResNet [23], MaxViT [47], EfficientNet [44, 45], and others [10, 12, 31, 51], leverage skip connections during training to improve their performance.

## 16 Limitation

In our experiments, we employed network architectures commonly used in MI research. Furthermore, we included a very recent network architecture, namely MaxViT. We observed consistent results across the range of network architectures we employed. Meanwhile, the study of additional network architectures may be included.

## 17 Ethical Consideration

Our study highlights the vulnerability of skip connections to privacy threats. We hope that our findings raise awareness of potential private data leaks associated with high-performance network architectures. We urge further research into privacy-safe network architectures.