Collaborative machine learning settings such as federated learning can be susceptible to adversarial interference and attacks. One class of such attacks is termed model inversion attacks, characterised by the adversary reverse-engineering the model into disclosing the training data. Previous implementations of this attack typically only rely on the shared data representations, ignoring the adversarial priors, or require that specific layers are present in the target model, reducing the potential attack surface. In this work, we propose a novel context-agnostic model inversion framework that builds on the foundations of gradient-based inversion attacks, but additionally exploits the features and the style of the data controlled by an in-the-network adversary. Our technique outperforms existing gradient-based approaches both qualitatively and quantitatively across all training settings, showing particular effectiveness against the collaborative medical imaging tasks. Finally, we demonstrate that our method achieves significant success on two downstream tasks: sensitive feature inference and facial recognition spoofing.

CCS Concepts:
- Security and privacy → Privacy protections; Privacy-preserving protocols;

Additional Key Words and Phrases: Federated learning, model inversion, collaborative image analysis

ACM Reference format:
Dmitrii Usynin, Daniel Rueckert, and Georgios Kaissis. 2023. Beyond Gradients: Exploiting Adversarial Priors in Model Inversion Attacks. *ACM Trans. Priv. Sec.* 26, 3, Article 38 (June 2023), 30 pages. https://doi.org/10.1145/3592800

1 INTRODUCTION

Machine learning (ML) models have been deployed in contexts ranging from medical image analysis [5] to stock market prediction [30]. The effective training of such models depends upon the availability of large quantities of representative training data. One paradigm that permits models to leverage larger and more diverse datasets is collaborative machine learning (CML) [39], which permits model training on geographically distributed datasets. CML includes a number of approaches ranging from direct data sharing to transfer learning on publicly available data, all of which allow institutions to share their data with other contributors and thus to train models which generalise better. However, as such data is often sensitive in nature, procuring these datasets directly can be problematic due to data protection and governance regulations, which specifically forbid collaborators from exchanging data with each other. This can be particularly problematic in

Authors’ address: D. Usynin, D. Rueckert, and G. Kaissis, TranslaTUM, Einsteinstraße 25, 81675 München; emails: du216@ic.ac.uk, d.rueckert@tum.de, g.kaissis@tum.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2471-2566/2023/06-ART38 $15.00
https://doi.org/10.1145/3592800

ACM Transactions on Privacy and Security, Vol. 26, No. 3, Article 38. Publication date: June 2023.
ML contexts that rely on data that is difficult to obtain (e.g., medical image analysis). Thus, newer paradigms of CML such as federated learning (FL) [21] have been introduced and enable distributed model training without exchanging the data itself. Instead, ML models are trained locally and only the model updates are shared with the rest of the federation of data owners. This approach, however, was shown to be exploitable by adversaries [25, 28, 38, 40], particularly those that can obtain access to shared model updates, as these contain information about the data used to train the model. Such information can be reverse-engineered, allowing the adversary to recover the original data behind the captured update, thus disclosing sensitive information. The most prominent class of such attacks is termed reconstruction attacks (or model inversion attacks) [11], which can exploit intermediate model updates in forms of shared activations or gradients. The former was previously presented in He et al. [16], where authors showed that collaborative inference is vulnerable to an honest-but-curious (HbC) adversary in the form of a central server, which can invert the shared activations to obtain images on which inference was performed. In this setting, the adversary is assumed to be one of the two collaborating parties and the proposed attack is typically constrained to shallow architectures. Another attack that employs the same threat model was presented by Zhu et al. and termed deep leakage from gradients (DLG) [47]. DLG is able to obtain a shared gradient either as the central server or as an adversarial client (when the number of data owners is relatively small). Gradient-based reconstruction attacks can, in comparison to activation-based attacks, obtain images that are either identical to or indistinguishable from the ones used to train the model, rendering them a significant threat to a large number of CML implementations.

While the aforementioned attacks can be threatening to CML implementations such as FL, they are often very brittle in practice and rely on a number of assumptions about the training protocol. Typical assumptions include that the adversary is limited to smaller models, datasets of lower dimensions (e.g., smaller images) or that specific layers are present in the target model (such as BN layers). Furthermore, the adversaries in these settings are highly encouraged to perform the attack at the start of the training procedure, while the shared gradients have high magnitudes and contain more information. These attacks, in addition, both assume that the adversary is part of the training consortium and/or is the aggregation server, which in most cases implies that they have access to or prior knowledge of the training data. However, in practice, gradient-based reconstruction attacks often do not consider this in their choice of the threat model: the adversary does not rely on their own data (that in a number of cases can be similar to the data that they are trying to reconstruct). As a result, in a number of contexts, these attacks fail to produce images with high fidelity or only provide incomplete reconstructions. We contend that this is an omission, as the possession of priors with features similar to the training dataset (or, in certain cases publicly available datasets with similar data distributions) is inherent to many forms of FL, and thus their exploitation should be considered.

In this work, we explore how principled adversarial prior exploitation allows to produce more accurate reconstructions both qualitatively and quantitatively. We analyse a number of learning contexts and model architectures, investigating not only the effects of using such data, but also how the content of the adversarial dataset affects the reconstruction. This issue is significantly more complex in heterogeneous datasets such as the ones used in facial recognition tasks (e.g., FACES [8]) with large intra-class variation [7], making reconstruction a challenging task for the adversary as even if they have the data of the same class, it might not be related to the data of their victim. Alternatively, datasets that contain images where inter-class variation is much smaller (such as abdominal computed tomography (CT) scans in medical image analysis) can be significantly more vulnerable to adversaries that can utilise their own data to facilitate the reconstruction. To exploit prior information, we utilise intermediate model activations from attacker-controlled
Fig. 1. Comparison of resulting reconstructions: our method (left) results in substantially improved reconstruction fidelity compared to the baseline method (right).

Fig. 2. Overview of a facial recognition downstream task. The adversarial priors used in our method are sourced from the same distribution as the target data.

datapoints, which we integrate in form of an additional reconstruction term in gradient-based model inversion setting. Similar to prior work [13], we assume an HbC threat model, and the only additional step we introduce is matching the activations of the image that is being reconstructed to those of the same class controlled by the attacker, similarly to a style transfer loss term from [18]. This allows us to maintain the same HbC attacker setting while enhancing the reconstructions for which the adversary controls images of the same class as the victim. We show that even an approach where such activations are used as a single, untuned penalty term, outperforms gradient-only reconstruction. We further study the properties of different datasets and architectures when it comes to assessing the severity of our method in various collaborative settings.

Finally, we discover that our method not only outperforms the existing baseline both quantitatively and qualitatively, but also leads to additional information leakage through a number of adversarial downstream tasks. One such task is facial recognition, where our method is able to produce facial reconstructions that are more similar to the original training data than the existing baseline, resulting in an adversary being able to use such images in facial spoofing attacks [10, 27] with higher fidelity (exemplary result shown in Figure 1). An overview of such a downstream attack can be found in Figure 2. Additionally, we are able to disclose a number of sensitive attributes
(such as age or gender characteristics) associated with the sensitive data even in cases where the reconstructions are incomplete. We summarise our contributions as follows:

- We explore the idea of using attacker-controlled data in a HbC setting to facilitate more accurate gradient-based reconstruction attacks;
- We evaluate a number of settings and models, showing quantitatively and qualitatively that our method outperforms attacks that rely on shared gradients alone;
- We present a number of improvements to the existing model inversion algorithm that result in more accurate inverted images (exemplary results shown in Figure 3) that can be used for further privacy-infringing downstream tasks, such as facial recognition or attribute inference attacks.

2 BACKGROUND AND PRIOR WORK

Our work extends gradient-based model inversion attacks to allow the adversary to achieve better reconstruction performance. This type of attack was first described in Zhu et al. [47] and allowed the adversary to reconstruct the training images in collaborative settings from shared gradient updates. However, this implementation was limited to images of low dimensions such as MNIST or CIFAR-10 and very shallow models.

However, both the label and the image have to be generated, forcing the adversary to perform two reconstruction tasks, significantly increasing the probability of an incorrect item being reconstructed. An advanced implementation of this attack was presented by, Zhao et al. [46], namely improved deep leakage from gradients (iDLG) that exploits the properties of the cross-entropy loss function allowing the adversary to always obtain the correct ground-truth label for the captured gradient, but still being limited to simple models and datasets. This is achieved through observation of the signs of the shared gradients with respect to the target label, as this label’s corresponding gradient vector is negative, whereas every other label has a positive corresponding gradient vector. This method significantly improved the attack’s performance, but does not scale to more complex models or datasets, with the L-BFGS optimiser still not converging in such settings for the aforementioned reasons. A further improvement was presented by Geiping et al., which [13] allowed the attacker to reconstruct images with a size of $224 \times 224$ on deep models such as ResNets.

The attack produces accurate results when the victim sends a single update per image, as otherwise the attacker is forced to attempt to reconstruct a batch of images corresponding to single
Beyond Gradients: Exploiting Adversarial Priors in Model Inversion Attacks

38:5

gradient, which in most cases is an infeasible task even with this implementation. Work by Yin et al. [42] allows to partially mitigate this issue through extraction of additional information about the training data from batch normalisation (BN) layers and allows reconstruction of higher quality on batches of up to 32 images. However, this method includes the (non-trivial) assumption that BN layers are present in the target model. While typically such layers are used in centralised ML, where they improve model convergence, their application in the context of collaborative learning has been shown to be sub-optimal [22]; thus this assumption may not be universally applicable. From this brief survey, we conclude that a number of existing attack implementations with an identical HbC adversary exist, but none have so far investigated the use of priors that are available to the attacker. It is of note that concurrent work was recently published by [15], which also exploits the adversarial priors similarly to our method, but is not conditioned on the activations or the style of the adversarial priors. Additionally, there exist other variations of model inversion attacks, which exploit intermediate activations produced by the victim such as He et al. [16] where the central server is used in an inference mode and the adversary inverts the target activations to obtain the original image. This approach relies on a much stronger assumption of a compromised central server as well as on a single image being passed for inference. Another attack was also proposed by Zhang et al. [45], which relies on the adversary having access to a suitable prior and a generative model in order to reconstruct the original image from the predictions returned by the target model. Similarly to He et al. [16], the adversary is assumed to be a corrupted central server in an inference setting. While both of these attacks rely on a looser threat model, unlike the gradient-based reconstruction methods, they actually leverage adversarial priors and heavily depend upon the data that is available to the attacker.

3 METHODS

3.1 Threat Model

The general threat model of our work corresponds to the one in Geiping et al. [13] (i.e., independent, identically distributed (IID) FL setting with an adversarial client/central server that is able to capture a gradient update generated by their victim and has access to prior images from the same distribution as the victim data). We note that the adversary does not need to know the ground-truth label of the image they are attacking in advance, as it can be extracted from the gradient as described in [46].

3.2 Gradient-based Reconstruction

The overview of the original gradient-only attack is identical to prior work by Geiping et al. [13] and can be summarised as follows:

1. The adversary randomly generates an image-model update pair;
2. The adversary captures the gradient update submitted by the victim;
3. Using a suitable cost function (in our case, cosine similarity), the adversary minimises the difference between the captured and the generated updates;
4. The algorithm is repeated until the final iteration is reached.

In [13], the adversary attempts to reconstruct the image by solving the following minimisation problem:

$$\arg \min_{x' \in [0,1]^n} \left\{ 1 - \frac{\langle \nabla_{\theta} L(x, y), \nabla_{\theta} L(x', y) \rangle}{\| \nabla_{\theta} L(x, y) \|_2 \cdot \| \nabla_{\theta} L(x', y) \|_2} + \alpha \mathrm{TV}(x) \right\},$$

where $x'$ is the reconstruction target, $x$ is the ground truth, $y$ is the label, $\nabla_{\theta} L$ is the gradient with respect to the weights, $\langle \cdot \rangle$ is the inner product in $\mathbb{R}^n$, and $\| \cdot \|_2$ is the $L_2$-norm. $\alpha$ is a...
hyperparameter scaling the total variation penalty over the image, TV(x) [33]. We set \( \alpha = 10^{-6} \) in all experiments. We will refer to the aforementioned objective as the gradient loss \( l_g \).

## 3.3 Activation Matching

The overall goal of the adversary in a gradient-based model inversion attack is approximating the image (or multiple images in [42]) controlled by the victim that produces the captured gradient update. Typically, the attacker starts with a randomly initialised image, which is then optimised to minimise the distance between the captured gradient and the one produced using the generated image. However, prior works [13, 46] (as well as the concurrent work by [15]) show that not only is this a non-optimal starting point for the adversary, but this can also significantly increase the computational cost of the attack, since a number of random initialisations would fail to converge to a meaningful reconstruction. As a result, we are interested in a method that allows for a more meaningful adversarial initialisation. This can significantly restrict the adversarial search space, reducing the number of unsuccessful reconstructions by initialising the optimisation process at a point on the image manifold, which is already close to the one where the target image lies. The core idea of this work is in exploitation of additional information that is already available to the adversary. The intuition behind our technique is that activations that belong to images of the same class are similar and conditioning the reconstruction on them can thus be used to guide the attack towards a result that looks similar to the original image.

### 3.3.1 Activation Matching

To achieve the above, we propose a novel “activation matching” penalty term. This additional penalty term allows us to employ a principled approach to penalise the images that look dissimilar from other images within the same class, reducing the adversarial search space, and enhancing the quality of reconstructions. This proposed term is conceptually close to the notion of the feature (or content) loss from [12]. We define this penalty term as follows:

\[
\|A' - A\|_2^2 = \frac{1}{N} \sum_{i=1}^{j} \left( a_i(x') - a_i(x) \right)^2 ,
\]

where \( A \) are the activations that correspond to the prior controlled by the attacker, \( A' \) are the activations that correspond to the generated image, and \( a_j \) are activations at layer \( j \) for the generated image \( (x') \) and the adversarial prior \( (x) \), respectively. \( N \) is the number of elements in \( A \). In contrast to [12] we use the \( L_1 \)-norm, as we found that relying on an absolute (rather than squared) difference loss results in better reconstructions (as we observed that the outlier pixels disproportionately affect the reconstruction making the \( L_1 \) loss a preferable option). Therefore, the final activation loss term is

\[
l_a := \|A' - A\|_1 = \frac{1}{N} \sum_{i=1}^{j} \left( a_i(x') - a_i(x) \right) .
\]

We then combine the \( l_g \) and \( l_a \) into a single loss term that is then passed to the adversarial optimiser. Note that initially, no scaling is performed on either of the terms. We discuss the limitations of this approach in Section 4, here, we briefly note that while such approach still improves reconstruction across all settings, it does not result in significant improvements for those images that were not previously reconstructed using the baseline implementation. As a result, we turn our attention to investigating how these additional activation values can be best used by the adversary.

### 3.3.2 Scaling \( l_a \)

We notice that when comparing the penalty terms there is a clear imbalance with \( l_a \) being approximately \( 10 \times \) larger than the \( l_g \) for the ConvNet architecture (we use the ConvNet name to describe the architecture described in [13]) and for ResNet architectures. As
a result, we decided to experiment with various scaling factors in order to investigate the relationship between these two terms when it comes to assessing the reconstruction quality. Initially, we perform an analysis of the relative magnitudes of the individual terms in order to determine the suitable scaling factors. We report the results for ResNet-9 on the ImageNet and Paediatric Pneumonia Prediction datasets (PPPD) (from [19]) in Figures 4 and 5. We chose these datasets for exemplary purposes (representing multi- and single-channeled data), and similar results were obtained on other datasets.

As evident from the magnitudes of the activation term weighted against the gradient penalty term, scaling is required to prevent one of the terms overpowering the contributions of the other. Thus, we selected a number of scaling factors $s_a$ and $s_g$ for both penalty terms where $s_a \in [0.1, 0.5, 1.0]$ and $s_g \in [1.0, 5.0, 10.0]$. We report results from these experiments in Section 4. In general, we find that our method is relatively insensitive to the choice of hyperparameters, allowing the adversary to benefit from a higher reconstruction quality across most scaling constants. We do, however, perform a larger study exploring the importance of individual coefficients in the Supplementary materials 7.1.

3.4 Style Reconstruction Penalty Term

Additionally, when considering the problem of image reconstruction, we hypothesise that in certain contexts (such as a number of medical image analysis settings [4, 9, 31, 41]) images of the same class only display a small amount of variation across different samples (i.e., minimal
intra-class variation). Thus, if an adversary possesses data that has a number of features very similar to the training sample, they would be able to extract it with higher fidelity as they have a strong prior. We perform this experiment similarly to the previous ones by creating a separate penalty term for the style loss that corresponds to style loss in a transfer learning setting \[18\].

We first define the Gram matrix \( G^a_j(x) \) as a \( C_j \times C_j \) matrix whose elements are given by

\[
G^a_j(x)_{c,c'} = \frac{1}{C_j \times H_j \times W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} a(x)_{h,w,c} a(x)_{h,w,c'},
\]

where \( a_j(x) \) are the activations of image \( x \) at layer \( j \), which is a feature map of shape \( C_j \times H_j \times W_j \). The style loss term \( l_s \) is then the squared Frobenius norm of the difference between the generated image and the adversarial image that produced the activations

\[
l_s := \| G^a_j(x') - G^a_j(x) \|_F^2,
\]

where \( x \) is the adversarial prior and \( x' \) is the generated image.

### 3.4.1 Scaling the Combined Terms.

Similarly to activation penalty term, we observe that style reconstruction loss is significantly lower than the gradient reconstruction loss (as seen in Figures 6 and 7) and we therefore perform scaling on this value in a similar manner. We note that on average (for CIFAR-10, ImageNet, FACES, and BraTS) the gradient reconstruction term \( l_g \) is \( 10^5 \times \) larger than the style reconstruction term \( l_s \) (for ConvNet- and ResNet models) and \( 10^6 \) for VGG models. The
The gradient reconstruction term $l_g$ is also 10× smaller than the activation loss term (with an exception of the ConvNet architecture, where it is $10^2$× smaller than the activation term). Thus, we select scaling factors $s_g$ for the style reconstruction term in a range of [1.0, 10.0, 10000.0] and record our results for these values. One notable standalone architecture is ResNet-34, for which no scaling is performed on the style penalty term and the activation penalty term is downscaled by a factor of $10^2$ (irrespective of the main experimental setup). This is due to the fact that the magnitudes of the additional terms are significantly larger than the one of the gradient reconstruction term. We report the reconstruction results with various scaling coefficients in Tables 15–17. We provide a more detailed discussion as well as the experimental details on the importance of the individual scaling factors in the Supplementary material (Tables 5–8).

### 3.4.2 Final Reconstruction Term

The final (configurable) reconstruction loss term ($L_r$) given an adversarial prior $x$ and the generated (or initialised) image $x'$ is given by:

$$L_r = s_g \cdot l_g + s_a \cdot l_a + s_s \cdot l_s,$$

(5)

with $l_g$, $l_a$, and $l_s$ defined as above and $s_g$, $s_a$, and $s_s$ representing the corresponding scaling parameters. With this loss term instead of the original gradient-based loss in [13], we proceed to execute the reconstruction for $n$ iterations, where $n$ depends on the dataset but is generally in a range $[2000, 16000]$ iterations.

The overall quality of the reconstructed image is measured in line with prior work in terms of mean square error (MSE), peak signal-to-noise (PSNR), and structural similarity metric (SSIM) between the target image and the generated image. We present our quantitative results in the Supplementary material (Tables 18–27).

### 4 EXPERIMENTS

Here, we compare our method against the baseline gradient model inversion implementation by Geiping et al. [13]. Alternative implementations of the model inversion attack exist such as the original DLG and iDLG methods, but these were superseded by [13]. We also note that while we do not directly compare our method to [15, 42] (as these are not model-agnostic and rely on the presence of BN layers, which is a more constrained assumption than our work), we, nonetheless, perform additional evaluation of the effects associated with various normalisation layers (and additional penalty terms) in Section 4.5.

#### 4.1 Experimental Setting

Our experimental setting was as follows: model architecture and datasets are selected in advance and shared in an IID manner across all participants. The batch size was set to 1 unless otherwise specified. The optimiser used by both the federation and the adversary was AdamW [23], the learning rate for the federation and the adversary was set to 0.1. Each adversarial class has between 3 and 10 images (the number is determined randomly). The best reconstructions for our method were selected based on the lowest combined loss value for a given adversarial prior.

All experiments were performed on a system running Linux 20.04 with 8 CPU cores at 2.4 GHz, 32 GB of RAM and an NVIDIA Quadro RTX 5,000 GPU.

We randomly selected a set of input images from the dataset not used for training (in batches of 32) and allocated these as the adversarial prior. Based on the number of classes in the target dataset, the adversary had between 3 and 12 priors per each individual captured gradient. We ran our experiments for each image in a batch and report the average results as well as the average difference with reconstructions made that rely solely on the gradients.
Fig. 8. Comparison of resulting reconstructions on ImageNet (ResNet-18): Ours (left) and baseline (right). Note that the baseline reconstruction is hallucinating 3 dogs instead of 2 present in the original image.

4.2 Datasets Descriptions

In this work, we performed an attack on 2 datasets identical to the ones deployed in [13, 42, 46], namely ImageNet (224 × 224 images) and CIFAR-10 (32 × 32 images). We additionally attack 3 more learning settings, namely: (A) the PPPD dataset (images resized to 224 × 224, 3 classes of normal/viral/bacterial), (B) the Brain Tumor Segmentation 2020 (BraTS) dataset [26] (images resized to 64 × 64 and the task changed to a binary classification of tumor/no tumor), and (C) the FACES [8] dataset (images resized to 64 × 64, 6 classes representing different emotions). For each of the “non-standard” tasks, we assign adversarial priors randomly: thus, an adversary can have a T₁ or a T₂ image for BraTS, for instance, as well as both a male and a female sample for FACES reconstruction. For smaller datasets (namely CIFAR-10, FACES and BraTS) the reconstruction was run for 2000 iterations and for larger datasets (namely ImageNet and PPPD) the number of iterations was 16000.

4.3 Targeting Different Architectures

We perform a wider study to investigate how our method performs on a larger variety of network architectures. We specifically consider architectures such as VGG11, VGG13, VGG16, ResNet-9, ResNet-18, and ResNet-34. We present results for CIFAR-10 in Table 18, for ImageNet in Table 19, PPPD in Table 20, BraTS in Table 21, and FACES in Table 22. From these results, we can deduce that our method outperforms the baseline in all settings, being particularly noticeable for popular ResNet-based architectures regardless of the learning task or the dataset, as shown in Figures 8, 3, 10, and 9. In general, VGG-based architectures perform well in terms of relative quantitative difference with the baseline attack. However, for a large number of samples, neither the baseline nor our reconstructions represented meaningful inversions. We attribute this to the fact that the attack is hyper-sensitive to differences in initialisations for larger models. Attacks executed with a smaller number of iterations (2000 or less in comparison to 16000 normally), but a larger number of restarts (8 instead of 1) are significantly more likely to produce images that are better qualitatively as well as quantitatively. This principle is true for any architecture and dataset combination, but for VGG-based architectures a smaller number of restarts essentially rendered most reconstructions incorrect. This result is in line with prior work by [48] on segmentation models with VGG backbones.

4.4 Attacking at Different Stages of Training

As discussed in [13], gradient norms get significantly smaller, and thus, the gradients themselves are less descriptive towards the end of the training procedure. As a result, reconstructions are often
unsuccessful for models that have previously been trained. We, therefore, investigate the effects our method has in these environments in order to determine if we are able to reconstruct images that were previously non-reconstructable by the adversary. Here, we attack models trained on CIFAR-10 and PPPD datasets and compare the results to their untrained counterparts. We report the results in Table 4 and exemplary reconstructions in Figure 11. In general, we see a trend similar to the results in [13], showing that trained models are less susceptible to gradient-based attacks. However, since we are able to leverage additional reconstruction terms that are independent from the gradient norm, we are able to produce reconstructions of higher quality even in uninformative settings. In general, the number of successfully reconstructed images is greatly reduced for the trained target models, resulting in a number of noisy images that do not resemble the original training data.

4.5 Additional Normalisation and Regularisation Terms
Similarly to [15, 42], we investigate how the reconstruction results can be improved by using additional penalty terms, which exploit additional information encoded in the model, e.g., the BN statistics. In this setting, we run the attack on the ResNet-18 architecture using the CIFAR-10 dataset. The attack is run 3 times comparing different additional reconstruction terms. Concretely,
we perform evaluation of the attack while employing the proposed “fidelity” term, which exploits the statistics used in BN layers. We omit the TV-based consistency term, since we already deploy it in our reconstruction term. This loss term can then be defined as

$$L_f = s_l \cdot l_l + s_b \cdot l_b.$$  

(6)

Here $l_l$ is defined as the $l_2$ norm penalty of the reconstructed image \([15, 42]\) and $l_b$ is defined as

$$l_b = \sum_l \|\mu_l(x) - \text{BN(mean)}\|_2 + \sum_l \|\sigma_l(x) - \text{BN(variance)}\|_2,$$  

(7)

where $\mu_l$ and $\sigma_l$ are the batch-wise mean and variance estimates of the feature maps at the $l^{th}$ convolutional layer. We use coefficients $s_l$ and $s_b$ to denote the scaling factors of the additional two terms. We set $s_l$ as $10^{-6}$ in line with \([42]\).

The final reconstruction term is then

$$L_r = s_g \cdot l_g + s_a \cdot l_a + s_s \cdot l_s + s_i \cdot l_i + s_b \cdot l_b$$  

(8)

We note, however, that since our method relies on a single prior for a single reconstruction principle, we expect that we cannot benefit from the batch statistics to the same extent as the aforementioned works can. As a result, we additionally experiment with the models that employ layer and group normalisation layers instead. It is of note that the authors of \([42]\) additionally rely on the “group consistency” term, which allows a joint simultaneous exploration of the adversarial search space from a number of random seeds. However, since our method essentially relies on constricting the adversarial search space instead, we do not use this term in our experiments.

Additionally, we have briefly explored the idea of adding random Gaussian noise to the reconstruction term (similarly to \([42]\)), but this has not yielded any additional benefits for the adversary. From these experiments, we determine that it is indeed possible to use additional penalty and regularisation terms to facilitate the reconstruction (some of which rely on the information from specific layers of the neural network). We note from the results in Table 1 (particularly when compared to Table 2, where $s_i = 0.0$), that the additional information contained in the normalisation layers on its own did not result in any substantial reconstruction improvements for our method, regardless of the type of normalisation. However, we also observe that the exploitation of additional regularisation in the fidelity term has resulted in improvements across all settings (particularly for the baseline approach). We hypothesise that our inability to benefit from normalisation layers occurs...
Table 1. Evaluation of the Additional Penalty Terms ($s_B = 1.0$, $s_L = 10^{-6}$)

| Method | Normalisation | PSNR (mean) | SSIM (mean) | MSE (mean) | PSNR (std) | SSIM (std) | MSE (std) |
|--------|---------------|-------------|-------------|------------|------------|------------|-----------|
| Ours   | None          | 27.5954     | 0.8813      | 0.0383     | 3.5663     | 0.0875     | 0.0319    |
| Baseline | None          | 26.5071     | 0.8611      | 0.0510     | 3.6058     | 0.0993     | 0.0458    |
| Ours   | Batch         | 27.3865     | 0.8784      | 0.0403     | 3.6818     | 0.0902     | 0.0306    |
| Baseline | Batch          | 26.7608     | 0.8664      | 0.0466     | 3.6194     | 0.0997     | 0.0380    |
| Ours   | Group         | 27.5620     | 0.8773      | 0.0410     | 3.9470     | 0.0861     | 0.0347    |
| Baseline | Group         | 26.8636     | 0.8708      | 0.0499     | 3.8760     | 0.0948     | 0.0543    |
| Ours   | Layer         | 27.4230     | 0.8735      | 0.0422     | 3.8641     | 0.1019     | 0.0373    |
| Baseline | Layer        | 26.8959     | 0.8728      | 0.0455     | 3.6726     | 0.0919     | 0.0359    |

32 images, CIFAR-10, ResNet-18.

Table 2. Evaluation of the Additional Penalty Terms ($s_B = 1.0$, $s_L = 0.0$)

| Method | Normalisation | PSNR (mean) | SSIM (mean) | MSE (mean) | PSNR (std) | SSIM (std) | MSE (std) |
|--------|---------------|-------------|-------------|------------|------------|------------|-----------|
| Ours   | None          | 27.0040     | 0.8762      | 0.0425     | 3.5455     | 0.0817     | 0.0292    |
| Baseline | None          | 25.6370     | 0.8518      | 0.0651     | 3.9255     | 0.1032     | 0.0622    |
| Ours   | Batch         | 26.5268     | 0.8690      | 0.0481     | 3.5760     | 0.0895     | 0.0355    |
| Baseline | Batch          | 25.8375     | 0.8529      | 0.0615     | 3.8743     | 0.1053     | 0.0596    |
| Ours   | Group         | 26.8030     | 0.8717      | 0.0444     | 3.3938     | 0.0858     | 0.0336    |
| Baseline | Group         | 25.8498     | 0.8604      | 0.0572     | 3.5902     | 0.0807     | 0.0451    |
| Ours   | Layer         | 27.0041     | 0.8763      | 0.0425     | 3.5456     | 0.0818     | 0.0292    |
| Baseline | Layer        | 25.6399     | 0.8518      | 0.0651     | 3.9255     | 0.1032     | 0.0622    |

32 images, CIFAR-10, ResNet-18.

due to the aforementioned problem of our method being most suitable for the individual image reconstruction, which makes BN layers uninformative in this context. However, exploration of how individual layers (or activations, similar to [13, 37]) carry additional information, which can be exploited by the adversary is a very promising avenue of future work. Finally, there is also room for exploration of the interplay between the individual penalty and regularisation coefficients and the results of the reconstruction. We briefly discuss this in Section 7.1.

4.6 Effects of DP-SGD on the Reconstruction

Similarly to prior work in the area, we are interested in the behaviour of our attack in contexts which are privacy-sensitive and employ formal methods of privacy protection. One of the most commonly used methods for private collaborative training is local differentially private training. This is often achieved through the use of differentially private stochastic gradient descent (DP-SGD) [1], where each individual model update is appropriately noised before being shared with the central server. In our case, the victim adds noise directly to their (captured) gradient in order to reduce the amount of information it reveals about their data and thus prevent the reconstruction from taking place. As seen in Table 3, DP-SGD can be very effective against gradient-based reconstruction attacks regardless of the reconstruction cost function across different privacy hyperparameters. It is of note, that even for a very modest privacy budget (clipping norm of 3.0 and noise multiplies of 0.05), both the baseline and our method can be mitigated with an average PSNR being below 10.0. While our results can be marginally better than the baseline reconstructions, we note that both methods can be effectively mitigated using methods that modify the gradient information. However, these results are not entirely unexpected, which was previously shown in a number of prior works, which have previously defined both the empirical [13, 48] and the theoretical baselines [2, 35] for effective model inversion mitigation using differentially private learning. Similarly to the results of [2], we demonstrate that it is possible to defend against such attacks even with very minimal noise addition.

ACM Transactions on Privacy and Security, Vol. 26, No. 3, Article 38. Publication date: June 2023.
4.7 Downstream Tasks based on the Reconstructions

While objective metrics of similarity between the original and the reconstructed image can be used effectively to assess the overall quality of reconstruction, it is sometimes the case that features of the image that are crucial for the main learning task do not have a significant contribution to the MSE between the two images. Therefore, in order to evaluate the effectiveness of certain tasks, we employed a number of downstream methods to assess the quality of our reconstructions for FACES datasets using alternative methods.

The first such task is an attribute inference attack, where we determine the approximate age and the attributes typically considered associated with the male or female gender of the victim through observation of the priors that were selected during the reconstruction. This essentially allows us to obtain sensitive characteristics of the victim regardless of the results of the reconstruction. The second downstream task is the clustering of similar facial images based on [34]. For the purposes of this work, we explore the idea of utilisation of inverted images to spoof the results of the clustering task: we insert a batch of images reconstructed by our method as well as the baseline method into a pool of data that contains facial images of the victim and perform clustering on this dataset. We then compare the $L_2$ distance between the embeddings produced from the images generated through our method and the baseline method. This distance can take values between 0 (identical images) and 4.0 (opposites of the spectrum), where the threshold of similarity lies between 1.0 and 1.1 (after which, in most cases, the images can be considered dissimilar [34]). This allows us to use a comparison that relies on additional features rather than on the image similarity between the reconstruction and the original image, showing that our method is capable of producing data that is not just more accurate at reconstructions, but also produces images that can be utilised for additional adversarial tasks, such as facial spoofing [27] (i.e., tricking facial recognition systems).

To further exemplify the advantages of our method when utilising the reconstructed images for the downstream tasks, we additionally include two further downstream tasks in the Supplementary material 7.2, which are run on the data obtained using the two reconstruction approaches. Concretely, we run image classification tasks using the ResNet-18 architecture (pre-trained on the respective datasets) on the reconstructions from the CIFAR-10 and ImageNet datasets (generated from the ResNet-18 model) and report the classification results in Table 9.

4.7.1 Attribute Inference from Priors. Out of 32 randomly selected images (repeated across 5 independent attack runs) for the ResNet-18 architecture, we are able to infer the approximate age in 22/32 images. We are also able to infer the attributes typically considered associated with the male or female gender of the victim in 12/32 cases. As we discuss below, while these results are very context specific, we nonetheless believe, that these findings show the additional threat associated with our model inversion method, allowing the adversary to infer the sensitive attributes of the training data regardless of the quality of the reconstruction itself. Additionally, we would like to point out that since most of the reconstructions were obtained from models that were untrained,
certain features would not have the same impact on the activations as the others. As a result, we see a number of correctly predicted, but less privacy infringing characteristics (such as hair colour) impacting other inference results (such as gender associated characteristics).

4.7.2 Facial Recognition Clustering. Finally, we perform facial clustering on the data generated by the two approaches as well as the original images in order to calculate the closeness between the reconstructions and the original data. For this we rely on the methodology from [34], where we firstly identify if the images constitute valid faces (as reconstructions are selected at random and can significantly differ in quality), then produce a corresponding embedding and perform clustering on these accordingly. A detailed description of the procedure can be found in the original work by [34]. We provide an overview diagram of the attack in Figure 2. We report our results in Figures 12 and 13. It is of note that out of 32 images only 28 constituted valid faces (for both methods). As evident from these results, our method produces a significantly larger number of imagesthat are similar to the original images (19/28 reconstructions). Furthermore, the $L_2$ distance between the original image embeddings and the ones generated by our method were, on average, 7.42% lower when compared to the baseline images.

5 DISCUSSION

In this work, we propose a novel formulation of a gradient-based model inversion attack, where an adversary that is part of the training consortium uses their own data to guide the inversion process through matching the activations and the style of the victim’s data. We see that this procedure allows the attacker to obtain reconstructions that are quantitatively and qualitatively superior to the baseline data-agnostic approach in all settings. During our experiments, we have discovered a number of independent novel insights in the context of this attack.

5.1 Not All Models are Affected Equally

As we see from Section 4, even within the same dataset, the results of the reconstruction are significantly affected by the selection of the shared model. There is a particularly noticeable divide between VGG-based architectures when compared to others: in general VGG models tend to produce much worse reconstructions and take significantly longer to produce any result that resembles the training data as seen in Figure 14. In our experimental setting, we hypothesise this comparative result arises due to 2 factors. First, (small) ConvNet models (and in fact models that are more shallow than ConvNet, such as LeNet5 or an adaptation of LeNet5 by [47] used in prior...
literature) are trivially invertible, as the number of parameters corresponding to a single image is much smaller (2,904,970 parameters) when compared to deeper models, resulting in fast adversarial convergence. Therefore, when compared to such architectures, VGG (the smallest of which has 128,807,306 parameters) nets perform worse due to their larger number of parameters and, thus, higher computational burden placed on the adversary. Second, while larger ResNet architectures are also significantly more complex (in terms of the width, depth, and the overall number of trainable parameters) than their ConvNet counterparts, we hypothesise that these do not suffer from the same issues as the larger VGG-nets due to the better signal flow through the residual components, resulting in more “useful” information available to the adversary. In addition, ResNet architectures have previously obtained the state-of-the-art performance in differentially private classification tasks [20, 32], further supporting this hypothesis. Therefore, in general, the larger the model, the less likely the reconstruction to be successful, but certain architectural components, such as residual layers, have a potential to alleviate this issue to an extent.

5.2 Metrics are Not Everything
While in the majority of cases, the reconstructions with higher PSNR and SSIM can be considered accurate in representing the target image, there are certain cases, for which this conclusion does not hold. This is of particular importance in datasets that put emphasis on smaller, but more
Beyond Gradients: Exploiting Adversarial Priors in Model Inversion Attacks

5.3 Influence of the Additional Loss Terms

When discussing Figures 5 and 7, for instance, we can visually see that both additional loss terms experience a degree of fluctuation, where they are not minimised to the same level as the gradient difference loss term. This is because despite the fact that the adversarial image and the target image share a number of characteristics (by being part of the same class), there are some variations, thus, these two terms are expected to converge to a specific point, which is not required (and not even expected) to be a minima. Otherwise, we would expect the resulting reconstruction to resemble the adversarial image, rather than the target, as while these terms are scaled to the same magnitude, the style, and the activation terms could dominate the reconstruction and dissolve the contribution of the gradient term. Therefore, the behaviour we identify above for these terms is explainable. However, one potential adaptation that remains unexplored is the effect of scaling decay. In this work, the scaling factors are pre-determined before the start of the training process and remain constant throughout. We hypothesise that their relative importance for the quality of the reconstruction might fluctuate, thus, making the exploration of individual term scaling decay an interesting investigation, potentially further enhancing the generated images further. We provide additional experimental results which can be used to better understand the influences of these scaling factors in Supplementary material 7.1.

5.4 Data Complexity Matters

In this study, we discuss two types of datasets: those with low and with high intra-class variations, where the former is represented in single-channel medical datasets and the latter in multi-channel datasets. We expect that the former dataset selection actually results in better reconstructions across the board, as the adversarial prior can often be similar to the target image both in features and style (due to the nature of medical data, as well as dependence on a single channel) when compared to their multi-channel counterparts. While this conclusion is typically true qualitatively, this is not always the case quantitatively as seen from the comparison of certain extreme values in Tables 23 and 24. This happens for a number of factors, such as the complexity of the task itself (as PPPD and BraTS are not only non-standard from a learning perspective in this study, but are also larger than CIFAR-10), larger penalty associated with the difference between two individual pixels and so on. Therefore, while we can visually (and contextually) support the assumption that settings with lower intra-class variation are much more vulnerable to this variation of model inversion attack, the relative difference between the generated and the original images in such setting can be a lot less significant when compared to datasets such as ImageNet or CIFAR-10.

5.5 Every Stage of Training is Vulnerable

One major limitation of the baseline approach is its dependence on the “descriptiveness” of the gradients, i.e., reconstructions can lack in quality towards the end of the training procedure, as the norms of the associated gradients are much smaller. As a result, for larger datasets, such as ImageNet, the attack can fail much more often towards the end of the training procedure. However, as we show in Section 4, it is still possible to attack the learning settings, where the model is already
Table 4. Attacks on Previously Trained Models (32 Images, ResNet-9, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Model Type          | CIFAR-10 (87.8% Acc.) | PPPD (77.3% Acc.) |
|---------------------|------------------------|-------------------|
| Ours (mean PSNR)    | 12.5                   | 16.7              |
| Baseline (mean PSNR)| 11.2                   | 15.7              |

trained. While it is possible under certain circumstances to successfully reconstruct images in such settings with a baseline method (as reported by [13]), we note that our method allows to target models that are not accessible to the baseline approach. We attribute this to the fact that we do not rely on the gradient difference alone, allowing the adversary to utilise 3 reconstruction loss items, allowing attacks even in settings with limited gradient data available.

5.6 On the Selection of Suitable Priors

During our experiments, we assigned the adversarial datasets in a random fashion (only making sure that the victim’s data does not overlap with the adversarial data), thus limiting the advantage that they can have should the priors be selected with a specific criteria in mind. While in contexts such as pneumonia prediction, where images of the same class tend to be very similar to each other, other tasks such as ImageNet classification can feature images within the same class that do not share many features with each other. This is particularly important for the BraTS learning task, as in our example, all modalities were available to the adversary and the best one was selected based on the lowest reconstruction loss value. However, we discovered that regardless of the results of the reconstruction (i.e., even in cases where the attack was unsuccessful for larger models), the correct adversarial prior always corresponded to the correct modality used by the victim. As a result, even if such attack never returns a complete reconstruction, the adversary essentially performs an attribute inference attack instead [14], where they can infer a sensitive attribute of the data record instead of reconstructing it in full. The impact on clients’ privacy becomes significantly more profound for the FACES dataset, since each class corresponds to an emotion, they all feature a participants of different genders and ages. We discovered that the adversary is able to infer both the age and the gender characteristics of victim’s image irrespective of the results of the reconstruction, thus inferring attributes that could be sensitive otherwise. This task can be significantly more challenging in biased non-IID settings, where the adversary only has limited access to images of the same class or such images share very little with the target data (which could be the case for certain facial recognition tasks). While an in-depth investigation of “what makes a good prior good?” remains an open challenge, we still note that our method not only performs better than the baseline in most cases, but also results in an unintended disclosure of auxiliary attributes that belong to the victim even when the main attack fails and no meaningful reconstruction is returned. We consider an in-depth exploration of this property in different datasets with various architectures a promising area of future work.

5.7 Generalisation of the Adversarial Prior

In this work, we rely on a fundamental assumption: that the adversary possesses a prior which is relevant to the main learning task in order to reduce the variance (and hence reduce the adversarial search space). Since the activation (as well as the style) term is tied to an individual prior image, our proposed method would inevitably propagate the bias contained in the adversarial prior into the final reconstruction. As a result, our method’s performance can vary significantly based on the kind of data that the adversary possesses, making this attack very dataset-specific. Nonetheless, we show that our method leads to better reconstruction performance both qualitatively and
quantitatively when compared to an image generated from a random adversarial initialisation. This also implies that the downstream tasks, which employ the same data as is used during the main learning task (potentially using the adversarial data as well) can incorporate the bias associated with these adversarial (as well as the target) images. In this work, we used the pre-trained models for the downstream tasks which we explore, which means that for datasets such as ImageNet and CIFAR-10, both the adversarial and the target data were used during training, propagating bias into the results of the downstream tasks. We regard an in-depth investigation of these effects as a promising future work direction.

5.8 Suitable Losses and Regularisers

In this work, we deployed two additional reconstruction penalty components, namely the activation term and the style term (based on the proposed style loss from [18]). Here, we briefly discuss the loss function candidates as well as any additional regularisation terms, which have the potential to enhance the results of the reconstruction.

One of the most straightforward losses we could employ is the MAE or the MSE between the target image and the resulting reconstruction in order to guide the image generation process [6, 17]. However, this approach is unlikely to succeed in practice, since it requires the adversary to have a number of ground-truth target images to compare their reconstructions against. One can argue that the adversary can use their previously recovered reconstructions for this purpose, but this could propagate the bias associated with the images (or parts of images) reconstructed incorrectly. Additionally, while the adversarial priors provide the attacker with a starting point for the reconstruction, it is unlikely that this image can be used to guide the reconstruction using this loss, as the target image is likely to be significantly different to the prior (particularly for datasets with high intra-class variation such as ImageNet or FACES).

We, in contrast, are interested in approaches which could benefit from data that is semantically close but has different high-level features. A number of works in the domain of style (and texture) transfer employ loss functions similar to the ones used in this work, namely the feature loss [24] and the style/texture loss [12]. The former is often described as the "content loss": it compares the feature activations at corresponding spatial locations, preserving the spatial information. The latter, in contrast, is agnostic to spatial information. A number of works in the domains of image generation [18] and style transfer [12, 36] employ such loss functions to capture the information required for generation of high-quality natural-looking images. As these permit us to utilise features which are relevant to the reconstruction without the need to obtain the ground-truth value, any similar loss function can –in theory– be used for this task.

In addition to using an appropriate loss function, it is possible to improve the results of the reconstruction by further restricting the adversarial search space through the introduction of regularisers during the reconstruction procedure. In this work, we rely on TV for these purposes, but there exist alternatives, which could, theoretically, aid the reconstruction. One straightforward approach involves the use of an $L_2$-decay, which prevents the generated pixels from taking extreme values that do not typically occur in natural images. Another approach (namely [43]) relies on using the $\alpha$-norm of the image, encouraging the reconstruction to stay within a pre-specified target region instead of diverging, thus further limiting the adversarial reconstruction space. Similarly to TV, same authors [43] proposed using Gaussian blur in order to penalise images with high frequency information (i.e., high activations), which are unrealistic and difficult to interpret [29]. Finally, [43] also propose clipping the pixels with small norms and activations, both of which discourage the contributions of pixels, which do not contribute to the main area of interest of the image (e.g., the background). In general, most of these approaches aim to remove the extreme values from the generated images, making the result look similar to naturally-occurring images. However, given a
large number of such regularisers, it is often very difficult to find the optimum trade-off between the values of the loss and the regulariser (or multiple regularisation terms), particularly without having access to the statistics of the training dataset [24].

Overall, while certain adversarial contexts can benefit from other formulation of penalty terms and the use of additional regularisation terms it is often very difficult to identify the optimal trade-off between these, particularly without having access to the training dataset (in certain cases the entire training dataset). As a result, we note that investigation of other suitable reconstruction terms is an important research area, which may open new exploitation opportunities for the adversary and is hence an interesting avenue for future work in the area.

5.9 Future Work

While this variation of the model inversion attack achieves promising results, there exist a number of further adaptations that are currently outside of scope of this study, but are otherwise potentially beneficial for the adversary. The main problem we are facing when using additional penalty terms is the issue of scaling: while in this work, we already present quantitative and qualitative improvements over the baseline attack, we do not investigate the effects of various loss terms relative to each other. One promising approach that could help us to move away from arbitrary scaling to a more guided approach is multiple gradient descent algorithm (MGDA) used in adversarial backdoor synthesis. The ability to optimise for various learning tasks at once (i.e., to penalise the model separately for each individual loss term) would very likely lead to a significant improvement over our existing method. Furthermore, we note that certain layers in the model contribute to activations unequally (as discussed in [44]: earlier layers can have a significantly more profound impact on the quality of the transferred, or in our case reconstructed, image. In general, we believe that our method can benefit significantly from a more in-depth investigation into the domain of transfer learning in order to determine how to most effectively utilise the additional penalty terms and how to most efficiently perform the post-processing of the adversarial output to match the input image. Our method is explicitly tailored to collaborative classification problems, but we note that model inversion has previously been extended into other domains, such as medical image segmentation [48]. We believe that as the results of the prior work in the field are not yet consistent across all architectures and datasets (reconstructions often look distorted and incomplete) as well as the fact that our method performs well in homogeneous data distributions that can be associated with such studies, could make our method significantly improve the adversarial performance for such tasks. We leave an investigation of the domains beyond image classification (such as image segmentation) as future work.

5.10 Limitations

We, however, outline one limitation of the proposed method that relates to the availability of a suitable prior. In particular, while the threat model remains identical to prior literature, we make an assumption that an adversary is placed in an IID setting, allowing them to control a small proportion of data that belongs to the same class as the victim’s. Alternatively, we show that for certain datasets whose features do not experience a large inter-class variation, it is sufficient to only possess data that comes from the same distribution (e.g., pneumonia prediction task). As a result, when neither of these conditions are met, the attacker falls back to the original method and does not gain any benefit from this approach. For instance, we found that even in a non-IID setting, utilisation of the style penalty term can reduce the accuracy of the reconstruction quantitatively without having a consistent impact on the quality of the reconstruction.

Additionally, as our proposed method is an extension of a gradient-based inversion attack, it is susceptible to most of the limitations of this class of attacks, namely sensitivity to variations of
the “effective” batch size as well as to perturbations of the gradient. By “effective” batch size, we mean that any method that is capable of increasing the number of images encoded in a gradient (whether an individual gradient at each participant node or an aggregated gradient at a central server) can impair our ability to reconstruct images with high fidelity. As a result, strategies such as secure update aggregation [3] or model adaptations [37] would likely mitigate our attack. When discussing the effects of gradient perturbations, we note that differentially private model training, as demonstrated in 4.6, has proven to be highly successful in preventing model inversion attacks even for contexts where only little DP-noise is added [2, 35]. Therefore, we outline the design of gradient-attacks that are resistant to these defences as a very interesting avenue for future work. Finally, it is of note that while our method is capable of creating more accurate reconstructions, this ability comes at a much higher computational cost than some of the baselines attacks. In order to determine the optimal prior for a reconstructed image, one must perform multiple reconstructions (in our experiments 32 times, identical to the number of images in a batch) and compare them, which can result in much larger attack time (up to 8 hours to identify the best prior for a ResNet-18 ImageNet image).

6 CONCLUSION

In this work, we propose a novel model inversion attack against CML, which shares the threat model with previously discussed gradient-based attacks, but offers more accurate reconstruction results. We achieve this by leveraging the data available to the adversary obtaining the activations associated with the data class that is shared with the other participants. We empirically demonstrate that even the un-tuned implementation of this algorithm yields better reconstruction results and has a significant potential for future application to various domains including transfer learning, and multi-objective optimisation. Additionally, we demonstrate the adverse effects our attack may have in real world contexts when applied to collaborative medical image segmentation or emotion prediction, where the datasets contain particularly sensitive information. We hope that our work can be used by both the privacy research community as well as the general ML community in order to better understand the adversarial perspective when designing collaborative systems and to facilitate the development of privacy-preserving ML systems.

7 SUPPLEMENTARY MATERIAL

7.0.1 Activation Matching. Here, we present the results of the attack when we use both the gradient loss term and the activation loss term in the generation procedure. We notice that even reliance on a simple approach that involves using the difference in activations as an unscaled and unnormalised additional penalty term, we managed to obtain more accurate reconstructions in a number of settings. Here, we experiment with a number of scaling factors in order to determine the relative importance of these on the quality of the reconstruction. We present an extensive overview of our results in Tables 15 below. We note that across all experiments, our numerical results are largely contributing towards better reconstruction, while requiring a very minor adaptation of the reconstruction algorithm.

7.0.2 Activation Matching with Style Penalty. We observe that since the magnitude of the style transfer loss is negligible unless normalised and scaled, setting $s_s$ to 1.0 only has a limited effect on the results of the reconstruction. As we discuss in Section 7.0.3; however, experimenting with an increasing scaling factors allows us to further enhance the quality of the reconstructed image: both qualitatively and quantitatively in most cases. We report our findings in Table 16. As evident from results, appropriate scaling should be selected very carefully, taking the other two terms into account, as otherwise, the reconstruction quality can decrease. Additionally, we note, that under
Table 5. Evaluation of the Importance of Scaling Constants (32 Images, CIFAR-10, $s_a = 1.0, s_g = 0.0, s_s = 0.0$)

| Architecture | Method | PSNR (mean) | SSIM (mean) | MSE (mean) | PSNR (std) | SSIM (std) | MSE (std) |
|--------------|--------|-------------|-------------|------------|------------|------------|-----------|
| ResNet-9     | Ours   | 11.2836     | 0.0010      | 1.2426     | 1.3428     | 0.0027     | 0.3838    |
| ResNet-9     | Baseline | 45.3259    | 0.9933      | 0.0007     | 4.0449     | 0.0063     | 0.0007    |
| ResNet-18    | Ours   | 11.0449     | 0.0005      | 1.3808     | 2.0245     | 0.0023     | 0.5978    |
| ResNet-18    | Baseline | 26.5528    | 0.8376      | 0.04458    | 3.1230     | 0.1013     | 0.0313    |

Table 6. Evaluation of the Importance of Scaling Constants (32 Images, CIFAR-10, $s_a = 10.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | Method | PSNR (mean) | SSIM (mean) | MSE (mean) | PSNR (std) | SSIM (std) | MSE (std) |
|--------------|--------|-------------|-------------|------------|------------|------------|-----------|
| ResNet9      | Ours   | 45.8691     | 0.9927      | 0.0006     | 4.1883     | 0.0115     | 0.0006    |
| ResNet9      | Baseline | 45.2273    | 0.9909      | 0.0008     | 4.4354     | 0.0184     | 0.0012    |
| ResNet18     | Ours   | 27.0827     | 0.8479      | 0.0410     | 3.3644     | 0.1022     | 0.0311    |
| ResNet18     | Baseline | 26.5527    | 0.8375      | 0.0445     | 3.1230     | 0.1012     | 0.0312    |

a biased non-IID distribution, there is a potential for the style loss term to alter the reconstructed image, which is undesirable. We discuss this further in Section 5.

7.0.3 Scaling the Penalty Terms. We adapted the same methodology as in the experiments above, while only altering the scaling factors by which the activations, the gradient and the style terms are multiplied to investigate the dependency between the separate penalty terms and the resulting image. We report the results in Table 17.

From the results above, we determine that scaling factors of $(1.0, 10.0, 10000.0)$ for $(s_a, s_g, s_s)$ result in the largest mean performance increase and we hence deploy these coefficients in the rest of the study.

7.1 Investigating the Importance of Scaling Terms

In this work, we used logarithmically-spaced coefficients for our reconstruction terms in order to investigate their relative empirical importance. While an in-depth investigation of the optimal scaling factors is outside the scope of this work, we, nonetheless, perform a number of experiments to determine the relative importance of individual scaling factors. Specifically, we consider the cases with different dominating penalty terms (e.g., in one setting the style norm is an order of magnitude larger than the other terms). We are particularly interested in whether or not there is a possibility for the attacker to mistakenly converge to its own prior (or another unrelated image with similar style and activations) should the coefficients be poorly balanced.

We perform a number of reconstructions on the CIFAR-10 dataset using ResNet-9 and ResNet-18 architectures, as we found this setting to be the most vulnerable during our experiments. Specifically we perform three reconstructions (per architecture) where there is only a single non-zero scaling factor present (scaled to be within the range of $0$–$1.0$) as well as settings where all three factors are non-zero, but one is a factor of 10 larger than the rest. We report the results in Tables 5–8.

It is of note that when the adversary is only relying on its own adversarial prior (regardless of whether the dominating factor is the style or the activations), they are unable to achieve a meaningful reconstruction and their image converges to random noise. As a result, we conclude that our method does indeed produce images which are guided by the adversarial prior, but cannot otherwise “overtake” the reconstruction procedure and reproduce the adversarial prior itself. Additionally, we find that while it is possible to achieve marginal improvements by optimising the coefficients, there is no clear dependency between the quality of the reconstruction and the
Table 7. Evaluation of the Importance of Scaling Constants (32 Images, CIFAR-10, $s_a = 0.0, s_g = 0.0, s_s = 10000.0$)

| Architecture | Method | PSNR (mean) | SSIM (mean) | MSE (mean) | PSNR (std) | SSIM (std) | MSE (std) |
|--------------|--------|-------------|-------------|------------|------------|------------|-----------|
| ResNet9      | Ours   | 11.2835     | 0.0010      | 1.2425     | 1.3428     | 0.0027     | 0.3837    |
| ResNet9      | Baseline | 44.7783     | 0.9916      | 0.0008     | 4.1979     | 0.0144     | 0.0010    |
| ResNet18     | Ours   | 11.0449     | 0.0004      | 1.3807     | 2.0244     | 0.0023     | 0.5977    |
| ResNet18     | Baseline | 26.5527     | 0.8375      | 0.0445     | 3.1230     | 0.1012     | 0.0312    |

Table 8. Evaluation of the Importance of Scaling Constants (32 Images, CIFAR-10, $s_a = 1.0, s_g = 10.0, s_s = 100000.0$)

| Architecture | Method | PSNR (mean) | SSIM (mean) | MSE (mean) | PSNR (std) | SSIM (std) | MSE (std) |
|--------------|--------|-------------|-------------|------------|------------|------------|-----------|
| ResNet9      | Ours   | 45.8384     | 0.9928      | 0.0006     | 4.1457     | 0.0144     | 0.0008    |
| ResNet9      | Baseline | 45.3172     | 0.9914      | 0.0008     | 4.5094     | 0.0174     | 0.0011    |
| ResNet18     | Ours   | 27.0827     | 0.8479      | 0.0410     | 3.3644     | 0.1022     | 0.0311    |
| ResNet18     | Baseline | 26.5527     | 0.8375      | 0.0445     | 3.1230     | 0.1012     | 0.0312    |

Table 9. Evaluation of the Reconstructions on the Downstream Classification Tasks (32 Images, Pre-trained ResNet-18)

| Original | Ours | Geiping |
|----------|------|---------|
| CIFAR-10 | 65.6%| 59.4%   | 56.3%    |
| ImageNet | 37.5%| 25.0%   | 6.25%    |

relative magnitudes of the individual scaling factors. Thus, we leave the investigation on coefficient optimisation as future work.

7.2 Additional Downstream Tasks

In order to provide more quantitative evaluation of our results, we perform a set of additional downstream classification tasks using CIFAR-10 and ImageNet datasets. In these experiments, we use a pre-trained ResNet-18 classifier (which means that it was exposed to both the target and the adversarial data before). We randomly select a batch of 32 images reconstructed by both methods, compare the top-1 accuracy for CIFAR-10, top-5 accuracy for ImageNet. The results can be found in Table 9.

One thing to note about these results is the performance of the target model on the original data. While the top-1 accuracy is of an acceptable magnitude on CIFAR-10 dataset, this does not seem to be the case for ImageNet. As a result, multiple reconstructions were not correctly predicted by the target model, as the original images were similarly misclassified by the same target model. We observe that while we substantially outperform the baseline approach in this downstream task, in most cases (including the results on the original data) the correct predictions were not obtained.

Overall, we observe that our method has a slight quantitative advantage over the baseline when applied to these downstream tasks, further highlighting the improvements of using our proposed penalty terms.

7.3 Statistical Testing

To test for statistical significance, we used the independent samples Student $t$-Test at a pre-defined threshold of $p = 0.05$. All experiments except for CIFAR-10 are averaged across 5 independent trials. The CIFAR-10 experiments (due to much faster convergence) were averaged across 10 independent trials.
We observe a number of interesting findings. Firstly, we note that while only 5 trials are considered for the majority of datasets, the improvement across the popular ResNet-based architectures can be considered significant in most cases. This further illustrates the improvements obtained using our reconstruction strategy. Secondly, it is of note that while the results of T-testing is similar across VGG-based and the ConvNet architectures for instance (e.g., the resulting values are very similar to one another), this does not illustrate the key difference between the reconstruction results, namely that the former produce adequate reconstructions much less often than the latter. As a result, it becomes very difficult to evaluate the improvements provided by our method during such evaluation. The opposite is also true: for a number of ResNet-34 reconstructions, while our improvements are considered to be statistically significant, for more complex datasets (such as ImageNet or FACES), these improvements have often resulted in noisy meaningless reconstructions, invalidating the potential advantages of such evaluation.

Overall, while some of these results are promising, the statistical significance alone cannot put the evaluation of such attacks into a realistic context for a number of reasons. Firstly, as seen in Table 14, for instance, it is sometimes the case that only a number of metrics show statistical significance (e.g., the improvement in SSIM and PSNR leads to statistically significant results more often than the improvements in MSE in most cases) for the same model and target dataset. This is particularly noticeable for more complex high-dimensional datasets, such as ImageNet, where the MSE value may present a sub-optimal method of evaluating the resulting reconstruction. Secondly, a number of results can show substantial visual improvements when using our method, but due to the nature of the metrics, this would not be captured in the discussion on the statistical significance. Furthermore, for a number of architectures the results were not statistically significant, but it is likely due to the fact that these architectures were difficult to attack using either of the two methods, producing poor reconstruction results overall. Since an effect size is present, obtaining statistical significance would merely require repeating the experiments across all settings many more times, which would not yield additional insights but only pose additional computation challenges.

### 7.4 Quantitative Reconstruction Results

Here, we present the best-case tables of our reconstruction against the baseline. As per Section 4, an ↑ represents an increase in a value (e.g., PSNR) when compared to Geiping et al. [13] and a ↓ represents a decrease in a value. We additionally provide the absolute best-case reconstruction result with its corresponding parameters in Figure 15.
Table 12. Statistical Significance Testing, FACES Dataset

|            | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 | p-value |
|------------|---------|-------|-------|-------|----------|-----------|-----------|---------|
| MSE        | 0.0728  | 0.2936| 0.5364| 0.7748| 0.0618   | 0.9803    | 0.0857    | 0.05    |
| PSNR       | 0.0855  | 0.2702| 0.4856| 0.7914| 0.0476   | 0.8480    | 0.0399    | 0.05    |
| SSIM       | 0.0200  | 0.2201| 0.1918| 0.2861| 0.0133   | 0.8396    | 0.0481    | 0.05    |

Table 13. Statistical Significance Testing, Pneumonia Dataset

|            | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 | p-value |
|------------|---------|-------|-------|-------|----------|-----------|-----------|---------|
| MSE        | 0.4624  | 0.3170| 0.6809| 0.7438| 0.0490   | 0.0022    | 0.1871    | 0.05    |
| PSNR       | 0.6460  | 0.2823| 0.6514| 0.7911| 0.0459   | 0.0023    | 0.1925    | 0.05    |
| SSIM       | 0.2178  | 0.3478| 0.2803| 0.6827| 0.0063   | 0.0006    | 0.0726    | 0.05    |

Table 14. Statistical Significance Testing, ImageNet Dataset

|            | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 | p-value |
|------------|---------|-------|-------|-------|----------|-----------|-----------|---------|
| MSE        | 0.8795  | 0.8942| 0.8730| 0.9592| 0.3387   | 0.2910    | 0.8475    | 0.05    |
| PSNR       | 0.2830  | 0.8468| 0.8606| 0.9700| 0.0173   | 0.0248    | 0.6441    | 0.05    |
| SSIM       | 0.2101  | 0.5406| 0.6037| 0.7124| 0.0104   | 0.0224    | 0.5788    | 0.05    |

Table 15. Reconstruction with an Added Activation Matching Term (32 Images, CIFAR-10, ConvNet)

| Coefficients | (1.0, 1.0, 0.0) | (1.0, 5.0, 0.0) | (1.0, 10.0, 0.0) | (0.5, 1.0, 0.0) | (0.5, 5.0, 0.0) | (0.5, 10.0, 0.0) | (0.1, 1.0, 0.0) | (0.1, 5.0, 0.0) | (0.1, 10.0, 0.0) |
|--------------|----------------|-----------------|-----------------|----------------|----------------|-----------------|----------------|----------------|------------------|
| MSE Difference (Mean) | 0.0092 ↑ | 0.0158 ↑ | 0.0095 ↑ | 0.0125 ↑ | 0.0129 ↑ | 0.0132 ↑ | 0.0131 ↑ | 0.0152 ↑ | 0.0153 ↑ |
| PSNR Difference (Mean) | 0.3379 ↑ | 1.1389 ↑ | 0.3384 ↑ | 0.6117 ↑ | 0.6963 ↑ | 1.0023 ↑ | 1.2151 ↑ | 1.3249 ↑ | 1.5308 ↑ |
| SSIM Difference (Mean) | 0.0110 ↑ | 0.0412 ↑ | 0.0115 ↑ | 0.0192 ↑ | 0.0253 ↑ | 0.0264 ↑ | 0.0393 ↑ | 0.0411 ↑ | 0.0579 ↑ |

| Coefficients | (1.0, 1.0, 0.0) | (1.0, 5.0, 0.0) | (1.0, 10.0, 0.0) | (0.5, 1.0, 0.0) | (0.5, 5.0, 0.0) | (0.5, 10.0, 0.0) | (0.1, 1.0, 0.0) | (0.1, 5.0, 0.0) | (0.1, 10.0, 0.0) |
|--------------|----------------|-----------------|-----------------|----------------|----------------|-----------------|----------------|----------------|------------------|
| MSE Difference (Max) | 0.0633 ↓ | 0.1007 ↓ | 0.0633 ↓ | 0.0852 ↓ | 0.0910 ↓ | 0.0883 ↓ | 0.0740 ↓ | 0.1093 ↓ | 0.0945 ↓ |
| PSNR Difference (Max) | 3.1667 ↑ | 5.4371 ↑ | 3.1867 ↑ | 3.4645 ↑ | 3.4049 ↑ | 7.7177 ↑ | 3.2014 ↑ | 4.5041 ↑ | 3.8626 ↑ |
| SSIM Difference (Max) | 0.1409 ↑ | 0.2130 ↑ | 0.1409 ↑ | 0.0946 ↑ | 0.1386 ↑ | 0.1783 ↑ | 0.1673 ↑ | 0.2783 ↑ | 0.2022 ↑ |

Coefficients represent \((s_a, s_d, s_s)\) respectively. We demonstrate the difference in various metrics when comparing images produced by our method against the ones produced by Geiping et al.’s method. An increase in the value is demonstrated by ↑ and a decrease in a value is demonstrated by ↓. Best values are highlighted like this.

Fig. 15. Best reconstruction result: ours (above, PSNR of 42.1019), baseline (below, PSNR of 21.3547) (CIFAR-10, ResNet-9, \(s_a = 1.0, s_d = 10.0, s_s = 1.0, 6\) restarts, 2000 iterations). Relative difference in MSE, PSNR, and SSIM are 0.1156 ↓, 20.7472 ↑, 0.4381 ↓, respectively.
Table 16. Scaling $s$ Term (32 Images, CIFAR-10, ConvNet)

| Coefficients | MSE Difference (Mean) | PSNR Difference (Mean) | SSIM Difference (Mean) |
|--------------|-----------------------|------------------------|------------------------|
| (1.0, 1.0, 1.0) | 0.0003 ↓ | 0.0021 ↓ | 0.0248 ↑ |
| (1.0, 1.0, 100.0) | 0.0258 ↓ | 0.0070 ↓ | 0.0621 ↓ |
| (1.0, 10.0, 100.0) | 0.0254 ↓ | 0.0254 ↓ | 0.0323 ↑ |
| (1.0, 10.0, 10000.0) | 0.0126 ↓ | 0.0126 ↓ | 0.0401 ↑ |

Coefficients represent $(s_a, s_g, s_s)$ respectively.

Table 17. Scaling Individual Penalty Terms (32 Images, CIFAR-10, ConvNet)

| Coefficients | MSE Difference (Mean) | PSNR Difference (Mean) | SSIM Difference (Mean) |
|--------------|-----------------------|------------------------|------------------------|
| (1.0, 10.0, 5.0) | 0.0004 ↓ | 0.0193 ↓ | 0.0017 ↑ |
| (1.0, 10.0, 500.0) | 0.0028 ↓ | 0.0070 ↓ | 0.0281 ↓ |
| (1.0, 10.0, 50000.0) | 0.0159 ↑ | 0.2150 ↑ | 1.7756 ↑ |
| (1.0, 10.0, 1.0) | 0.0043 ↑ | 0.0148 ↑ | 0.0253 ↑ |
| (1.0, 100.0, 10000.0) | 0.0126 ↓ | 0.0126 ↓ | 0.0401 ↑ |

Coefficients represent $(s_a, s_g, s_s)$ respectively.

Table 18. Activation Matching for Various Architectures (32 Images, CIFAR10, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | MSE Difference (Mean) | PSNR Difference (Mean) | SSIM Difference (Mean) |
|--------------|-----------------------|------------------------|------------------------|
| ConvNet | 0.0095 ↓ | 0.1247 ↑ | 0.0658 ↓ |
| VGG11 | 0.0528 ↓ | 0.0329 ↓ | 0.0038 ↓ |
| VGG13 | 0.0070 ↓ | 0.0107 ↓ | 0.2100 ↓ |
| VGG16 | 0.1756 ↑ | 0.2150 ↑ | 0.6963 ↑ |
| ResNet-9 | 0.2126 ↑ | 0.0257 ↑ | 1.5366 ↑ |
| ResNet-18 | 0.0253 ↑ | 0.0043 ↑ | 0.0148 ↑ |
| ResNet-34 | 0.0287 ↑ | 0.0159 ↑ | 0.0043 ↑ |

Table 19. Activation Matching for Various Architectures (32 Images, ImageNet, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | MSE Difference (Mean) | PSNR Difference (Mean) | SSIM Difference (Mean) |
|--------------|-----------------------|------------------------|------------------------|
| ConvNet | 0.0088 ↓ | 0.0318 ↓ | 0.0329 ↓ |
| VGG11 | 0.0528 ↓ | 0.0107 ↓ | 0.0038 ↓ |
| VGG13 | 0.0070 ↓ | 0.0107 ↓ | 0.2100 ↓ |
| VGG16 | 0.1756 ↑ | 0.2150 ↑ | 0.6963 ↑ |
| ResNet-9 | 0.2126 ↑ | 0.0257 ↑ | 1.5366 ↑ |
| ResNet-18 | 0.0253 ↑ | 0.0043 ↑ | 0.0148 ↑ |
| ResNet-34 | 0.0287 ↑ | 0.0159 ↑ | 0.0043 ↑ |

Table 20. Activation Matching for Various Architectures (32 Images, PPPD, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | MSE Difference (Mean) | PSNR Difference (Mean) | SSIM Difference (Mean) |
|--------------|-----------------------|------------------------|------------------------|
| ConvNet | 0.0227 ↓ | 0.0207 ↓ | 0.0532 ↓ |
| VGG11 | 0.0529 ↓ | 0.0529 ↓ | 0.0510 ↓ |
| VGG13 | 0.0316 ↓ | 0.0316 ↓ | 0.0316 ↓ |
| VGG16 | 0.1756 ↑ | 0.2150 ↑ | 0.6963 ↑ |
| ResNet-9 | 0.0257 ↑ | 0.0043 ↑ | 0.0148 ↑ |
| ResNet-18 | 0.0043 ↑ | 0.0148 ↑ | 0.0043 ↑ |
| ResNet-34 | 0.0162 ↑ | 0.0162 ↑ | 0.0162 ↑ |

Table 21. Activation Matching for Various Architectures (32 Images, BraTS, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | MSE Difference (Mean) | PSNR Difference (Mean) | SSIM Difference (Mean) |
|--------------|-----------------------|------------------------|------------------------|
| ConvNet | 0.0222 ↓ | 0.0850 ↓ | 0.0583 ↓ |
| VGG11 | 0.0001 ↓ | 0.0001 ↓ | 0.2560 ↓ |
| VGG13 | 0.0029 ↓ | 0.0029 ↓ | 0.0010 ↓ |
| VGG16 | 5.4543 ↑ | 2.5245 ↑ | 5.5825 ↑ |
| ResNet-9 | 0.2746 ↑ | 1.1381 ↑ | 0.9124 ↑ |
| ResNet-18 | 0.0929 ↑ | 0.0398 ↑ | 0.0333 ↑ |
| ResNet-34 | 0.0163 ↑ | 0.0163 ↑ | 0.0163 ↑ |

ACM Transactions on Privacy and Security, Vol. 26, No. 3, Article 38. Publication date: June 2023.
### Table 22. Activation Matching for Various Architectures (32 Images, FACES, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 |
|--------------|---------|-------|-------|-------|----------|----------|----------|
| MSE Difference (Mean) | 0.0241 ↓ | 0.0558 ↓ | 0.0317 ↓ | 0.0239 ↓ | 0.0933 ↓ | 0.0239 ↓ | 0.1042 ↓ |
| PSNR Difference (Mean) | 0.4374 ↑ | 0.5039 ↑ | 0.1974 ↑ | 0.0614 ↑ | 1.2821 ↑ | 0.2981 ↑ | 0.2997 ↑ |
| SSIM Difference (Mean) | 0.0409 ↑ | 0.0387 ↑ | 0.0108 ↑ | 0.0013 ↑ | 0.0835 ↑ | 0.0022 ↑ | 0.0030 ↑ |

### Table 23. Activation Matching for Various Architectures (32 Images, CIFAR10, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 |
|--------------|---------|-------|-------|-------|----------|----------|----------|
| MSE Difference (Max) | 0.0633 ↓ | 0.7705 ↓ | 0.2833 ↓ | 0.2270 ↓ | 0.1156 ↓ | 0.0159 ↓ | 0.0182 ↓ |
| PSNR Difference (Max) | 3.1867 ↑ | 7.0077 ↑ | 1.2924 ↑ | 1.0595 ↑ | 20.7472 ↑ | 0.6963 ↑ | 0.7261 ↑ |
| SSIM Difference (Max) | 0.1409 ↑ | 0.4007 ↑ | 0.0844 ↑ | 0.0487 ↑ | 0.4381 ↑ | 0.0253 ↑ | 0.0287 ↑ |

### Table 24. Activation Matching for Various Architectures (32 Images, BraTS, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 |
|--------------|---------|-------|-------|-------|----------|----------|----------|
| MSE Difference (Max) | 0.1168 ↓ | 0.2404 ↓ | 0.2357 ↓ | 0.1128 ↓ | 1.2421 ↓ | 0.0941 ↓ | 0.1146 ↓ |
| PSNR Difference (Max) | 2.0597 ↑ | 2.9632 ↑ | 0.7108 ↑ | 1.2157 ↑ | 9.3620 ↑ | 2.5959 ↑ | 0.7946 ↑ |
| SSIM Difference (Max) | 0.1176 ↑ | 0.1957 ↑ | 0.0171 ↑ | 0.0298 ↑ | 0.4062 ↑ | 0.0598 ↑ | 0.0626 ↑ |

### Table 25. Activation Matching for Various Architectures (32 Images, ImageNet, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 |
|--------------|---------|-------|-------|-------|----------|----------|----------|
| MSE Difference (Max) | 0.1308 ↓ | 0.0576 ↓ | 0.3083 ↓ | 0.1168 ↓ | 0.1301 ↓ | 0.0586 ↓ | 0.0644 ↓ |
| PSNR Difference (Max) | 2.8425 ↑ | 5.4895 ↑ | 2.1320 ↑ | 2.3886 ↑ | 4.6085 ↑ | 1.9259 ↑ | 3.2286 ↑ |
| SSIM Difference (Max) | 0.1891 ↑ | 0.2210 ↑ | 0.1964 ↑ | 0.1456 ↑ | 0.2548 ↑ | 0.1438 ↑ | 0.1158 ↑ |

### Table 26. Activation Matching for Various Architectures (32 Images, PPPD, $s_a = 1.0, s_g = 10.0, s_s = 10000.0$)

| Architecture | ConvNet | VGG11 | VGG13 | VGG16 | ResNet-9 | ResNet-18 | ResNet-34 |
|--------------|---------|-------|-------|-------|----------|----------|----------|
| MSE Difference (Max) | 0.0799 ↓ | 0.1785 ↓ | 0.1238 ↓ | 0.0239 ↓ | 0.3472 ↓ | 0.1350 ↓ | 0.5757 ↓ |
| PSNR Difference (Max) | 1.4266 ↑ | 1.8020 ↑ | 0.7910 ↑ | 0.0614 ↑ | 4.4027 ↑ | 2.1986 ↑ | 1.8718 ↑ |
| SSIM Difference (Max) | 0.1029 ↑ | 0.1656 ↑ | 0.0383 ↑ | 0.0013 ↑ | 0.3054 ↑ | 0.1118 ↑ | 0.0396 ↑ |
REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 308–318.

[2] Borja Balle, Giovanni Cherubin, and Jamie Hayes. 2022. Reconstructing training data with informed adversaries. In 2022 IEEE Symposium on Security and Privacy (SP). (2022), 1138–1156.

[3] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2016. Practical secure aggregation for federated learning on user-held data. In NIPS Workshop on Private-Multi-Party Machine Learning (2016).

[4] Joseph E. Burns, Jianhua Yao, Didier Chalhoub, Joseph J. Chen, and Ronald M. Summers. 2020. A machine learning algorithm to estimate sarcopenia on abdominal CT. Academic Radiology 27, 3 (2020), 311–320.

[5] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. 2019. Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis 58 (2019), 101539.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaou Tang. 2015. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 2 (2015), 295–307.

[7] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. 2018. Pairwise confusion for fine-grained visual classification. In Proceedings of the European conference on computer vision. 70–86.

[8] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. 2010. FACES–A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods 42, 1 (2010), 351–362.

[9] Mohamed Abd Elaziz, Khalid M. Hosny, Ahmad Salah, Mohamed M. Darwish, Songfeng Lu, and Ahmed T. Sahlol. 2020. New machine learning method for image-based diagnosis of COVID-19. Plos One 15, 6 (2020), e0235187.

[10] Nesli Erdogmus and Sebastien Marcel. 2013. Spoofing 2D face recognition systems with 3D masks. In Proceedings of the 2013 International Conference of the BIOSIG Special Interest Group. IEEE, 1–8.

[11] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 1322–1333.

[12] Leon Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. Advances in Neural Information Processing Systems 28 (2015), 262–270.

[13] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems 33 (2020), 16937–16947.

[14] Neil Zhenqiang Gong and Bin Liu. 2016. You are who you know and how you behave: Attribute inference attacks via users’ social friends and behaviors. In Proceedings of the 25th USENIX Security Symposium. 979–995.

[15] Ali Hatamizadeh, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wengqi Li, Prerna Dogra, Andrew Feng, et al. 2023. Do gradient inversion attacks make federated learning unsafe? IEEE Transactions on Medical Imaging (2023).

[16] Zecheng He, Tianwei Zhang, and Ruby B. Lee. 2019. Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference. 148–162.

[17] Jia-Bin Huang, Archana Murali, H Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. arXiv:1610.05492 (2016). Retrieved from https://arxiv.org/abs/1610.05492.

[18] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. Fedbn: Federated learning on non-iid features via local batch normalization. In International Conference on Learning Representations (2021).

[19] Ilya Loshchilov and Hutter Frank. 2019. Decoupled weight decay regularization. Proceedings of ICLR 7 (2019).
Beyond Gradients: Exploiting Adversarial Priors in Model Inversion Attacks

[24] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5188–5196.

[25] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP’19). IEEE, 691–706.

[26] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 10 (2014), 1993–2024.

[27] Zuheng Ming, Muriel Visani, Muhammad Muzzamil Luqman, and Jean-Christophe Burie. 2020. A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices. Journal of Imaging 6, 12 (2020), 139.

[28] Milad Nasr, Reza Shokri, and Amir Hoomansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy. IEEE, 739–753.

[29] Anh Nguyen, Jason Yoinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 427–436.

[30] Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. 2015. Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications 42, 4 (2015), 2162–2172.

[31] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chestx: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv:1711.05225 (2017). Retrieved from https://arxiv.org/abs/1711.05225.

[32] Nicolas W. Remerscheid, Alexander Ziller, Daniel Rueckert, and Georgios Kaisissis. 2022. SmoothNets: Optimizing CNN architecture design for differentially private deep learning. arXiv:2205.04095 (2022). Retrieved from https://arxiv.org/abs/2205.04095.

[33] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60, 1–4 (1992), 259–268.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 815–823.

[35] Pierre Stock, Igor Shilov, Ilya Mironov, and Alexandre Sablayrolles. 2022. Defending against reconstruction attacks with R\’enyi differential privacy. arXiv:2202.07623 (2022). Retrieved from https://arxiv.org/abs/2202.07623.

[36] Dmitriy Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. 2016. Texture networks: feed-forward synthesis of textures and stylized images. In Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML’16). 48 (2016) 1349–1357.

[37] Dmitrii Usynin, Daniel Rueckert, Jonathan Passerat-Palmbach, and Georgios Kaisissis. 2022. Zen and the art of model adaptation: Low-utility-cost attack mitigations in collaborative machine learning. Proceedings on Privacy Enhancing Technologies 2022, 1 (2022), 274–290.

[38] Dmitrii Usynin, Alexander Ziller, Marcus Makowski, Rickmer Braren, Daniel Rueckert, Ben Glocker, Georgios Kaisissis, and Jonathan Passerat-Palmbach. 2021. Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. Nature Machine Intelligence 3, 9 (2021), 749–758.

[39] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. 2020. A survey on distributed machine learning. ACM Computing Surveys (CSUR) 53, 2 (2020), 1–33.

[40] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2019. Dbfa: Distributed backdoor attacks against federated learning. In Proceedings of the International Conference on Learning Representations.

[41] Yuan Xu, Yuxin Wang, Jie Yuan, Qian Cheng, Xueding Wang, and Paul L. Carson. 2019. Medical breast ultrasound image segmentation by machine learning. Ultrasoundics 91 (2019), 1–9.

[42] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. 2021. See through deep visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16337–16346.

[43] Jason Yoinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding networks through deep visualization. In ICML Workshop on Deep Learning (2015).

[44] Chiyuan Zhang, Samy Bengio, and Yoram Singer. 2022. Are All Layers Created Equal? Journal of Machine Learning Research 23, 67 (2022), 1–28.

[45] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wexiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 253–261.
[46] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. arXiv:2001.02610 (2020). Retrieved from https://arxiv.org/abs/2001.02610.

[47] Ligeng Zhu and Song Han. 2020. Deep leakage from gradients. In Federated Learning. Springer, 17–31.

[48] Alexander Ziller, Dmitrii Usynin, Nicolas Remerscheid, Moritz Knolle, Marcus Makowski, Rickmer Braren, Daniel Rueckert, and Georgios Kaissis. 2021. Differentially private federated deep learning for multi-site medical image segmentation. arXiv:2107.02586 (2021). Retrieved from https://arxiv.org/abs/2107.02586.

Received 13 July 2022; revised 3 November 2022; accepted 7 March 2023