This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

Deep Face Leakage: Inverting High-quality Faces from Gradients Using Residual Optimization

Xu Zhang, Tao Xiang, *Senior Member, IEEE,* Shangwei Guo, *Member, IEEE,* Fei Yang, Tianwei Zhang, *IEEE Member*

***Abstract*—Collaborative learning has gained significant traction for training deep learning models without sharing the original data of participants, particularly when dealing with sensitive data such as facial images. However, current gradient inversion attacks are employed to progressively reconstruct private data from gradients, and they have shown successful in extracting private training data. Nonetheless, our observations reveal that these methods exhibit suboptimal performance in face reconstruction and result in the loss of numerous facial details. In this paper, we propose DFLeak, an effective approach to boost face leakage from gradients using residual optimization and thwart the privacy of facial applications in collaborative learning. In particular, we first introduce a superior initialization method to stabilize the inversion process. Second, we propose to integrate blind face restoration results into the gradient inversion optimization process in a residual manner, which enriches facial details. We further design a pixel update schedule to mitigate the adverse effects of image regularization terms and preserve fine facial details. Comprehensive experimentation demonstrates the effectiveness of our approach in achieving more realistic and higher-quality facial image reconstructions, surpassing the performance of state-of-the-art gradient inversion attacks.**

***Index Terms*—gradient inversion attack, collaborative learning, face reconstruction, data privacy**

I. Introduction

Facial data has become an integral part of various applications, ranging from the fundamental tasks of face detection [1], [2] and face recognition [3]–[5] to the creative realm of face generation [6]–[8]. This ubiquity highlights the critical role that facial data plays in our digital lives. However, the extensive use of facial data also raises concerns about privacy, especially in the context of distributed facial applications.

To address these privacy concerns, collaborative learning techniques, exemplified by federated learning, have gained prominence in the landscape of distributed facial applications [9]–[14]. In a typical collaborative learning system, a parameter server collaborates with multiple clients, each contributing its share of knowledge while safeguarding its local data privacy. This collaborative process involves the exchange of intermediate model gradients, preserving the confidentiality of individual data. The training loop, fundamental to this approach, comprises three essential steps: the server initially disseminates global model parameters to all clients; the clients then leverage their local data samples to train their local models and transmit the resulting local gradients to the server, and finally, the server updates the global model parameters using the aggregated gradients from the clients.

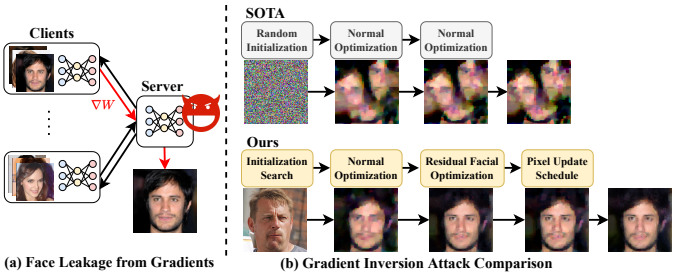


Fig. 1: Illustration of (a) the procedure of the face leakage from gradients in collaborative learning and (b) the comparison of existing and proposed gradient inversion attacks.

Despite the rigorously maintained privacy measures, recent advancements in gradient inversion attacks have cast a shadow over the privacy-preserving paradigm [15]–[20]. These attacks have demonstrated the ability to reconstruct training samples even without direct access to the private data. Fig. 1 visually portrays the intricate process involved in such reconstructions. We consider a hypothetical scenario where an attacker, characterized as an honest-but-curious server, endeavors to recover private samples by exploiting model information from clients. The attacker initiates this process by creating a dummy sample and subsequently iteratively optimizing the gradient distance between the dummy sample and the target sample. Besides, some image regularization techniques (i.e. Total Variation [17], L2-norm [18]) are introduced to further enhance the quality and fidelity of the reconstructed images.

However, the current landscape of gradient inversion attacks [15]–[20] primarily focuses on the reconstruction of generic images, leaving a pivotal question unanswered: *Can these gradient inversion attacks also yield high-quality face reconstructions?* To explore this, we embark on a systematic examination of the performance of these methods in reconstructing facial samples, revealing two notable limitations. Firstly, the reliance on random noise initialization proves suboptimal for face reconstruction due to the substantial distribution disparity between random noise and human faces. This discrepancy leads to instability in the face reconstruction process and results in unrecognizable outcomes. Secondly, while commonly-used image regularization terms offer assistance in the initial stages of image reconstruction, they falter in producing clear and detailed facial images in the later stages. This is chiefly

Tao Xiang is the corresponding author.
Xu Zhang, Tao Xiang, and Shangwei Guo are with College of Computer Science, Chongqing University, China (Email: {xuzhang, txiang, swguo}@cqu.edu.cn).
Fei Yang is with Zhejiang Lab, China (Email:yangf@zhejianglab.com)
Tianwei Zhang is with School of Computer Science and Engineering, Nanyang Technological University, Singapore (Email: tianwei.zhang@ntu.edu.sg).

attributed to the smoothing effect these regularization terms impose on important facial details and textures, ultimately leading to a loss of fidelity in the reconstructed images.

Investigating the failures of existing gradient inversion attacks has led us to two key observations. One is that using high-quality human faces as an initialization point for the optimization process can be an effective approach. The structural similarity between human faces effectively reduces the difficulty of the optimization process. Another one is that the presence of shared facial features in human faces allows for incorporating existing facial restoration techniques to efficiently complete missing details. Prior works [21]–[24] propose blind face restoration (BFR) models, which benefit to restore face images in the face reconstruction process.

Inspired by our observations, we present DFLeak, a novel approach to reconstruct face images from gradients. Fig. 1 depicts three key components of DFLeak: 1) initialization search, 2) residual facial optimization, and 3) pixel update schedule. We first pick up the face image which shares a similar distribution to the ground truth (GT) to stabilize the optimization process in the early stage. Then, we utilize a BFR model to further restore the details of reconstructed images. However, the restoration does not suffice to produce high-fidelity results. To address this limitation, we devise a residual facial optimization module to progressively enrich facial details during the reconstruction process. Additionally, we propose a pixel update schedule to adjust the pixel optimization to retain facial details.

We conduct comprehensive experiments to evaluate the effectiveness of our approach. In particular, we, compare our DFLeak with four well-established gradient inversion attacks on two human face datasets (CelebAHQ [6] and LFW [25]). Our experiments encompass various attack configurations (e.g. model structures, hyperparameter settings), training modalities (e.g. batch sizes, local steps), and defensive strategies. We also provide a comprehensive analysis to substantiate the efficacy of our proposed methodology, alongside a detailed evaluation of its computational costs. The results demonstrate that our method leads to more significant face leakage compared to state-of-the-art gradient inversion attacks. Our codes are available at https://github.com/LuckMonkeys/DFLeak.

Our main contributions are summarized as follows:

- We systematically analyze the limitations of existing gradient inversion attacks against facial gradients and introduce a novel approach, DFLeak, designed to reconstruct high-quality faces from these gradients.
- We propose an initialization search method to stabilize the gradient inversion attacks for face reconstruction.
- We design a novel face optimization and pixel update schedule to obtain high-quality face reconstruction via a BFR model in a residual manner.
- We empirically evaluate our proposed method on two datasets and compare it against four baselines. The superior qualitative and quantitative results demonstrate the effectiveness of our approach in face reconstruction.

The structure of this paper is as follows: Section II provides a review of prior research on gradient inversion attacks, emphasizing their shortcomings in face reconstruction. In Section III, we introduce the basic system and attack models. Section IV elaborates on the pipeline and key components of our proposed face reconstruction approach. Comprehensive experimental evaluations of our method are presented in Section V, with further discussion in Section VI. Finally, Section VII concludes the paper.

TABLE I: Summarization and taxonomy of existing gradient inversion attacks.

Category	Method	Initialization	Model	Dataset
Optimization	DLG [15]	Gaussian	ResNet	MNIST, CIFAR100, LFW
	iDLG [16]	Uniform	LeNet	MNIST, CIFAR100
	InvertG [17]	Gaussion	ResNet	ImageNet, CIFAR100
	GrandInv [18]	Gaussian	ResNet	ImageNet
	BayesAdv [19]	Gaussian	ConvNet	MNIST, CIFAR10
	GIAS [26]	Latent	ResNet	ImageNet
	GGL [20]	Latent	ResNet	ImageNet, CelebA
Recursion	SPN [27]	N/A	ConvNet	MNIST, CIFAR100
	R-GAP [28]	N/A	ConvNet	MNIST, CIFAR10

II. Related work

Collaborative learning poses several privacy threats, typically represented in different types of data inference attacks: membership, property, and sample. Our research focuses specifically on sample inference attacks. These attacks aim to reconstruct training samples within clients from their shared gradients, known as gradient inversion attacks. Gradient inversion can be divided into two main approaches: optimization-based and recursion-based. We provide a detailed summary of state-of-the-art gradient inversion attacks in Table I .

Optimization-based attack. Zhu et al. [15] proposed DLG to gradually approximate the training data. They first constructed a pair of dummy samples and labels, and then simulated the training procedure to obtain the gradients of dummy samples. As the gradient distance between the dummy sample and GT decreases, the dummy sample gets closer to GT. iDLG [16] simplified the optimization process by directly extracting the GT label from the gradients in the last fully-connected layer and improved the quality of reconstruction. However, their approach is only available to the low-resolution sample with a single batch size. Subsequently, numerous works have focused on recovering the training samples on a large scale and batch size. InvertG [17] optimized the cosine similarity of gradients, and added the total variation term in the image reconstruction process as regularization. Consequently, they could recover high-resolution training samples. GrandInv [18] introduced the running mean and variance in batch normalization layers to improve the fidelity of reconstruction results. BayesAdv [19] formulated the gradient leakage problem as a Bayes optimal adversary. Their proposed solution performs effectively when they have prior knowledge of image regularization and underlying defensive mechanisms. Besides, some works (GIAS [26], GGL [20]) optimized a latent variable and utilized generative models to obtain reconstruction results.

Recursion-based attack. In addition to the optimization-based attack, some researchers recursively calculated the intermediate activation of the last layer to the first of the neural network and obtained the reconstruction result from the input of the first layer. Aono et al. [29] proposed that the input of the fully-connected layer with bias could be easily recovered

from the division of their gradients. And the training data for multi fully-connected layers could be reconstructed through the recursive calculation. SPN [27] extended the reconstruction to convolutional neural networks by transferring the convolutional layer to the linear layer via stacking the filters. R-GAP [28] utilized weight and gradient constraints in training steps to construct a linear system of equations. Besides, their approach does not rely upon the existence of bias.

Limitations. While existing gradient inversion attacks have shown satisfactory performance in generic image reconstruction, they are often limited in their effectiveness for face reconstruction. Optimization-based attacks perform unstably when optimizing from a noisy dummy sample. The facial details are missing due to the inappropriate image regularization terms. Moreover, the recursive computing process only works for simple linear and convolutional neural networks, and the reconstruction performance drops significantly in deep models.

III. Problem Statement

In this section, we begin by presenting the system model, followed by an explanation of the capability and goal of the attack model.

A. System Model

We consider a standard collaborative learning system with the SGD training procedure. Each client has its own human face training dataset $\mathcal{D}$ and jointly trains the classification model $\mathcal{M}$ with the server. Let $(x, y) \in \mathcal{D}$ be a data sample and label from $\mathcal{D}$, and let ℓ and W be the loss function and model parameters of $\mathcal{M}$, respectively. At each iteration, each client trains its local model with the selected data sample and then sends the gradient to the server.

$$\nabla W = \frac{\partial \ell(x, y)}{\partial W}. \tag{1}$$

For distributed training, ∇W_k is the gradient from client k. After the server receives gradients from all clients, it updates the model parameters via gradient descent:

$$W = W - \lambda \frac{1}{K} \sum_{k=1}^{K} \nabla W_k, \tag{2}$$

where K and λ represent the total number of clients and the learning rate respectively.

B. Attack Model

Adversary capability. We assume that the server in the collaborative learning system is an honest-but-curious adversary who adheres to the training rules but attempts to reconstruct the training data using the gradients received from any clients. This type of attack falls under the category of white-box attacks, where the adversary has knowledge of the model parameters, loss function, and hyperparameters. Consequently, the adversary can calculate the gradient of any arbitrary data sample (x', y'). Furthermore, we assume that the adversary can access a human face dataset $\mathcal{D}_c$ that is disjointed to clients' training datasets, which could be a public dataset or created using recent facial generative models. The adversary can also access an existing blind face restoration model $\mathcal{M}^r$.

Adversary goal. The adversary's goal is to reconstruct the training faces from received gradients. Formally, given the gradient $\nabla W(x, y)$ from a single client (the client index is omitted for simplicity). The adversary's objective can be described as a reverse problem $\mathcal{R}$:

$$x^*, y^* = \mathcal{R}(\nabla W(x, y), \mathcal{M}, W, \ell, \mathcal{M}^r, \mathcal{D}_c). \tag{3}$$

However, solving $\mathcal{R}$ directly is a challenging task. In gradient inversion attacks, the adversary tries to minimize the distance between $\nabla W(x, y)$ and the gradient of recovered face $\nabla W(x^*, y^*)$:

$$x^*, y^* = \arg\min_{x', y'} L, \tag{4}$$

$$= \arg\min_{x', y'} D(\nabla W(x, y), \nabla W(x', y')) + \alpha\omega(x'), \tag{5}$$

$$x'_{i+1} = x'_i - \gamma \nabla L_{x'_i}, \tag{6}$$

where D is some sort of distance measurement function. $\omega(x')$ denotes the image regularization that is used to keep the reconstruction away from the unrealistic image and α represents its weight. For convenience, we use L to denote the reconstruction loss including the gradient matching and image regularization term. i and γ represent the number of attacking iterations and the learning rate of the reconstruction process. The adversary has successfully recovered the training data if x^* is visually similar to x, which could be evaluated by metrics (e.g., PSNR [30] or LPIPS [31]) adopted in image quality assessment. This type of attack can easily be extended across clients, as the adversary is able to access the gradients uploaded by each client.

IV. Methodology

In this section, we first give the insight and pipeline of our face reconstruction method. Then we delve into three key components and conclude with an algorithm summation.

A. Insights and Pipeline

Key observations. Our investigation has revealed that existing optimization-based reconstruction attacks suffer from two major limitations in face reconstruction: 1) The random noise initialization scheme leads to instability of reconstruction optimization on account of the large distribution difference between random noise and human faces. 2) The image regularization terms used in reconstruction attacks have a detrimental effect on reconstructing the facial details.

Pipeline. Fig. 2 illustrates our proposed method, which consists of four modules: initialization search, normal optimization, residual facial optimization, and pixel update schedule. Firstly, the initialization search is to select appropriate samples from existing face datasets as the initialization input to the reconstruction attack algorithm. Secondly, the normal optimization adopts the gradient inversion attack technique (discussed in Sec. III-B) to recover the facial samples. Thirdly, the residual facial optimization module utilizes the BFR model to supplement the facial details of reconstruction results.

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

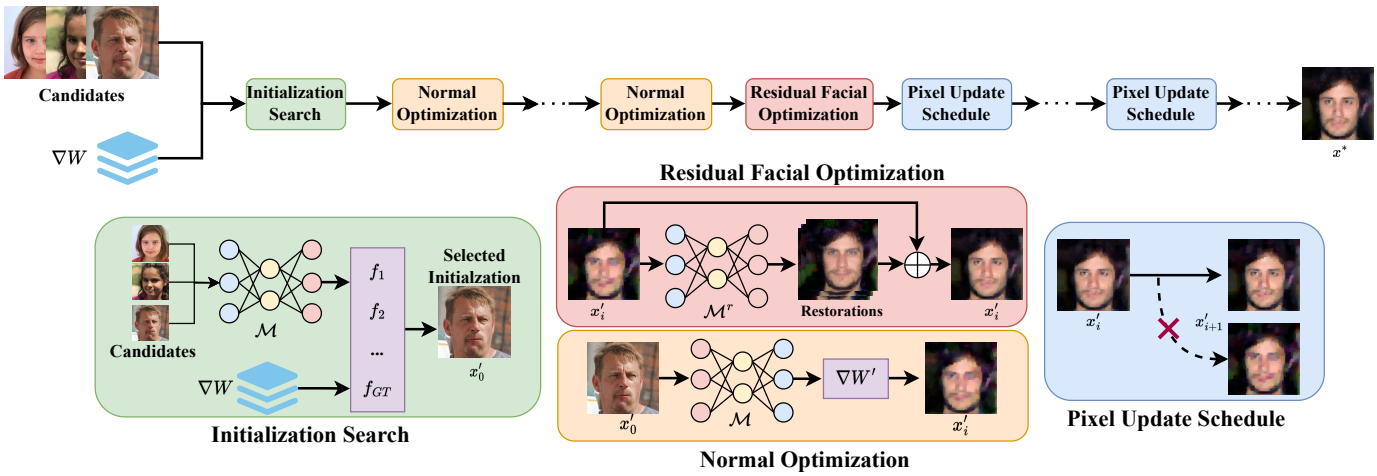


Fig. 2: Illustration of our proposed method, DFLeak. The initialization search module selects the initialization image with closer activation values with GT. The normal optimization module is the same as the common gradient inversion attacks. The residual facial optimization module blends the averaged restoration and the reconstruction of the normal optimization module to incorporate facial details into the reconstructed face. Additionally, the pixel update schedule module distributes differentweights to gradients to prevent quality degradation.

Lastly, the pixel update schedule module is to ensure that the subsequent optimization process does not discard supplemented facial details.

B. Initialization Search

Initial face search. While Gaussian noise initialization is commonly used, it may lack the necessary 'hints' or 'biases' toward desired facial features, which can lead to unstable performance in face reconstruction. Previous study [32] also points out that initializing with fewer random pixel values and closer to the GT can stabilize the optimization process and improve the reconstruction quality. Considering the similar structure of human faces, selecting a face image as the initialization is a better choice than Gaussian noise.

We use the activation distance to quantify whether both samples are semantically-close and select the image with the lowest activation difference as the initial sample x'_0:

$$x'_0 = \arg\min_{c \in \mathcal{D}_c} ||f_{\text{GT}} - f_c||_2, \tag{7}$$

where

$$f_{\text{GT}} = \nabla W^{(\text{FC})}_{m,n} / \nabla W^{(\text{FC})}_{n}, \tag{8}$$

where f_{GT} and f_c are the activation values of GT and the candidate. $\nabla W^{(\text{FC})}_{m,n}$ and $\nabla W^{(\text{FC})}_{n}$ denote the gradients of the weight and bias of the last fully-connected (FC) layer. m and n are the number of hidden features and classes respectively. To identify an appropriate initialized sample, we compute the activation distance between each candidate and GT, and select one with the minimal activation distance. We also empirically demonstrate that randomly-selected samples cannot achieve satisfactory performance compared to our proposed solution. (discussed in Section VI-A)

Label recovery. We recover the label information by identifying the negative signs of the gradient in the last fully-connected layer. Without the loss of generality, we consider recovering the label for the single-batch scenario. The computation process is as follows:

$$y' = \arg\min(\min_i \nabla W^{(\text{FC})}_{:,i}), \tag{9}$$

where i denotes the index of the class dimension. The key observation for the label recovery process is that the gradient values in the feature dimension are consistently negative $\nabla W^{(\text{FC})}_{:,i=i_y} < 0$ at the true label index i_y, and non-negative at other indices. Therefore, the negative signs serve as a robust indicator for accurate label recovery. Following the method suggested by [18], we utilize the column-wise minimum gradient values to identify the true label. Initially, we determine the minimal gradient values along the feature dimensions. The class index corresponding to the absolute minimum from these values is then identified as the recovered label y'. It is subsequently used to calculate the gradients of the dummy sample.

C. Residual Facial Optimization

We observe that the conventional reconstruction attack solutions often fail to capture sufficient facial details, resulting in lacking facial details (LFD). Recent blind face restoration techniques have shown promise in alleviating this issue by utilizing high-quality face priors [21]–[24]. Therefore, we harness the potential of the BFR model to attain facial details reconstruction (FDR). However, directly applying the BFR model to address LFD issues does not yield satisfactory results. As shown in Fig. 3, the restored image exhibits high visual quality but possesses a different identity from the GT. Moreover, the performance of BFR is sensitive to random seeds. Consequently, we propose residual facial optimization, which uses the BFR model to gradually fill the facial details

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

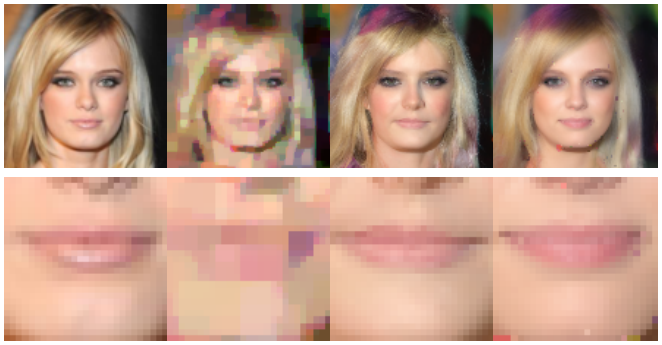


Fig. 3: Visual comparisons are conducted using normal optimization, restoration of BFR model [24], and our residual facial optimization. The second row presents an enlarged view of the mouth areas. The result of normal optimization is blurry, and restoration of the BFR model [24] holds a different identity from GT. In contrast, our residual facial optimization achieves excellent completion of facial details while accurately preserving the original features.

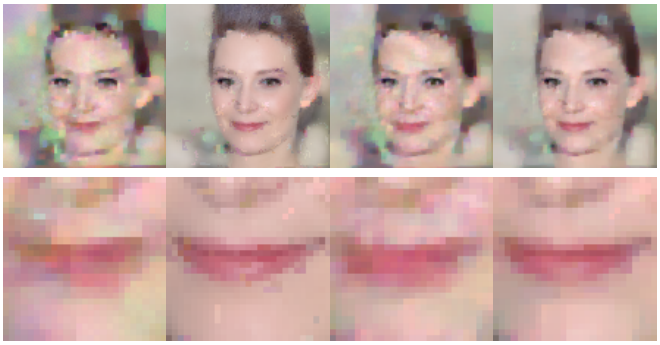


Fig. 4: Illustration of the effectiveness of our pixel update schedule. Columns 1 and 2 are the LFD images and the FDR after our residual facial optimization. Columns 3 and 4 are the subsequent reconstruction results with normal optimization and our pixel update schedule module. The second row presents an enlarged view of the mouth areas. The normal optimization blurs the image while our pixel update schedule module still preserves the facial details.

while maintaining the similarity with GT. The success of residual facial optimization depends on two aspects: 1) when to apply the BFR model and 2) how to integrate it into the face reconstruction process.

To answer the first question, we monitor the loss and apply the BFR model when the loss falls below a threshold, denoted as ϵ. This step primarily preserves the high similarity of the reconstruction result with GT. As for the second question, we progressively incorporate the restoration results, rather than directly replacing the reconstruction results with them. Specifically, the BFR model functions as a residual module to provide facial details in the reconstruction process. The residual facial optimization is formulated as follows:

$$x' = \mathbf{M}(\eta) \odot x^r + (1 - \mathbf{M}(\eta)) \odot x', \tag{10}$$

where

$$x^r = \mathbb{E}(M^r(x', s)), \tag{11}$$

$$\eta = \{p \mid ||x^r_p - x'_p||_2 \leq t_\tau\}. \tag{12}$$

We first obtain the expectation of face restoration x^r from the reconstruction result x' using the BFR model $\mathcal{M}^r$ with different random seeds s. Then, we compute the residual area $\mathbf{M}(\eta)$, where η indicates the set of pixel positions where the pixel difference between x' and x^r is equal or less than the τ-th smallest pixel difference, denoted by t_τ. Besides, $\odot$ is the element-wise product. We use p to denote the pixel position and τ represents the proportion of pixels that form the residual area. We increase τ with the progress of attacking iterations. Additionally, $\mathbf{M}$ is a function that converts a set of pixel positions to a binary matrix with the same shape as x'.

Algorithm 1: Face Leakage from Gradients Using Residual Optimization

Input : Gradients $\nabla W(x, y)$, $\mathcal{M}^r$, $\mathcal{D}_c$, ϵ, α_p
Output: Reconstructed image and label: x^*, y^*

x'_0, $y' \leftarrow$ InitializationSearch($\mathcal{D}_c$, $\nabla W(x, y)$);
x^*, $y^* \leftarrow x'_0$, y';
$L_{min} \leftarrow +\infty$;
$\eta \leftarrow \phi$;
for $i \leftarrow 1$ *to* T **do**
  $L \leftarrow D(\nabla W(x, y), \nabla W(x'_i, y')) + \alpha\omega(x'_i)$;
  if $L < L_{min}$ **then**
    $x^* \leftarrow x'_i$;
    $L_{min} \leftarrow L$;
  $x'_{i+1} \leftarrow x'_i - \gamma(\alpha_p \nabla L_{\mathbf{M}(\eta_i)\odot x'_i} + \nabla L_{(1-\mathbf{M}(\eta_i))\odot x'_i})$;
  if $L \leq \epsilon$ **then**
    $x^r_{i+1} \leftarrow \mathbb{E}(M^r(x'_{i+1}, s))$;
    $\eta_{i+1} = \{p \mid ||(x^r_{i+1})_p - (x'_{i+1})_p||_2 \leq t_{\tau_{i+1}}\}$;
    $x'_{i+1} = \mathbf{M}(\eta_{i+1}) \odot x^r_{i+1} + (1 - \mathbf{M}(\eta_{i+1})) \odot x'_{i+1}$;
return x^*, y^*;

D. Pixel Update Schedule

In reconstruction attacks, image regularization terms are typically introduced to yield visually-friendly reconstruction results. Fig. 4 presents that incorporating these terms would also blur facial details. To avoid this issue, we multiply a decay value (< 1) on the gradient in the residual area. This approach aims to alleviate the negative effect of normal optimization on facial details. The gradient of the face reconstruction is formulated as:

$$\nabla L_{x'} = \alpha_p \nabla L_{\mathbf{M}(\eta)\odot x'} + \nabla L_{(1-\mathbf{M}(\eta))\odot x'}, \tag{13}$$

where α_p denotes the pixel update decay value.

E. Attack Overview

To facilitate a better understanding of our proposed method, we illustrate the correlations among the three key components

as follows. The initialization search module plays a crucial role in stabilizing the optimization process during the early stages of the reconstruction attack. It provides a robust starting point for subsequent optimization, which can function independently. The residual facial optimization module and the pixel update schedule module are interconnected and complement each other. The former focuses on enhancing the facial details in the reconstructions, while the latter plays a vital role in preserving these details and preventing blurring caused by the image regularization terms. In combination, these components synergistically promote the effectiveness and efficiency of our approach in successfully reconstructing the training faces.

The details of our proposed method are summarized in Algorithm 1. Given gradient $\nabla W(x, y)$, we first perform the initialization search to obtain the starting point x'_0 and label recovery y'. During each iteration of the attack, if the reconstruction loss is lower than the loss threshold ϵ, we perform the residual facial optimization. We use three random seeds to calculate the averaged restoration and the proportion τ_i is equidistantly sampled from the interval $[0, 1]$ according to current iteration i. Then, the pixel value update in the residual area is decreased by α_p in subsequent attack iterations. At the end of the attack process, we can produce the high-quality face reconstruction x^* with minimal reconstruction loss and the label recovery y^*.

V. Experiment

In this section, we start by describing the experimental setup. We then proceed to present the overall performance of DFLeak, and end by thoroughly analyzing its effectiveness under various system settings.

A. Experimental Setup

Datasets and models. We conduct our experiments on two human face datasets: CelebAHQ [6] and LFW [25]. We attack ResNet18 [33] and MobileNetV2 [34] trained on these datasets for gender classification. To demonstrate the generalizability of our method, we also conduct an experiment on ImageNet [35] with image size 224x224. The publicly available dataset $\mathcal{D}_c$ (100 face images) is randomly sampled from the FFHQ dataset [6] for our initialization search.

Baselines. We implement four state-of-the-art optimization-based gradient inversion attacks as the baseline methods. We do not involve recursion-based gradient inversion attacks because these attacks are only applicable to simple linear and convolutional neural networks that have limited capabilities.

- *iDLG-Adam* [16]: an optimization-based attack with the Adam optimizer and a L2 gradient matching loss.
- *InvertG* [17]: the attack that is similar to iDLG-Adam but with a cosine distance loss and a total variation regularization term.
- *GradInv* [18]: the attack that improves iDLG with a total variation and L2-norm regularization terms.
- *GGL* [20]: an input-enhanced optimization-based attack that conducts latent searches using generative models.

Implementation. Following the experimental setups of previous studies [17], [18], [20], we construct a collaborative learning system with 100 clients, each receiving a random split of data from the validation dataset. Face datasets generally possess incompatible characteristics (e.g. image sizes, facial attributes). Therefore, we ensure that both the training and validation datasets are from a single face dataset. The batch size for each client is 1 if not specified. We reconstruct single batch data from every client. Besides, we also investigate scenarios with larger batch sizes to analyze their impact on the inversion performance of the involved attacks.

We implement the baselines primarily following the methodologies outlined in their respective papers [16]–[18], [20], and adjust the hyperparameters to maximally improve the face reconstruction performance of these baselines and ensure fair comparison. For iDLG-Adam, we switch to the Adam optimizer due to the challenges presented by L-BFGS in the context of large networks, as highlighted in [17]. In the case of InvertG, we decrease the weight of total variation term to 0.01, resulting improved face reconstruction outcomes. As for GradInv, we exclude the incorporation of statistics from batch normalization layers to ensure a fair comparison. The initial learning rate of all the three methods is set as 0.1 and we record the results of these attacks after 5000 iterations. For GGL, we train DCGAN on the training samples of each dataset to achieve the optimal attack results. The search dimension of DCGAN is 128 and the attack iterations is 200.

For our approach, we employ DiFace [24] as the BFR model that uses the diffusion model [36] trained on the FFHQ dataset to accomplish face restoration. We follow the attacking procedure with InvertG until first applying the BFR model. The loss threshold ϵ is set to 0.3 and the pixel update decay α_p to 0.1. The residual face optimization has been applied a total of four times, balancing the computational cost and reconstruction performance.

Evaluation Metrics. We report the averaged results of following metrics for quantitative evaluation of the similarity between GT and the face reconstruction: (1) Mean Square Error (MSE $\downarrow$), (2) Peak Signal-to-Noise Ratio (PSNR $\uparrow$), (3) Learned Perceptual Image Patch Similarity (LPIPS $\downarrow$) [31], and (4) Structural Similarity Index Measure (SSIM $\uparrow$) [37]. Note that "$\downarrow$" means the lower the metric the higher relative image quality, while "$\uparrow$" represents the higher the metric the higher image quality.

B. Overall Evaluation

We first present an overall performance comparison between our method and the involved baselines on the CelebaHQ and LFW datasets. For each dataset, we evaluate the reconstruction of data at two image resolutions (56x56 and 112x112). As shown in Table II, our method outperforms the baselines across all image quality metrics. For other methods, we notice that InvertG [17] achieves higher image quality than iDLG-Adam [16] due to the additional image regularization term. GrandInv [18] has lower reconstruction quality for the lack of batch normalization information. Compared to iDLG-Adam, GGL [20] has an advantage in the LPIPS score, but it does not perform as well in the other three metrics. Furthermore, we have observed that performing face reconstruction on data with

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

TABLE II: Quantitative comparison of the face reconstruction results on CelebAHQ and LFW datasets, where our proposed method outperforms all prior methods across all metrics.

Dataset	Attack	56x56				112x112			
		PSNR ↑	MSE ↓	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓	LPIPS ↓	SSIM ↑
CelebAHQ	iDLG-Adam [16]	16.09	0.028	0.3399	0.6554	12.63	0.0661	0.7591	0.1671
	InvertG [17]	22.58	0.0064	0.0925	0.7769	19.34	0.0252	0.2763	0.6124
	GradInv [18]	11.64	0.0743	0.596	0.2469	11.32	0.0787	0.741	0.2396
	GGL [20]	12.48	0.0633	0.2053	0.4551	11.21	0.0836	0.3933	0.3684
	Ours	**24.49**	**0.0048**	**0.0631**	**0.8074**	**22.66**	**0.0078**	**0.1600**	**0.6848**
LFW	iDLG-Adam [16]	15.65	0.0335	0.3671	0.6694	12.11	0.0686	0.8373	0.4706
	InvertG [17]	21.10	0.0139	0.1754	0.7510	18.34	0.0267	0.4086	0.5871
	GradInv [18]	12.93	0.0608	0.5653	0.2540	11.57	0.0741	0.7922	0.2576
	GGL [20]	11.97	0.0732	0.2882	0.4186	10.52	0.1005	0.5315	0.3252
	Ours	**21.88**	**0.0102**	**0.1377**	**0.7656**	**20.35**	**0.0109**	**0.3082**	**0.6458**

Fig. 5: The 112x112 visual comparison of face reconstructions using our proposed method and baselines on the CelebAHQ dataset. Row 1: GT, Row 2: iDLG-Adam [16], Row 3: InvertG [17], Row 4: GradInv [18], Row 5: GGL [20], Row 6: Ours. The face reconstructions of our method are more realistic and closer to GT.

TABLE III: Quantitative comparison of face reconstruction results (InvertG/DFLeak) with the different number of classes, conducted on the CelebAHQ dataset with 100 clients.

Classes	PSNR ↑	MSE ↓	LPIPS ↓	SSIM ↑
10	19.32/20.90	0.0340/0.0249	0.3941/0.2923	0.6018/0.6499
20	19.62/20.82	0.0298/0.0232	0.3751/0.2768	0.6061/0.6500
50	24.22/26.26	0.0052/0.0034	0.1593/0.0828	0.7220/0.7545

TABLE IV: Quantitative results on different batch sizes of the first 10 clients on the CelebAHQ dataset with image size 112x112, left: InvertG [17], right: Ours.

Batch Size	PSNR↑	MSE ↓	LPIPS ↓	SSIM ↑
2	21.95/**24.46**	0.0065/**0.0038**	0.2827/**0.1444**	0.6733/**0.7211**
4	20.34/**22.34**	0.0132/**0.0104**	0.3423/**0.2090**	0.6151/**0.6646**
8	18.54/**19.94**	0.0164/**0.0126**	0.4073/**0.2602**	0.5782/**0.6262**

higher image resolution can be more challenging. Nevertheless, our method consistently delivers outstanding performance and offers substantial performance improvements.

Fig. 5 and Fig. 6 provides the visual comparisons of face reconstructions, which also demonstrates that our method could recover more facial details and achieve the best visual quality. Furthermore, without the restriction of image regu-

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

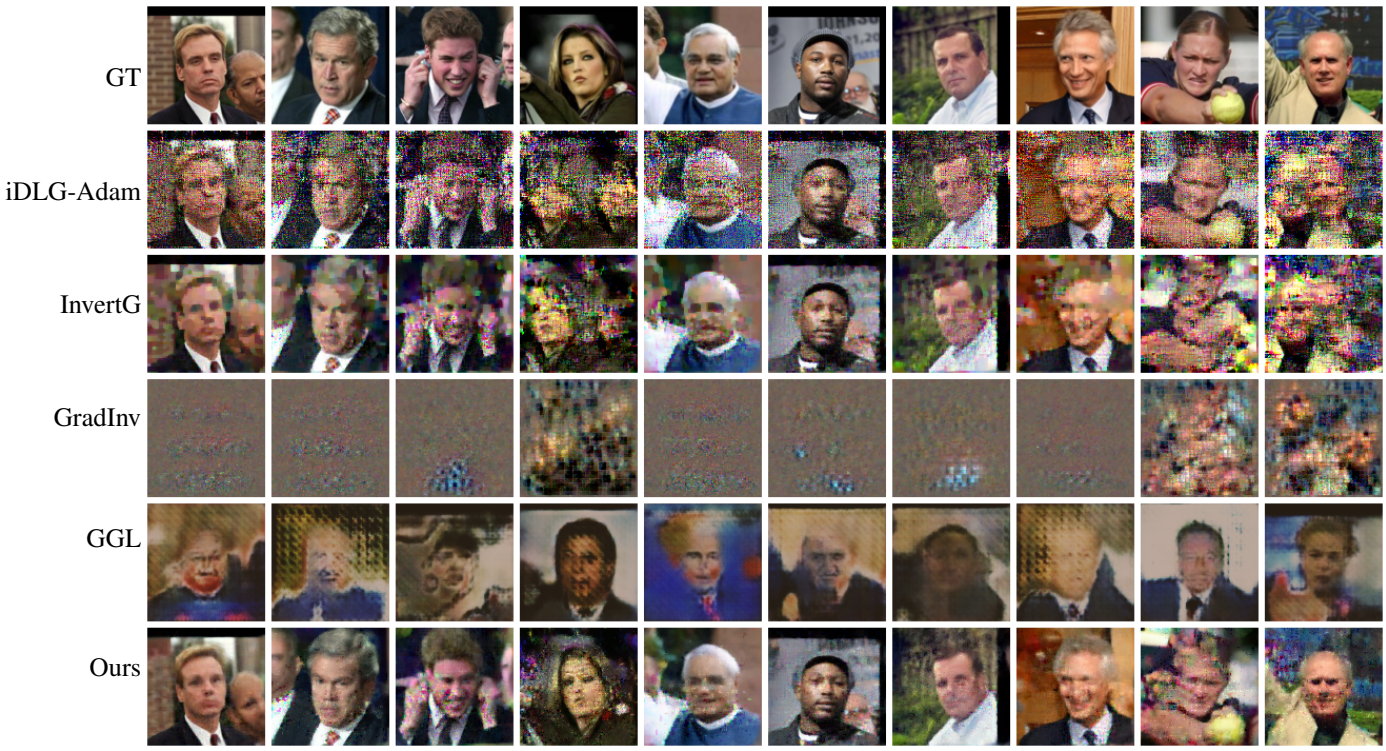
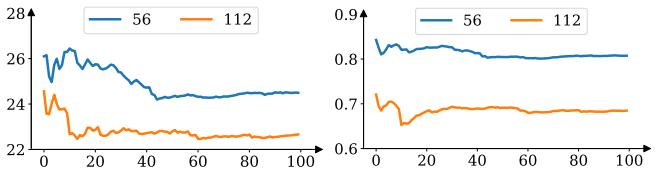


Fig. 6: The 112x112 visual comparison of face reconstructions using our proposed method and baselines on the LFW dataset. Row 1: GT, Row 2: iDLG-Adam [16], Row 3: InvertG [17], Row 4: GradInv [18], Row 5: GGL [20], Row 6: Ours. The face reconstructions of our method are more realistic and closer to GT.



(a) PSNR (b) SSIM

Fig. 7: Quantitative evaluation of the reconstruction quality (y-axis) with different client numbers (x-axis) on the CelebAHQ dataset, two image sizes: 56×56 and 112×112.

Fig. 8: The influence of different local gradient descent steps on the face reconstruction quality.

Fig. 9: Reconstructed faces with batch size 8 and image size 112x112 on the CelebAHQ dataset. Row 1: GT, Row 2: InvertG [17], Row 3:Ours.

TABLE V: Ablation study of the initialization search and residual facial optimization modules. Results are conducted on the CelebAHQ dataset with 100 clients.

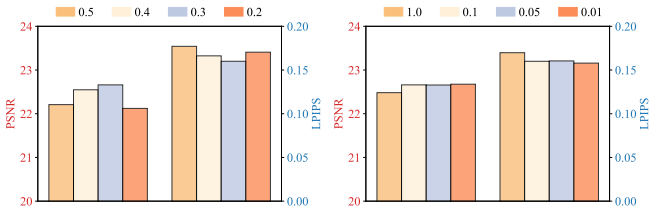
Initialization	PSNR↑	MSE ↓	LPIPS ↓	SSIM ↑
InvertG	19.34	0.0252	0.2763	0.6124
Ours	**22.66**	**0.0078**	**0.1600**	**0.6848**
w/o initialization search	20.54	0.0213	0.2272	0.6362
w/o residual facial optimization	21.56	0.0096	0.2118	0.6647
w/o pixel update schedule	22.48	0.0078	0.1698	0.6825

larization terms, the reconstructions in iDLG-Adam contain much more noise than InvertG [17]. However, such restriction also prevents further facial details reconstruction. GradInv has difficulty in recovering the facial features from the gradients without batch normalization information, as previously noted in [38]. On the other hand, GGL [20] improves the image quality via GANs, but it loses fidelity with GT.

Number of clients. Theoretically, the execution of our attack would be not impacted by the number of clients, as the reconstruction of each client's training data is handled individually. However, since our experimental results are calculated by averaging the reconstruction results of each client, an increase

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

Fig. 10: Visual comparison of the reconstruction results based on InvertG [17] with different initial strategies on the CelebAHQ dataset with image size 112x112. Row 1: GT, Row 2: Initialized from random noise, Row 3: Initialized from the searched sample.



(a) Loss threshold ϵ (b) Pixel update decay α_p

Fig. 11: Impact of hyperparameters on the performance of our method computed on 100 clients on the CelebAHQ dataset with image size 112x112.

in the number of clients contributes to more stable statistical results. We conduct experiments to measure the reconstruction quality (PSNR/SSIM) with different client numbers (from 1 to 100). As illustrated in Figure 7, the statistical results for reconstruction quality exhibit significant variability with fewer clients and stabilize when the number of clients reaches 100.

FedAvg setting. We further investigate the most practical setting of Federated Averaging in which clients share model parameter updates with multiple local gradient descent steps. We assumed that the adversary has access to the local hyperparameters (e.g. the number of local gradient descent steps, local learning rate) and performs the same operation while executing the gradient inversion attack. Fig. 8 shows the face reconstruction outcomes under multiple local gradient steps with a local learning rate of 0.001. We have observed that as the number of local gradient steps increases, the performance of our approach slightly decreases. Nevertheless, our approach still outperforms the state-of-the-art attacks.

Classes Number. We primarily assess the attack performance on the gender classification task involving two classes. To further investigate the potential impact of the classes number, we conduct additional experiments on face recognition tasks involving different numbers of identities (10, 20, and 50). We employ the ResNet18 model trained on the CelebAHQ dataset. The results, detailed in Table III, indicate that our proposed attack maintains efficacy as the number of classes increases.

Batch size. Our proposed method has been demonstrated to be effective in reconstructing individual faces. Nevertheless, the batch size is an important factor to consider during the reconstruction process. To ensure accurate label recovery, it is necessary to avoid duplicate labels within the same batch [18]. We consider the face recognition task with 50 identities on the CelebAHQ dataset and assume that no repeat labels exist within the same batch. The outcomes of various batch sizes are presented in Table IV. As the batch size increased, the quality of face reconstruction reduced due to the increased search space. However, our method still gets excellent reconstruction performance across all metrics. Fig. 9 presents the visual reconstruction results on batch size 8, our method can produce recognizable face recoveries while the InvertG [17] only gets the serious degradation without facial details.

TABLE VI: Quantitative comparison of face reconstruction with different BFR models on the CelebAHQ dataset with image size 112x112. Results are conducted with 100 clients.

BFR Models	PSNR ↑	MSE ↓	LPIPS ↓	SSIM ↑
DiFace [24]	**22.66**	**0.0078**	**0.1600**	**0.6848**
CodeFormer [21]	22.13	0.0084	0.1770	0.6791
RestoreFormer [22]	21.93	0.0088	0.1932	0.6710

TABLE VII: Performance comparison between state-of-the-art blind face restoration techniques and our DFLeak. Results are conducted on the CelebAHQ dataset with 100 clients.

Combination	PSNR↑	MSE ↓	LPIPS ↓	SSIM ↑
InvertG	19.34	0.0252	0.2763	0.6124
+ DiFace [24]	19.23	0.0978	0.2201	0.5550
+ CoderFormer [21]	19.29	0.1001	0.2400	0.5806
+ RestoreFormer [22]	18.76	0.1141	0.2767	0.5455
Ours	**22.66**	**0.0078**	**0.1600**	**0.6848**

C. Ablation Study

We provide quantitative comparisons to ablate the effectiveness of our proposed three modules in Table V. The initialization search from real face samples greatly improves the face reconstruction quality, as shown in Fig. 10, it effectively avoids failure cases in the face reconstruction process. The residual facial optimization module further enhances the quality of the recovery by introducing more facial details. Moreover, the pixel update schedule module is also beneficial to the quality of face reconstruction by maintaining facial details that are not eliminated during the reconstruction process.

D. Varying Hyperparameters and BFR Models

Hyperparameters. As illustrated in Fig. 11, we evaluate the performance of our method using varying loss thresholds ($\epsilon = 0.5, 0, 4,\ 0.3, 0.2$) and pixel update decay values ($\alpha_p = 1.0, 0.1, 0.05, 0.01$) on the CelebAHQ dataset. The face reconstructions produced with a moderate loss threshold ($\epsilon = 0.3$) exhibit the best image quality. The loss threshold plays a vital role in controlling the facial reconstruction quality upon application of the BFR model. A higher loss threshold

TABLE VIII: Quantitative comparison of the face reconstructions on CelebAHQ with image size 56x56 and MobileNetV2. Results are conducted on 100 clients.

Attack	PSNR ↑	MSE ↓	LPIPS ↓	SSIM ↑
iDLG-Adam [16]	9.21	0.1204	0.9777	0.4803
InvertG [17]	15.73	0.0290	0.3875	0.5920
GradInv [18]	12.64	0.0596	0.5801	0.3730
GGL [20]	11.02	0.0877	**0.2366**	0.4014
Ours	**16.94**	**0.0237**	0.2591	**0.6731**

TABLE IX: Quantitative comparison of InvertG [17] and our method on the ImageNet dataset. Results are conducted with 50 clients.

Attack	PSNR↑	MSE ↓	LPIPS ↓	SSIM ↑
InvertG [17]	12.73	0.0611	0.5621	**0.3869**
Ours	**12.78**	**0.0605**	**0.5248**	0.3759

leads to reconstructions with insufficient optimization, making the BFR model unsuitable for generating usable results. Conversely, a lower loss threshold reduces the number of optimization iterations after the residual facial optimization, potentially compromising the fidelity of the face reconstruction. Furthermore, our method demonstrates robustness to variations in α_p. When α_p equals 1.0, indicating no pixel update schedule, there is a noticeable decline in performance compared to other settings.

BFR models. In addition to DiFace [24], several other state-of-the-art BFR techniques are dedicated to improving the quality of images that have suffered complex degradation in real-world scenarios. We evaluated the performance of our method alongside two other notable BFR models, namely CodeFormer [21] and RestoreFormer [22]. These models utilize facial structures or facial component dictionaries derived from high-quality images as inference priors and subsequently enrich these dictionaries with diverse and detailed facial features. It's worth noting that both CodeFormer and RestoreFormer are trained on the FFHQ dataset. As demonstrated in Table VI, all three BFR models yield impressive reconstruction results. The advantage of DiFace [24] may come from the superior generative capabilities of the diffusion model.

Comparison with gradient inversion attacks with BFR models. It's worth noting that, to the best of our knowledge, our work represents the first attempt to specifically address facial leakage through gradient inversion attacks. Nevertheless, in order to explore alternative approaches, we have created several new facial leakage baselines by combining existing gradient inversion attacks with various face restoration techniques. Specifically, we have chosen InvertG as the foundation and integrated the reconstruction results with different blind face restoration methods. As shown in Table VII, we have observed that the performance of the new baselines declined due to the restoration introducing a different identity compared to GT. The results also demonstrate the effectiveness of our strategies and our method is capable of reconstructing high-quality faces from the corresponding gradients.

TABLE X: Quantitative comparison of InvertG [17] and our method with the ATS defense on the CelebAHQ dataset with image size 112x112. Results are conducted with 100 clients.

Attack	PSNR↑	MSE ↓	LPIPS ↓	SSIM ↑
InvertG [17]	19.61	0.0204	0.3143	0.5911
Ours	**20.36**	**0.0183**	**0.2710**	**0.6060**

Fig. 12: Visual comparison of the reconstruction images on the ImageNet dataset with image size 224x224. Row: 1 GT, Row 2: InvertG [17], Row 3: Ours.

E. Extending to Other Model and Dataset

MobileNetV2. In addition to ResNet18 [33], we extend our evaluation to include MobileNetV2 [34], as indicated in Table VIII. Our method demonstrates outstanding performance in face reconstruction, surpassing the baseline models. Although GGL [20] marginally outperforms our method in LPIPS score, our approach significantly outperforms it in the other three metrics. This underscores that while GGL can generate visually appealing faces, they may lack fidelity with GT.

ImageNet. We extended our method to a more general dataset, ImageNet [35] with image size 224x224, which is a frequently used dataset in previous gradient inversion attacks. We utilize the pre-trained guided diffusion model [39] in ImageNet as the "BFR" model and follow the same restoration procedure described in DiFace [24]. To save computational resources, we train the victim ResNet18 model on only 25 out of the total 1000 classes and use the same sample acceleration technique as in DiFace (100 steps with DDIM [40]). Following the same attack procedure of face reconstruction, we initialize a dummy image and optimize it by aligning its gradients with the ground truth. We employ the noisy initialization strategy on ImageNet to investigate the lower bounds of our reconstruction performance. Fig. 12 presents the visual comparison of the reconstructions in ImageNet, demonstrating that our method could effectively replace the noisy areas with high-quality image features (columns 1 to 5) and even complete the missing parts (column 6), thus enhancing the visual quality of the reconstructed images. Additionally, we provide a quantitative comparison in Table IX. The results indicate that our adapted method offers an advantage in most performance metrics, particularly LPIPS.

This article has been accepted for publication in IEEE Transactions on Image Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIP.2025.3533210

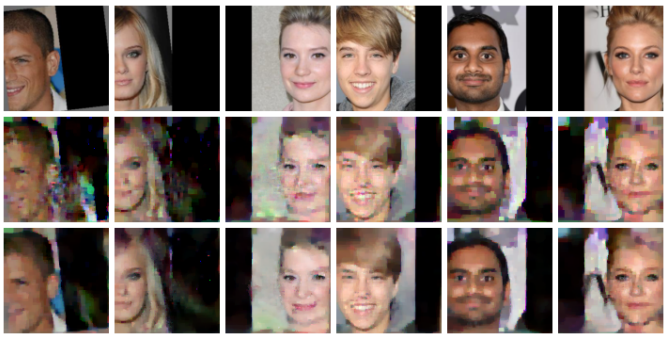


Fig. 13: Visual reconstruction results with the ATS defense on the CelebAHQ dataset with image size 112x112. Row 1: GT with ATS, Row 2: InvertG [17], Row3: Ours.

TABLE XI: Quantitative comparison of face reconstruction results with different initialization methods. "GD" denotes L2 distance between gradients and "Random Selection" indicates that initialization is randomly sampled from $\mathcal{D}_c$. Results are conducted on the CelebAHQ dataset with 100 clients.

Initialization	Start		End	
	GD ↓	SSIM/PSNR ↑	GD ↓	SSIM/PSNR ↑
Random Noise	0.8067	0.0083/7.706	0.0343	0.6124 (91.0%)/19.34 (87.8%)
Red	0.8009	0.0977/5.907	0.0365	0.6063 (89.8%)/19.06 (86.5%)
Green	0.8494	0.0878/4.845	0.0303	0.6289 (93.5%)/20.01 (90.8%)
Goldfinch	0.7312	0.1315/9.316	0.0331	0.6200 (92.2%)/19.64 (89.2%)
Random Selection	0.6508	0.1715/8.863	0.0273	0.6517 (96.9%)/21.04 (95.5%)
Initialization Search	**0.5311**	**0.1851/9.671**	**0.0231**	**0.6647 (98.8%)/21.56 (97.9%)**
Ground Truth (Ideal)	0.0	1.0/∞	0.0208	0.6728 (100%)/22.03 (100%)

F. Attack Effectiveness under Defense

Inspired by the application of data transformations in training robust deep learning models [41]–[43], we implement the recent Automatic Transformation Search (ATS) [44] as the defensive strategy. It has shown a satisfactory defense effect against gradient inversion attacks through its searched augmentation policies. We employ their strongest hybrid policies ("21-19+3-2-38") searched on the CelebA dataset, which is the combination of translation, sharpness, rotation, and solarization augmentations. As shown in Fig. 13, ATS presents poor defense effectiveness against face leakage, and our method further enhances the quality of face reconstruction. Table X presents the quantitative comparison of InvertG [17] and our method with the ATS defense on the CelebAHQ dataset. We notice that the effectiveness of our approach decreased slightly due to the reduced benefit from the initialization search scheme in the perturbed face images.

VI. Discussion

A. Explanation of the Effectiveness

The effectiveness of our method can be attributed to the two carefully designed components: the initialization search and the residual facial optimization. The initialization search strategy involves identifying an initialization with a lower gradient difference and higher visual similarity to the GT faces. We conducted experiments to assess the gradient difference and visual similarity at the initial and final stages of the reconstruction process, comparing noisy, non-facial image initializations with our facial initialization search. The results in Table XI illustrate that our facial initialization exhibits a closer proximity to the GT faces in both gradient difference and visual quality, thereby enhancing reconstruction performance. Furthermore, we use the GT images as the ideal initialization to estimate the maximum reconstruction benefits from the initialization phase. Our method achieves over 98% of the performance (SSIM) of this ideal scenario. It also shows a nearly 2% enhancement compared to Random Selection, a similar improvement observed between the Random Selection and non-facial initializations (2.1% over ' Green'). Thus, the advantage of our Initialization Search method over the Random Selection is notable.

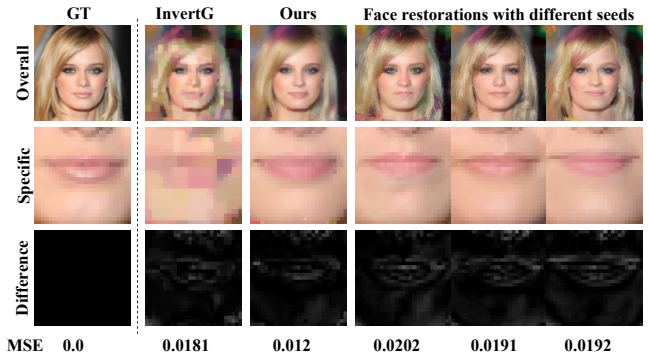


Fig. 14: Understanding the effectiveness of our residual facial optimization.

TABLE XII: Performance comparison of different methods with the same inference time.

time(s)	iDLG-Adam [16]	InvertG [17]	GradInv [18]	GGL [20]	Ours
0	0.0140	0.0141	0.0135	0.2137	0.1934
5	0.0738	0.2690	0.0163	0.2023	0.3396
10	0.0896	0.3468	0.0177	0.1556	0.4763
15	0.0949	0.3755	0.0180	0.1572	0.4943
20	0.1006	0.4000	0.0182	0.1574	0.5138
25	0.1041	0.4205	0.0189	0.1530	0.5142

On the other hand, our residual facial optimization module focuses on enriching the facial details in the reconstructed faces. One of the main challenges we encounter is the low fidelity of the facial features reconstructed through blind face restoration techniques. Consequently, these features fail to produce satisfactory reconstruction results (as depicted in Table VII). To address this issue, we progressively incorporate the restoration results, selectively leveraging the facial details that exhibit lower differences with the intermediate reconstructions. The experimental results demonstrate that our residual facial optimization technique achieves remarkable completion of facial details while maintaining the minimum deviation with the GT faces (as illustrated in Fig. 14).

B. Computation Burden Analysis

We also include a comparison of computational burden (the inference time of face reconstruction). Specifically, such burden can be interpreted from two different perspectives: time efficiency and iteration efficiency. The former is the time needed to reach equivalent performance levels, while the latter is the time taken for the same number of iterations.

TABLE XIII: Performance and inference time comparison with increased attacking iterations.

Iterations	iDLG-Adam [16]	InvertG [17]	GradInv [18]	GGL [20]	Ours
5000	0.4501/117.1	0.6145/117.3	0.2429/128.9	0.3603/63.2	0.6883/145.1
10000	0.4513/226.5	0.6095/226.2	0.2498/247.6	0.3574/125.3	-
15000	0.4477/337.2	0.6096/337.8	0.1632/368.5	0.3653/186.6	-

We first conduct the performance comparison (measured by SSIM) of different methods with the same inference time. The results in Table XII have shown that our approach achieves comparable reconstruction results with less computational time. In other words, regarding achieving the same level of performance, our DFLeak actually demonstrates a lower computation burden compared to existing attacks.

We then extend the number of attacking iterations for baseline methods and present the average SSIM value and the time required for single reconstruction in Table XIII. Even with an increased number of iterations, the existing methods fall short in achieving the same level of performance as our proposed approach. The results highlight that while our method introduces additional steps, it is not merely a trade-off between computation time and performance.

VII. Conclusion

This paper introduces DFLeak, a novel approach that analyzes facial leakage from gradients by leveraging existing facial images and blind face restoration models. Our proposed method includes an intelligent initialization scheme to reduce the possibility of failure in the face reconstruction process. Additionally, we incorporate an existing blind face restoration network into our reconstruction pipeline to provide additional facial details and improve the quality of the reconstructed image. We evaluate our proposed method on two human face datasets and demonstrate that it outperforms state-of-the-art methods in terms of both visual presentation and image quality. Moreover, We extend our method to the ImageNet dataset, which demonstrates its effectiveness on a higher resolution and more general dataset.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant No. 2022YFB3103500; the National Natural Science Foundation of China under Grants No. 62472057, U21A20463; the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grants No.2024SSYS0002; the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office.

References

[1] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[2] C.-Y. Yang and H. H. Chen, "Efficient face detection in the fisheye image domain," *IEEE Transactions on Image Processing*, 2021.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[4] W. Liu, K. I. Kou, J. Miao, and Z. Cai, "Quaternion scalar and vector norm decomposition: Quaternion pca for color face recognition," *IEEE Transactions on Image Processing*, 2022.

[5] X. Luan, Z. Ding, L. Liu, W. Li, and X. Gao, "A symmetrical siamese network framework with contrastive learning for pose-robust face recognition," *IEEE Transactions on Image Processing*, 2023.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[7] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *European Conference on Computer Vision*, 2018, pp. 282–297.

[8] J. Huo, X. Liu, W. Li, Y. Gao, H. Yin, and J. Luo, "Cast: Learning both geometric and texture style transfers for effective caricature generation," *IEEE Transactions on Image Processing*, 2022.

[9] D. Aggarwal, J. Zhou, and A. K. Jain, "Fedface: Collaborative learning of face recognition model," in *IEEE International Joint Conference on Biometrics*, 2021, pp. 1–8.

[10] S. Guo, X. Zhang, F. Yang, T. Zhang, Y. Gan, T. Xiang, and Y. Liu, "Robust and privacy-preserving collaborative learning: A comprehensive survey," *arXiv preprint arXiv:2112.10183*, 2021.

[11] S. Guo, T. Zhang, H. Yu, X. Xie, L. Ma, T. Xiang, and Y. Liu, "Byzantine-resilient decentralized stochastic gradient descent," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[12] S. Guo, T. Zhang, G. Xu, H. Yu, T. Xiang, and Y. Liu, "Topology-aware differential privacy for decentralized image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[13] G. Xu, G. Li, S. Guo, T. Zhang, and H. Li, "Secure decentralized image classification with multiparty homomorphic encryption," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[14] Y. Mi, Y. Huang, J. Ji, H. Liu, X. Xu, S. Ding, and S. Zhou, "Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain," in *ACM International Conference on Multimedia*, 2022, pp. 6755–6764.

[15] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, 2019.

[16] B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved deep leakage from gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[17] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients–how easy is it to break privacy in federated learning?" in *Advances in Neural Information Processing Systems*, 2020.

[18] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[19] M. Balunović, D. I. Dimitrov, R. Staab, and M. Vechev, "Bayesian framework for gradient leakage," *arXiv preprint arXiv:2111.04706*, 2021.

[20] Z. Li, J. Zhang, L. Liu, and J. Liu, "Auditing privacy defenses in federated learning via generative gradient leakage," *arXiv preprint arXiv:2203.15696*, 2022.

[21] S. Zhou, K. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 599–30 611, 2022.

[22] Z. Wang, J. Zhang, R. Chen, W. Wang, and P. Luo, "Restoreformer: High-quality blind face restoration from undegraded key-value pairs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 512–17 521.

[23] Y. Zhao, Y.-C. Su, C.-T. Chu, Y. Li, M. Renn, Y. Zhu, C. Chen, and X. Jia, "Rethinking deep face restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7652–7661.

[24] Z. Yue and C. C. Loy, "Difface: Blind face restoration with diffused error contraction," *arXiv preprint arXiv:2212.06512*, 2022.

[25] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[26] J. Jeon, K. Lee, S. Oh, J. Ok *et al.*, "Gradient inversion with generative image prior," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 898–29 908, 2021.

[27] L. Fan, K. W. Ng, C. Ju, T. Zhang, C. Liu, C. S. Chan, and Q. Yang, "Rethinking privacy preserving deep learning: How to evaluate and thwart privacy attacks," *arXiv preprint arXiv:2006.11601*, 2020.

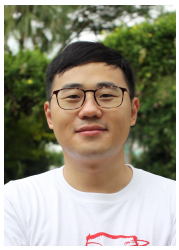
[28] J. Zhu and M. Blaschko, "R-gap: Recursive gradient attack on privacy," *arXiv preprint arXiv:2010.07733*, 2020.

[29] Y. Aono, T. Hayashi, L. Wang, S. Moriai *et al.*, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2017.

[30] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Pattern Recognition*, 2010.
[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[32] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, and Y. Wu, "A framework for evaluating gradient leakage attacks in federated learning," *arXiv preprint arXiv:2004.10397*, 2020.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[38] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7232–7241, 2021.
[39] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
[40] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
[41] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[42] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International conference on machine learning*, 2019.
[43] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," *arXiv preprint arXiv:1906.00945*, 2019.
[44] W. Gao, X. Zhang, S. Guo, T. Zhang, T. Xiang, H. Qiu, Y. Wen, and Y. Liu, "Automatic transformation search against deep leakage from gradients," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Xu Zhang received the B.E. degree from Chongqing University, China in 2021. He is currently pursuing the PhD degree at Chongqing University. His research interests include machine learning security and distributed computing security.

Tao Xiang received the BEng, MS and PhD degrees in computer science from Chongqing University, China, in 2003, 2005, and 2008, respectively. He is currently a Professor of the College of Computer Science at Chongqing University. Prof. Xiang's research interests include multimedia security, cloud security, data privacy and cryptography. He has published over 100 papers on international journals and conferences. He also served as a referee for numerous international journals and conferences.



Shangwei Guo is an associate professor in College of Computer Science, Chongqing University. He received the Ph.D. degree in computer science from Chongqing University, Chongqing, China at 2017. He worked as a postdoctoral research fellow at Hong Kong Baptist University and Nanyang Technological University from 2018 to 2020. His research interests include machine learning security, multimedia security and privacy.

Fei Yang received a B.S. and M.S. degree in computer science from Shanghai Jiao Tong University in 2011 and 2014, and the Ph.D. degree in computer science from Eindhoven University of Technology, the Netherlands, in 2018.From 2019 to 2020, He worked as a research fellow at the Cyber Security Lab in the Department of Computer Science and Engineering at Nanyang Technological University, Singapore. Since 2020, he has worked at Zhejiang Lab as an advanced research specialist. His major research interest at Zhejiang Lab includes deep learning framework, distributed computing technique, and intelligent computing platform. He is the software architect of the Digital Reactor OS project.

Tianwei Zhang is an assistant professor in School of Computer Science and Engineering, at Nanyang Technological University. His research focuses on computer system security. He is particularly interested in security threats and defenses in machine learning systems, autonomous systems, computer architecture and distributed systems. He received his Bachelor's degree at Peking University in 2011, and the Ph.D degree in at Princeton University in 2017.