Gradients protection in federated learning for Biometric authentication model

Anonymous authors
Paper under double-blind review

Abstract

In federated learning (FL) environments, biometric authentication systems encounter a distinct challenge: safeguarding user privacy without sacrificing the precision necessary for identity confirmation. Although previous FL privacy research has primarily addressed broad-spectrum protections, this paper concentrates on the particular weaknesses of biometric authentication models, especially those susceptible to gradient inversion and deep gradient leakage (DGL) attacks. We introduce an innovative privacy-preserving framework specifically designed for federated biometric authentication. Our approach employs a dual strategy: (1) an authentication model that is trained on both original and modified biometric samples to maintain resilience against input perturbations, and (2) a client-side obfuscation technique that alters biometric data prior to gradient computation, efficiently preventing reconstruction attempts. The obfuscation is adaptive and privacy-aware, selectively preserving critical biometric features necessary for authentication while discarding nonessential components to reduce input size and improve accuracy. Simultaneously, this process increases the gradient distance between the original and shared data, enhancing protection against reconstruction. Additionally, block-wise shuffling is employed to disrupt the semantic structure, ensuring that any reconstructed image lacks meaningful visual content. To validate its practical use, our framework is tested in a multibiometric context using facial and fingerprint information. The blockwise transformation strategy ensures superior authentication efficiency while reducing privacy risks. Experiments conducted in various adversarial FL settings reveal that our method significantly enhances defenses against reconstruction attacks, outperforming traditional measures.

1 Introduction

Federated Learning (FL) Rodríguez-Barroso et al. (2023) has emerged as an important framework for decentralized machine learning, enabling collaborative model training across multiple devices while keeping raw data local. This approach is particularly beneficial for applications involving sensitive biometric information, such as facial recognition and fingerprint matching. However, recent studies have highlighted significant privacy concerns within FL systems. Specifically, shared gradients may inadvertently expose sensitive information, allowing attackers to recreate the original inputs and their detailed biometric traits Geiping et al. (2020); Zhu et al. (2019); Zhao et al. (2020); Melis et al. (2019); Shokri et al. (2017).

Recent advances in gradient inversion attacks have underscored the severity of these privacy risks. As demonstrated by Wu et al. (2023), simple adaptive attacks can effectively compromise current defenses, revealing the inadequacies of existing privacy-preserving strategies in Federated Learning (FL). Additionally, Dimitrov et al. (2024) introduced SPEAR, an approach capable of accurately reversing gradients for batches exceeding a single instance, questioning the notion that increasing batch sizes inherently improves privacy.

In response to these challenges, we present a novel FL architecture tailored to maintain privacy within biometric authentication systems. This framework is composed of two separate modules: one is a client-side perturbation technique that utilizes saliency-aware obfuscation on the original biometric data, and the other is an authentication model refined to authenticate individuals using these perturbed images. By retaining raw biometric data on the client device, this architecture significantly reduces the risk of data breaches associated with shared gradients.

Unlike traditional methods that adjust each pixel separately, our perturbation technique employs block-level transformations guided by saliency maps, enabling efficient and targeted obfuscation of non-essential features. This method preserves significant identity information while enhancing computational performance, making it suitable for resource-constrained devices.

We evaluate the effectiveness of our privacy-preserving technique by applying it to the three most commonly used model architectures for classification tasks, ResNet He et al. (2015), Vision Transformer (ViT) Dosovitskiy et al. (2021), and Jigsaw ViT Chen et al. (2023). In our implementation each model is combined with ArcFace Sun et al. (2019) embeddings to make the model compatible for authentication task and provide robust identity representation for various biometric inputs.

Our approach begins by empirically generating saliency maps through experiments that assess how modifying different segments of biometric images impacts authentication accuracy. This process allows us to identify the most critical regions that contribute to reliable authentication. Building on these insights, we create new perturbed images by combining the highly crucial segments from two different biometric samples while removing less important areas. This selective mixing not only boosts authentication accuracy by preserving key identity features but also effectively destroys the semantic meaning of the reconstructed images, increasing the privacy protection. To further maximize the difference between the gradients of the original and obfuscated data, we apply a shuffling technique inspired by Jigsaw ViT Chen et al. (2023), which rearranges image patches to disrupt spatial coherence. Our perturbation methods are additionally guided by frameworks such as PuzzleMix Kim et al. (2020) and InstaHide Huang et al. (2020), ensuring a strong balance between privacy and utility.

1.1 Related Work

Gradient perturbation methods protect user data by altering the gradients exchanged during training, typically through adding noise to the gradient Zhu et al. (2019); Sun et al. (2020). Although these methods can reduce the danger of recovering the original inputs from the gradients, recent research Huang et al. (2021); Yang et al. (2022) indicates that effective protection against gradient inversion attacks (such as deep gradient leakage (DGL) Zhu et al. (2019)) requires extensive gradient perturbation. This often results in reduced model accuracy, especially in biometric authentication tasks where fine details are crucial to precision. For instance, Huang et al. (2021) showed that noise addition, sufficient to prevent inversion attacks, considerably reduces recognition accuracy.

Differential privacy (DP) improves the perturbation of gradients by mathematically determining the noise added, aiming to minimize any reduction in accuracy Bonawitz et al. (2016); McMahan et al. (2017). While DP-based federated learning (FL) has gained significant popularity, it introduces specific noise that could potentially affect authentication performance Liu et al. (2023). Balancing privacy and accuracy is an ongoing challenge, particularly when high precision is required.

Homomorphic encryption (HE) allows for processing encrypted data without disclosing the original information Cheon et al. (2017). This approach maintains privacy by ensuring that raw biometric data is never exposed to potential attackers. Despite its privacy benefits, HE is burdened by significant computational and communication overhead, making it challenging for use in real-time biometric authentication Ma et al. (2022).

Various strategies prioritize the obfuscation of raw images before training models to minimize privacy breaches. InstaHide Huang et al. (2020) achieves this by combining several images with random transformations to mask sensitive elements. Likewise, PuzzleMix Kim et al. (2020) and SaliencyMix Uddin et al. (2022) merge important and unimportant sections from different images, resulting in augmented samples that enhance generalization and robustness. These mixing techniques improve model training by merging image features, which involves integrating pixel-level or patch-level data from multiple images to form new training samples. Although these augmentations can somewhat obscure data, they are mainly tailored for centralized training and are susceptible to sophisticated gradient inversion attacks Carlini et al. (2021) as they do not modify the gradient structures that adversaries exploit.

Recent studies focus on safeguarding feature representations instead of raw inputs or gradients. For instance, Chen et al. (2024) adjusts data augmentations to reshape the loss landscape, thereby obstructing inversion

attacks while maintaining accuracy. Schwethelm et al. (2025) employ diffusion-based reconstruction attacks to expose weaknesses in differential privacy protections, stressing the necessity for defenses that consider actual image priors. However, these methods often apply globally or uniformly across all features, potentially leading to a loss of essential identity information or causing significant computational demands.

Table 3 (see Appendix A) provides an overview of the principal characteristics of existing privacy-preserving approaches, analyzed in terms of privacy effectiveness, computational expense, and effect on authentication accuracy for deployment within biometric federated learning contexts.

To overcome these limitations, our study introduces a selective obfuscation method (see figure 1) at the feature level aimed at isolating and hiding repetitive, basic features that are particularly vulnerable to inversion attacks, while maintaining the identifying information crucial for authentication precision. This obfuscation is applied locally on the client device before gradient calculations, eliminating the necessity of sharing either raw data or heavily altered gradients, thereby decreasing computational load and privacy risks. Additionally, our approach uses blockwise localized changes instead of pixelwise noise, keeping the feature structure intact and enhancing defense against recognized threats such as Deep Gradient Leakage Zhu et al. (2019) and model inversion Geiping et al. (2020). This specific obfuscation effectively balances privacy protection with high authentication accuracy, making it appropriate for federated biometric scenarios in the real world.

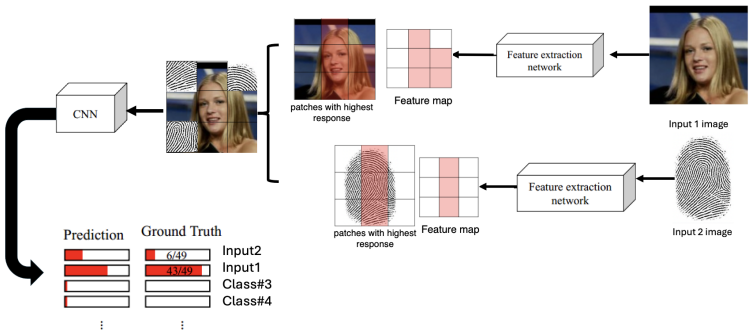

Figure 1: Overview of the proposed privacy-preserving pipeline. The architecture integrates saliency-aware feature selection, randomized data augmentation (e.g., erasing, noise, block swapping), and multi-biometric fusion with fingerprint patches. During training, the model leverages local data masking and federated updates to enhance resistance to gradient inversion attacks while preserving authentication accuracy.

2 Preliminaries

In federated learning architecture, we focus on biometric authentication through the use of facial images as input. Each image is characterized by $x \in \mathbb{R}^{H \times W \times C}$, where H, W, and C denote the image's height, width, and the number of channels, respectively. The identity label associated with each image is denoted as $y \in \mathcal{Y}$. The authentication model, denoted by $f_\theta(x)$ with parameters θ, is developed cooperatively by clients. FL framework ensures that raw data remains on local devices, thus preserving data locality and reducing potential privacy risks.

Although in federated learning, a significant threat is posed by gradient inversion attacks. These occur when an adversary captures gradients $\nabla_\theta \ell(f_\theta(x), y)$ exchanged during the training phase and uses them to reconstruct the input data x. The attacker solve the equation:

$$\hat{x} = \arg\min_{x'} \|\nabla_{x'} \ell(f_\theta(x'), y) - G\|^2 ,$$

with G being the intercepted gradient. Such attacks exploit the strong correlation between gradients and inputs, particularly in structured datasets like facial images.

2.1 Augmentation as Defense

To mitigate the risk of inversion attacks, we introduce a customized biometric data augmentation technique that obscures input samples while preserving essential attributes for authentication. Our method initiates with a mix-up strategy Pang et al. (2020) and then evolves into a spatially adaptive and more expressive enhancement.

In **Classic Mixup**, Given two samples (x_0, y_0) and (x_1, y_1), the standard mixup creates a convex combination as follows:

$$\tilde{x} = (1 - \lambda)x_0 + \lambda x_1, \quad \tilde{y} = (1 - \lambda)y_0 + \lambda y_1,$$

The parameter λ is calculated from a Beta distribution, specifically $\text{Beta}(\alpha, \alpha)$, to introduce randomness into the mixing process. However, this uniform approach (i.e. every pixel is mixed the same way, regardless of its meaning) does not account for the semantic layout of the data. The hyperparameter α determines the shape of the distribution: when $\alpha < 1$, extreme λ values close to 0 or 1 are favored, leading to a predominance of one image in the mix. In contrast, when $\alpha > 1$, the λ values cluster closer to 0.5, promoting a more even mix. Typically, α is set between 0.2 and 0.4 to allow one image to have greater influence. However, this pixel-based uniform mixing inadequately reflects the importance of semantic regions in structured data, such as facial images, where particular areas (like the eyes and nose) are more crucial to identity.

2.2 Saliency-Guided Obfuscation

We enhance the mixup methodology by presenting a saliency-based spatial augmentation specifically for biometric images. Consider x_0 as a face image and x_1 as another biometric input, such as a different face or a fingerprint. We implement a saliency mask $\mathbf{z} \in [0, 1]^{H \times W}$ to direct spatial blending:

$$\tilde{x} = (1 - \mathbf{z}) \odot x_0 + \mathbf{z} \odot x_1,$$

where $\odot$ signifies element-wise multiplication. The saliency mask $\mathbf{z}$ is influenced by a saliency map $S(x_0) \in [0, 1]^{H \times W}$, indicating the importance of each pixel in facial authentication processes. This map is empirically learned by moving a spatial window over the image and observing the decrease in authentication accuracy when obscuring that section; areas that cause a more significant dip in accuracy receive higher saliency values. To determine $\mathbf{z}$, we optimize it to reduce the saliency loss:

$$\mathcal{L}_{\text{saliency}}(\mathbf{z}; S, \tau) = \sum_{i,j} \left(S(x_0)_{i,j} \cdot \mathbf{z}_{i,j} - \tau\right)^2.$$

The threshold τ is an adjustable parameter defining the obfuscation range: larger values τ promote a higher degree of masking of the salient sections, while lower values retain them. The loss function encourages $\mathbf{z}_{i,j}$ to be minimal in areas of high saliency (to conserve critical regions of identity) and maximal where saliency is low (to boost privacy through perturbation).

2.3 Block-wise Obfuscation and Permutation

To boost diversity and introduce non-linearity in the transformation process, we implement a block-wise mixing technique combined with permutation. Let x_0 be the original face image, and x_1 be an auxiliary biometric sample (e.g., a face or fingerprint from a different identity). We divide each image into N distinct, non overlapping blocks, labeled as $\{x_0^{(k)}\}_{k=1}^N$ and $\{x_1^{(k)}\}_{k=1}^N$, with corresponding saliency-aware masks $\{\mathbf{z}^{(k)}\}_{k=1}^N$ applied to each block.

To enhance obfuscation, we apply block-level permutation using transport matrices Π_0 and Π_1, which randomly shuffle the blocks in x_0 and x_1, respectively. The final augmented image is computed as:

$$\tilde{x} = \sum_{k=1}^{N} \left[(1 - \mathbf{z}^k) \odot (\Pi_0 x_0)^k + \mathbf{z}^k \odot (\Pi_1 x_1)^k\right].$$

This strategy increases spatial variability and conceals the spatial arrangement of features, making it significantly harder for inversion attacks to reconstruct the original structures.

2.4 Training Objective

In our federated learning paradigm, the global model is trained with perturbed images as described above. Each client aims to reduce classification loss for the masked inputs while also refining the perturbation parameters. The objective is formulated as:

$$\min_{\theta,\,\{\mathbf{z}^{(k)}\},\,\Pi_0,\,\Pi_1} \mathbb{E}_{(x_0,y_0),(x_1,y_1)\sim\mathcal{D}}\,\mathbb{E}_{\lambda\sim q(\lambda)}\left[\ell\big(\tilde{x},(1-\lambda)y_0+\lambda y_1;\theta\big)\right],$$

Here, $\mathcal{D}$ denotes the data distribution over input-label pairs (x, y), $\ell(\cdot)$ is the classification loss (e.g., cross-entropy), θ represents the global model parameters, $\mathbf{z}^{(k)}$ are the local saliency masks used by client k, and Π_0, Π_1 are the spatial block permutations applied to inputs x_0 and x_1, respectively. The mixing coefficient λ is sampled from a symmetric Beta distribution $q(\lambda) = \mathrm{Beta}(\alpha, \alpha)$.

In order to enhance defense against gradient based inversion attacks, we apply a jigsaw puzzling permutation to image segments. This method generates spatial disruption, increasing the gradient gap between $\tilde{x}$ and the original x_0, consequently decreasing the comprehensibility of any reconstructed images.

2.5 Privacy Impact

Within the framework of federated learning, our augmentation method injects uncertainty into the gradient space. By introducing perturbed inputs $\tilde{x}$, we produce gradients that diverge significantly from those calculated with the original inputs x_0, thus reducing the success of gradient inversion attacks. We utilize three metrics: the Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018), which evaluates perceptual similarity through deep feature activations from pre-trained neural networks (where lower scores signify greater perceptual similarity); the Mean Squared Error (MSE), a conventional metric for pixel-wise differences (where lower values are preferable); and the Peak Signal-to-Noise Ratio (PSNR) Keleş et al. (2021), which measures the ratio of signal to noise in pixel intensities (where higher scores indicate superior visual fidelity). These metrics are widely used in the evaluation of gradient inversion attacks, including Inverting Gradients (IG) Geiping et al. (2020), Fishing Wen et al. (2022), and GIAS Jeon et al. (2021), and therefore provide a consistent basis to compare the effectiveness of our defense. Our augmentation process is deliberately irreversible, applied in shuffled spatial segments, and varies between samples. This disrupts spatial and semantic consistency, invalidating reconstruction algorithms' assumptions of alignment and locality, thus boosting privacy without compromising the model's authentication accuracy.

3 Methods

Our approach advances biometric authentication within federated learning by integrating perturbation-based privacy methods with the merging of various biometric characteristics. We employ a dual-model system: one model, focused on perturbation, generates saliency-enhanced inputs, while the authentication model ensures dependable identification even when data is obscured. For the primary biometric image x_0, typically a facial photograph, and the secondary biometric x_1, such as a fingerprint from the same individual, we first derive feature representations from each using a shared encoder $f_\theta(\cdot)$. This encoder, possibly based on a ResNet He et al. (2015) or Vision Transformer Chen et al. (2021) framework, transforms the inputs into spatial embeddings $f_0 = f_\theta(x_0)$ and $f_1 = f_\theta(x_1)$, which are subsequently used in fusion processes.

To achieve spatially adaptive perturbation, the input image is divided into a 3×3 grid, following the patch layout from Dosovitskiy et al. (2021), producing 9 spatially organized segments for each image. For every index i where $1 \le i \le 9$, each pair of spatial blocks $(x_0^{(i)}, x_1^{(i)})$, from the face and fingerprint modalities, is processed by a compact encoder g_ϕ, resulting in semantic embeddings $e_0^{(i)}$ and $e_1^{(i)}$. The cosine similarity $s^{(i)} = \mathrm{cosine}(e_0^{(i)}, e_1^{(i)})$ quantifies the semantic alignment between the modalities within that block. Only blocks corresponding to the same spatial location i are compared and combined; no mixing or substitution occurs between different locations. This restriction ensures spatial coherence and avoids semantic misalignment, which could impair the clarity and functionality of the combined representation.

A threshold τ is learned to decide whether to keep the face block or substitute it with the fingerprint block. If $s^{(i)} < \tau$, indicating unreliable or sensitive information in the face block, it is replaced; otherwise, it is

preserved. Here, unreliable or sensitive blocks are those with low semantic alignment between face and fingerprint embeddings, implying either privacy risks from identifiable facial features or diminished quality due to noise, justifying replacement with fingerprint data for improved privacy and reliability.

This strict masking strategy may be softened to a learnable mix, where each block is combined using a weight $\beta^{(i)} = \sigma\big(W[e_0^{(i)}; e_1^{(i)}]\big)$. Here, the matrix W serves as a learnable weight of a linear layer that converts the concatenated embeddings $[e_0^{(i)}; e_1^{(i)}]$ into a scalar, which is then passed through a sigmoid function σ, creating a blending coefficient between 0 and 1.

The combined block is calculated as $\tilde{x}^{(i)} = (1-\beta^{(i)}) \cdot x_0^{(i)} + \beta^{(i)} \cdot x_1^{(i)}$. The weight $\beta^{(i)}$ is derived by applying a sigmoid function to a trained linear transformation of the concatenated embeddings $[e_0^{(i)}; e_1^{(i)}]$, allowing the model to dynamically adjust the blend ratio for each block according to their semantic similarity. This flexible blending approach offers subtle control over privacy preservation and preservation of informative content, facilitating a gradual blend rather than a fixed choice. This method allows the model to selectively protect or reveal information with spatial awareness.

Our authentication system utilizes a Vision Transformer (ViT) structure Chen et al. (2021), specifically the Jigsaw-ViT Chen et al. (2023) variant, which is adapted to tolerate spatial interruptions resulting from our augmentation techniques. Throughout the training phase, the model is exposed to both original and obfuscated (augmented) images, allowing it to authenticate accurately even with spatially altered or perturbed inputs. The data sets used for training include CASIA-WebFace, CelebA, FaceScrub, and the FVC2004 fingerprint database, all of which are elaborated on in subsequent sections. For evaluation, the model is tested exclusively on obfuscated images to confirm its robustness in handling privacy-preserving transformations.

The Jigsaw-ViT operates by segmenting each input image $\tilde{x}$ into several non-overlapping blocks, embedding these blocks, and then passing them through transformer layers that utilize multi-head self-attention. During training, a jigsaw-style loss prompts the model to leverage global context instead of exact pixel patterns, allowing it to perform well in biometric authentication even with spatially shuffled inputs. The ultimate identity prediction is determined by the transformer's class token output.

The training approach aims to simultaneously refine the perturbation mechanism and the authentication model by minimizing an integrated loss function composed of two parts: a cross-entropy loss for ensuring precise identity classification and a regularization term that influences the perturbation method. Formally, the objective can be expressed as $\min_{\theta,\phi,W} \mathbb{E}_{(x_0,y_0),(x_1,y_1)} \Big[\ell(f_{\text{auth}}(\tilde{x}), y_0) + \lambda \sum_{i=1}^{9} (s^{(i)} - \tau)^2\Big]$. Here, ℓ represents the cross-entropy loss between the predicted outcome and the actual label y_0, with λ adjusting the regularization term's weight. The regularization component $\sum_{i=1}^{9} (s^{(i)} - \tau)^2$ discourages similarities between blocks $s^{(i)}$ from aligning with the threshold τ, promoting the model to confidently decide on maintaining or altering each block. This dual-purpose objective ensures robust recognition while promoting privacy-conscious perturbation patterns.

In order to succinctly explain our privacy-preservation framework, Algorithm 1 is introduced, which outlines the complete training process. This begins at the client level, where each biometric image is divided into spatial segments and processed through a saliency-based obfuscation technique. Each segment is evaluated with a binary decision, retain the face content as is or replace it, guided by semantic equivalence with a different biometric modality, like a fingerprint. The obfuscated image may then optionally undergo block permutation for added randomization prior to being used in federated training. On the server side, the authentication model is adapted to handle these altered inputs, incorporating a regularization loss that ensures effective feature masking. This algorithm illustrates the modular nature of our system, emphasizing the integration of block-level augmentation, biometric incorporation, and federated learning to ensure both secure authentication and improved privacy.

4 Model Architecture and Experimental Setup

4.1 Model Architecture

The proposed framework for federated biometric authentication consists of two integrated components, shown in Figures 9 and 10 (see Appendix A). The local component, located on client devices, handles raw biometric

Algorithm 1 Privacy-Preserving Federated Biometric Authentication

Require: Biometric images x_0 (face), x_1 (fingerprint), labels y_0, y_1
Ensure: Obfuscated input $\tilde{x}$ and trained authentication model f_{auth}

Client-Side: Local Obfuscation
Compute saliency map $S(x_0)$
for each spatial block $x_0^{(i)}$, $x_1^{(i)}$ in a 3×3 grid **do**
    Extract embeddings $e_0^{(i)} = g_\phi(x_0^{(i)})$, $e_1^{(i)} = g_\phi(x_1^{(i)})$
    Compute semantic similarity $s^{(i)} = \text{cosine}(e_0^{(i)}, e_1^{(i)})$
    if $s^{(i)} < \tau$ **then**
        Replace: $\tilde{x}^{(i)} \leftarrow x_1^{(i)}$ ▷ Sensitive region obfuscated
    else
        Retain: $\tilde{x}^{(i)} \leftarrow x_0^{(i)}$
    end if
end for
Concatenate blocks: $\tilde{x} = \bigcup_{i=1}^{9} \tilde{x}^{(i)}$
Optional: Randomize blocks with transport matrices Π_0, Π_1
$\tilde{x} \leftarrow \sum_{i=1}^{N} [(1 - \mathbf{z}^{(i)}) \odot (\Pi_0 x_0)^{(i)} + \mathbf{z}^{(i)} \odot (\Pi_1 x_1)^{(i)}]$
Server-Side: Authentication Model Training
Compute class prediction: $\hat{y} \leftarrow f_{\text{auth}}(\tilde{x})$
Compute loss:

$$\mathcal{L} = \ell(\hat{y}, y_0) + \lambda \sum_{i=1}^{9} (s^{(i)} - \tau)^2$$

Update model parameters θ, obfuscation module ϕ, and blending weights W using federated averaging
Repeat across clients and rounds until convergence

data like facial images or fingerprints. It applies particular methods to obscure sensitive information while preserving essential identity features. This strategy guarantees that raw biometric data stays on the user's device, significantly enhancing privacy. In contrast, the global module functions on the server side by collecting gradient updates from various client devices. It uses a unique data set that includes both original and obfuscated biometric images, allowing the global model to learn features and their relations. The focus on these modified images helps to assess the significance of features, enabling accurate authentication purely with these privacy-protected data during evaluation. This dual-level strategy successfully harmonizes privacy with authentication accuracy in federated learning settings.

In our assessment of approaches aimed at protecting privacy in identity recognition, we implemented and examined three core architectures for biometric authentication (i.e. Global model): ResNet18 He et al. (2015), Vision Transformer (ViT) Chen et al. (2021); Dosovitskiy et al. (2021), and Jigsaw ViT Chen et al. (2023).

ResNet18 serves as a traditional convolutional model, consisting of 17 convolutional layers with residual connections. These connections facilitate the learning of more complex representations. The final fully connected layer offers class probabilities for predicting identities.

ViT (Vision Transformer) utilizes an attention mechanism that focuses on image patches to grasp the global interactions across various parts of the image. The image is divided into fixed-size patches, which are subsequently embedded into a transformer encoder. Inside this encoder, self-attention layers facilitate the depiction of contextual relationships. This model is particularly sensitive to structured obfuscations and patch-level augmentations.

Jigsaw ViT is a modified Vision Transformer aimed at improving accuracy in machine learning. It includes a jigsaw mechanism that randomly rearranges patch positions during both the training and testing stages.

This disrupts spatial continuity while maintaining the semantic meaning, thus protecting against gradient inversion attacks while still allowing for identity discrimination Chen et al. (2023).

Recently, DNNs have seen extensive application in face recognition tasks. The work of Schroff et al. (2015) employed a Triplet loss to enhance performance on challenging facial recognition datasets. Subsequently, Liu et al. (2016) proposed an enhancement to the traditional Softmax loss, known as L-Softmax loss. It encourages intra-class compactness and separation between classes in the learned embedding features. Liu et al. (2017) introduced the idea of combining an angular margin loss with the standard Softmax function. The approach of Wang et al. (2018), which involved replacing Softmax with a Cosine margin loss in relation to the target logit, resulted in superior performance compared to prior methods. ArcFace Deng et al. (2019) utilized a geodesic distance on a hypersphere, leading to better discriminative capabilities and a more stable training process. Additionally, Kim et al. (2022) argued that optimizing ArcFace based on each sample's quality, as judged by the embedding norm, results in considerable improvements. Their newly proposed loss achieved state-of-the-art outcomes on both challenging low-quality and high-quality datasets Kim et al. (2022). We implemented all three model architectures with backbone of Arcface for increasing accuracy in authentication task.

4.2 Datasets

We tested our models using three commonly employed face recognition datasets:
FaceScrub: This dataset, accessible to the public, contains over 100,000 aligned and cropped images from 500 celebrity identities. It offers moderate class variability and is widely used in privacy research Harvey (2021).
CelebA: Featuring more than 200,000 images across 10,177 unique identities, this dataset includes extensive attribute labels and exhibits a wide array of facial expressions, poses, and lighting conditions Liu et al. (2015).
CASIA-WebFace: With over 490,000 images of 10,575 individuals, this set presents notable variability and challenging intra-class differences, making it ideal for testing model robustness Cao et al. (2018). Utilizing the FVC2004 Fingerprint Verification Competition (2004) database, consisting of grayscale fingerprint images obtained under various conditions, we gathered fingerprint data. In the preprocessing phase for multi-biometric fusion, these fingerprint images were resized and combined with the low-saliency parts of facial images.

4.3 Training and Evaluation Protocol

In order to train models, multi-class identity classification was conducted using cross-entropy loss. Within federated learning frameworks, the data was distributed across 20 clients, each carrying out local training on their allocated data segment through several communication rounds.

The model's training hyperparameters were chosen via empirical tuning to optimize between performance and resource utilization. Initially, we used parameters frequently applied in the latest face authentication systems and adjusted them through preliminary trials to assess authentication accuracy and convergence stability within our perturbation pipeline. Pretrained weights from AdaFace Kim et al. (2022) were employed to initialize our models, as they offer a robust foundation for identity classification tasks. We followed their recommended training configuration as a baseline, modifying batch size, learning rate, and training time to integrate the increased complexity presented by our jigsaw and perturbation modules.

Training specifics: We utilized the Adam optimizer with a learning rate of 1×10^{-4}, and a batch size of 64. The training was conducted over 50 epochs, which corresponds to approximately 500 communication rounds in the federated learning setup. To enhance generalization and robustness, we applied a variety of data augmentations, including random cropping, flipping, random erasing, Gaussian noise, and mixup.

A batch refers to a subset of the training dataset that is handled simultaneously during a single forward/backward pass in model training. Utilizing batches aids in balancing both computational efficiency and the stability of convergence.

After conducting tuning experiments aimed at optimizing accuracy, we set the batch size to 64 to ensure both efficient training duration and manageable GPU memory consumption. Comparable batch sizes have been used in other face recognition studies, like Kim et al. (2022), from which we also utilized pretrained parameters to enhance initialization.

In privacy-preserving experiments involving random block swapping and biometric fusion, the augmentations helped improve model robustness and generalization by exposing it to diverse variations of the input data. The test set used for identity classification evaluation comprised approximately 20% of the full dataset. All models were implemented using PyTorch and trained on GPUs from Compute Canada. To ensure reproducibility, experiments were repeated across three different random seeds.

4.4 Metrics

We evaluated model performance and privacy protection using the following metrics:
Accuracy (Acc) — measures the correctness rate of identity authentication.
PSNR (Peak Signal-to-Noise Ratio), **LPIPS (Learned Perceptual Image Patch Similarity)**, **MSE (Mean Squared Error)**, and **Cosine Similarity** — assess the similarity between original and reconstructed images in gradient inversion attacks, following prior work such as Zhu et al. (2019); Yin et al. (2021). Improved privacy corresponds to lower PSNR and cosine similarity values, and higher LPIPS and MSE values.

4.5 Data Augmentation Techniques

Our model training leveraged multiple augmentation methods to enhance feature invariance and robustness as shown in figure 1(see A:
Mixup, Generates convex combinations of pairs of training samples and labels, encouraging the model to behave linearly between training examples Pang et al. (2020). **CutMix**, Combines patches from different images, forcing the model to rely on local discriminative features Yun et al. (2019). **PuzzleMix**, Applies a saliency-guided mixing of image blocks, promoting robustness to spatial perturbations Kim et al. (2020). **Random Block Swapping**, Randomly rearranges blocks within images to obscure spatial correlations, enhancing resistance to inversion attacks.

Diverse augmentations are utilized on both the pixel and feature levels to enhance input variability, assisting the model in learning more generalizable features. Within our training framework, we have implemented various methods like Mixup, CutMix, PuzzleMix, and Random Block Swapping to expand the dataset and mimic realistic distortions and occlusions. These enhancements are applied separately or in a combined scheduling throughout the training batches to strike a balance between regularization and specific learning for the task. By embedding these augmentation techniques into the training process, the model is introduced to a broad spectrum of spatial and semantic variations, which ultimately boosts its robustness and generalization capabilities.

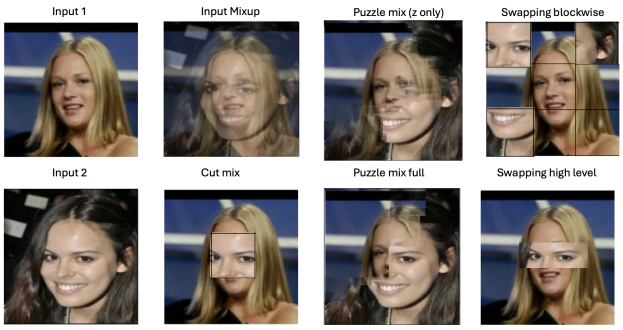

Figure 2: Visualization of the data augmentation techniques used in our experiments. First column : Input 1 and Input 2 (original samples), Input Mixup (pixel-wise interpolation), Puzzle Mix (just pixel level mixing), and Swapping Blockwise (spatial patch swapping). Bottom row: CutMix (region replacement), Puzzle Mix Full (both pixel and feature level mixing), and Swapping High Level (semantic region exchange). These augmentations aim to increase the diversity of training samples and provide regularization against overfitting and privacy attacks.

We offer further evidence validating the efficacy of our proposed privacy-preserving strategies through a qualitative analysis of images reconstructed using gradient inversion attacks. Figure 3 displays attempts

to reconstruct augmented single-biometric (face) images, in which our obfuscation techniques significantly impair the clarity and recognizability of the images. Figure 4 displays attempts to reconstruct multi-biometric (face, fingerprint) perturbed images. we can see that our obfuscation techniques significantly impair the clarity and recognizability of the images. The optimization attack can only reconstruct the input and not the accurate prediction results. Within the jigsawVIT configuration, each image is initially jumbled, as illustrated in figure 2, and this jumbled formation is not integrated into the primary model. Consequently, for jigsawVIT, the attack reconstructs merely the mixed-up image rather than the images in their correct sequence, since arranging them correctly is part of the solution.

Figure 3: Visualization of gradient inversion attack reconstructions from augmented face images. Each column corresponds to a different data augmentation method applied prior to the attack. First column, Original face images without augmentation—reconstructions are visibly sharp and identifiable. Second column (top and bottom), Mixup (pixel-wise interpolation) and block-level swapping (between two images of the same person)—show moderate privacy gain via reduced visual clarity. Third column (top and bottom), CutMix (region replacement) and PuzzleMix (feature level mixing). Fourth column, important feature window swapping, produces highly distorted reconstructions, offering the best resistance to identity inference. These results illustrate how augmentation methods, especially spatially and semantically aware ones, effectively degrade inversion quality and enhance privacy protection.

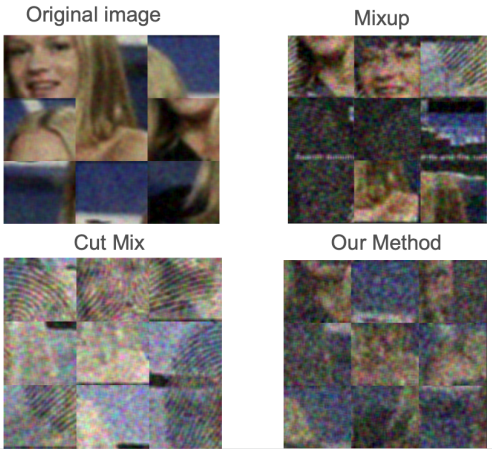

Figure 4: Comparison of augmentation methods in biometric fusion for privacy protection. Top-left: Original image is divided into 9 blocks (3x3). Top-right: Mixup interpolation introduces pixel-level blending of face and fingerprint blocks. Bottom-left: CutMix directly replaces regions with fingerprint patches. Bottom-right: Our proposed method uses saliency-aware block replacement, selectively inserting fingerprint patches only in low-saliency (less important) regions.

To clearly demonstrate our proposed augmentation and authentication workflow, we include a detailed diagram showcasing the entire procedure, accompanied by visual examples of various augmentation techniques in Appendix A.

5 Experiments

Through our experiments, we verify the effectiveness of our saliency-aware obfuscation method and its robustness against gradient inversion attacks. We assess how various augmentation techniques, multi-biometric fusion, and targeted masking affect authentication accuracy and privacy. Additionally, we introduce perceptual and reconstruction-focused evaluation metrics, including PSNR, LPIPS, MSE, and cosine similarity, to quantify privacy protection levels. To begin, we assessed the importance of different facial regions by methodically obscuring areas with high and low salience in a 3×3 grid pattern. Blocking low-salience parts like the forehead and cheeks led to a minimal impact on accuracy, whereas covering high-salience regions such as the eyes, nose, and mouth significantly decreased authentication performance. This evidence supports our design choice of preserving high-salience features while hiding low-salience areas to maintain accuracy and ensure privacy.

In order to further investigate spatial robustness, we executed experiments with both random and controlled block-swapping. When low-important (such as background textures or skin regions) semantic regions were interchanged between two pictures of the same person, the model's performance was hardly affected, indicating that the model generalizes beyond redundant textures. However, exchanging highly important semantic regions (such as the eyes, nose, or mouth) between images of different individuals caused a significant drop in performance, highlighting their importance for identity representation.

Considering these findings, we devised a multi-biometric fusion approach that replaces less significant facial regions with fingerprint sections. This unified representation is crafted via the following masking procedure:

$$h(x_0, x_1, \mathbf{z}) = (1 - \mathbf{z}) \odot x_0 + \mathbf{z} \odot x_1,$$

where x_0 represents the facial image, x_1 denotes the fingerprint image, and $\mathbf{z}$ signifies a learned binary mask. This fusion strategy preserves areas vital to identity while embedding biometric noise in less important zones, complicating the reconstruction of adversarial features.

We assess the privacy resilience of our approach against gradient inversion attacks, using the techniques outlined in Zhu et al. (2019); Yin et al. (2021). To evaluate the reconstructed images, we utilize four metrics widely used in previous studies for image recovery assessment: PSNR (Peak Signal-to-Noise Ratio), LPIPS (Learned Perceptual Image Patch Similarity), MSE (Mean Squared Error), and cosine similarity. These metrics evaluate both the pixel-level precision and the perceptual likeness of the images. More precisely, low PSNR and cosine similarity values together with high LPIPS and MSE values suggest a decline in reconstruction quality, thus indicating enhanced privacy protection. These same metrics were also applied in Geiping et al. (2020); Zhao et al. (2020); Bai et al. (2024) to evaluate the efficiency of inversion attacks and related defensive strategies, ensuring both fair and consistent comparisons.

Table 1: Privacy metrics for different augmentation and fusion strategies. Higher LPIPS and MSE, and lower cosine similarity indicate stronger privacy protection.

Method	**PSNR ↓**	**LPIPS ↑**	**MSE ↑**	**Cosine Sim. ↓**
Original (No Mask)	34.2	0.141	0.0018	0.98
Random Block Swapping	31.8	0.157	0.0023	0.79
CutMix	28.7	0.194	0.0035	0.82
Mixup	30.2	0.200	0.0030	0.73
PuzzleMix	30.8	0.159	0.0032	0.80
Biometric Fusion (Ours)	**26.3**	**0.245**	**0.0048**	**0.68**

The result table 1 shows that our saliency-focused augmentation and fusion method achieves a beneficial compromise between authentication accuracy and resistance to gradient inversion attacks. By integrating

fingerprint data into low-saliency regions, the complexity of reconstruction is significantly increased, offering an effective defense mechanism for federated identity recognition systems compared to just PuzzleMix on face biometric.

Moreover, we evaluate the accuracy of authentication methods across different techniques to ensure that improvements in privacy do not negatively impact performance.

Table 2: Authentication accuracy across different augmentation methods.

Method	**Authentication Accuracy (%)**
Original (No Mask)	98.5
Random Block Swapping	83.2
CutMix	50.7
Mixup	85.3 Pang et al. (2020)
PuzzleMix	78.7
Biometric Fusion (Ours)	**89.8**

Our approach to biometric fusion displays enhanced accuracy in authentication and privacy improvements, outperforming other augmentation methods in maintaining this balance.

6 Conclusion

Our experimental findings indicate that strategically blurring facial areas based on their importance significantly enhances privacy while only slightly impacting identity recognition effectiveness. Through comprehensive testing, we discovered that hiding low-importance regions (such as the forehead and cheeks) had negligible effects on authentication accuracy, suggesting these areas provide redundant or less vital information for identity verification. Conversely, preserving high-importance regions (like the eyes, nose, and mouth) was crucial for maintaining accuracy. This targeted obfuscation approach achieved an optimal balance, allowing the model to offer robust privacy protection with minimal compromise in authentication accuracy.

Our framework's resilience was enhanced by integrating multi-biometric fusion, where fingerprint data is embedded in non-critical regions of facial information. This method increased the complexity for adversaries using gradient-based reconstruction and expanded the variety of biometric characteristics available for authentication. Consequently, our fusion technique made reconstruction more difficult, as evidenced by significant declines in reconstruction quality measures like PSNR and cosine similarity, accompanied by increases in distortion measures such as LPIPS and MSE.

The customized Jigsaw ViT model, which includes random patch swapping in its training process, showed excellent performance in undermining the spatial consistency that attackers rely on for successful gradient inversion. Unlike traditional ViT or ResNet models, the Jigsaw ViT maintained high performance even under considerable spatial confusion. This underscores its suitability for federated learning environments, where the danger of gradient leakage is a significant security concern. Our method reliably outperformed existing privacy-preserving techniques such as random erasing, Gaussian noise addition, and conventional mixup approaches.

References

Li Bai, Haibo Hu, Qingqing Ye, Haoyang Li, Leixia Wang, and Jianliang Xu. Membership inference attacks and defenses in federated learning: A survey. *ACM Computing Surveys*, 57(4):1–35, 2024.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

Si Chen, Feiyang Kang, Nikhil Abhyankar, Ming Jin, and Ruoxi Jia. Data-centric defense: Shaping loss landscape with augmentations to counter model inversion. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=r8wXaLJBIS.

Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023.

Jung Hee Cheon, Andrey Kim, Miran Kim, and Yongsoo Song. Homomorphic encryption for arithmetic of approximate numbers. In *Advances in cryptology–ASIACRYPT 2017: 23rd international conference on the theory and applications of cryptology and information security, Hong kong, China, December 3-7, 2017, proceedings, part i 23*, pp. 409–437. Springer, 2017.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Dimitar I Dimitrov, Maximilian Baader, Mark Müller, and Martin Vechev. Spear: Exact gradient inversion of batches in federated learning. *Advances in Neural Information Processing Systems*, 37:106768–106799, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Fingerprint Verification Competition. Fingerprint verification competition 2004 (fvc2004). http://bias.csr.unibo.it/fvc2004/, 2004. Accessed: 2025-06-03.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33: 16937–16947, 2020.

Google Cloud Skills Boost. Google cloud skills boost: Machine learning with tensorflow on google cloud - official documentation, 2024. URL https://www.cloudskillsboost.google/paths/17/course_templates/1036/video/513292. Accessed: 2025-05-13.

Adam Harvey. Exposing.ai, 2021. URL https://exposing.ai.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instance-hiding schemes for private distributed learning. In *International conference on machine learning*, pp. 4507–4518. PMLR, 2020.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34:7232–7241, 2021.

Jinwoo Jeon, Kangwook Lee, Sewoong Oh, Jungseul Ok, et al. Gradient inversion with generative image prior. *Advances in neural information processing systems*, 34:29898–29908, 2021.

Onur Keleş, M Akın Yılmaz, A Murat Tekalp, Cansu Korkmaz, and Zafer Doğan. On the computation of psnr for a set of images or video. In *2021 Picture Coding Symposium (PCS)*, pp. 1–5. IEEE, 2021.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International conference on machine learning*, pp. 5275–5285. PMLR, 2020.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

Shiao Liu, Yunfei Yang, Jian Huang, Yuling Jiao, and Yang Wang. Zkfl: Zero-knowledge proofs for federated learning. In *Advances in Neural Information Processing Systems*, 2023. URL `https://proceedings.neurips.cc/paper/2023/hash/66be31e4c40d676991f2405aaecc6934-Abstract.html`.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Minghui Ma, Jiali Deng, Meiyi Yang, Xuan Cheng, Nianbo Liu, Ming Liu, and Xiaomin Wang. Cost ensemble with gradient selecting for gans. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 1194–1200, 2022. URL `https://www.ijcai.org/proceedings/2022/167`.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks, 2020.

Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Kristian Schwethelm, Johannes Kaiser, Moritz Knolle, Sarah Lockfisch, Daniel Rueckert, and Alexander Ziller. Visual privacy auditing with diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=D3DA7pgpvn`.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Provable defense against privacy leakage in federated learning from representation perspective. *arXiv preprint arXiv:2012.06043*, 2020.

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00481. URL https://arxiv.org/abs/1801.07698.

Md Amir Uddin, Md Rifat Hassan, Jihwan Park, and Sangyoun Lee. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11177–11186, 2022. URL https://arxiv.org/abs/2006.01791.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

Yuxin Wen, Jonas Geiping, Liam Fowl, Micah Goldblum, and Tom Goldstein. Fishing for user data in large-batch federated learning via gradient magnification. *arXiv preprint arXiv:2202.00580*, 2022.

Wu, Ruihan, Chen, Xiangyu, Guo, Chuan, Weinberger, and Kilian Q. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Uncertainty in Artificial Intelligence*, pp. 2293–2303. PMLR, 2023.

Ceyuan Yang, Yujun Shen, Yinghao Xu, Dingdong Zhao, Bo Dai, and Bolei Zhou. Improving gans with a dynamic discriminator. In *Advances in Neural Information Processing Systems*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6174c67b136621f3f2e4a6b1d3286f6b-Abstract-Conference.html.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16337–16346, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019. URL https://arxiv.org/abs/1905.04899.

Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Appendix

To further support our findings and demonstrate how augmentations can preserve user privacy while impacting model performance, we offer additional figures. These figures analyze the effects of occluding both low-level and high-level features and assess semantic consistency via feature swapping.

Figure 5 presents several augmentations implemented during the training and testing stages using images from the CelebA dataset. The left column showcases examples of three key methods: random square masking, eye masking, and the implementation of synthetic noise around the eye area. These augmentations include removing low-level details (random black squares), applying high-level semantic masking (eye coverage), and adding perceptually structured perturbations (colored noise). The middle and right columns show graphs indicating the effect of each augmentation on face authentication accuracy at various similarity thresholds.

Our analysis shows that implementing augmentations during the training phase helps the model retain decent recognition skills, albeit with some reduction in performance. Conversely, applying augmentations only at the testing stage (illustrated in the right plots) leads to a more substantial drop in performance, particularly with noise-based methods and semantic eye masking. This contrast suggests that consistent augmentation at both training and testing phases is vital for maintaining effectiveness. Additionally, out of the techniques assessed, high-level semantic masking (such as covering the eyes) leads to a more stable and understandable performance trajectory, whereas random square masking causes more irregular results. This underscores that strategic, region-specific masking techniques offer better privacy protection than random occlusion.

Figure 6 explores the importance of semantic attributes by analyzing the impact of swapping facial features either within the same person or with different individuals. The plots in the top left emphasize the results of exchanging eye regions across different individuals. Utilizing this method on both training and testing datasets results in a significant decline in accuracy, highlighting the essential role the eye region plays in identifying individuals.

Interestingly, swapping eyes with the same individual results in minimal disruption, suggesting that despite changes, structural consistency maintains model predictions. The right side of the figure explores the outcomes of replacing random image patches, both within the same individual and across different identities. While exchanging patches between different individuals causes a moderate reduction in performance, doing so within the same individual (bottom right) leads to a noticeable and inconsistent decrease in accuracy. These findings reinforce the idea that disrupting semantic coherence, either through component misalignment or arbitrary alterations, significantly impairs the model's capacity to recognize identities.

Together, these experiments emphasize the contrast between altering low-level and high-level characteristics. High-level, semantic augmentations like eye masking or identity swapping prove more effective in preserving privacy and impairing model inference than basic, unstructured techniques such as random blocking. These visual and numerical results not only validate our proposed augmentation strategy but also reveal which facial traits are crucial for face authentication systems. The consistent decline in performance with specific augmentations highlights the effectiveness of our approach in protecting biometric privacy while maintaining a fair trade-off with model utility.

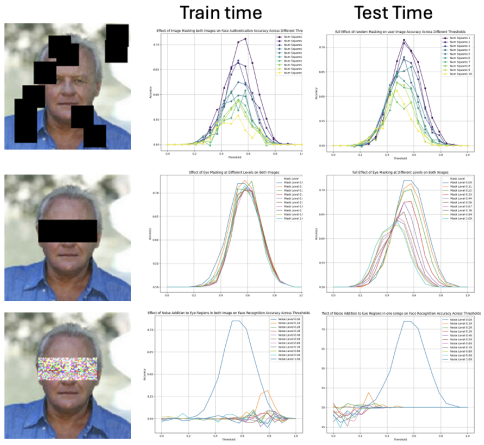

Figure 5: Effect of various augmentations (random square occlusion, eye masking, and eye noise) on recognition performance at train and test time.

To better grasp how different data augmentation techniques affect security and reconstruction quality across various architectures (e.g., ResNet, ViT, and Jigsaw ViT), we present a heatmap visualization in Figure 7.

Each cell in heatmap 7 displays the normalized measurement of metrics such as Security (Distance), PSNR (dB), LPIPS, and MSE for particular block configurations. Enhanced performance is indicated by higher security and PSNR scores along with reduced LPIPS and MSE values. This visualization serves as a valuable tool for illustrating the trade-offs between improving security and preserving visual fidelity.

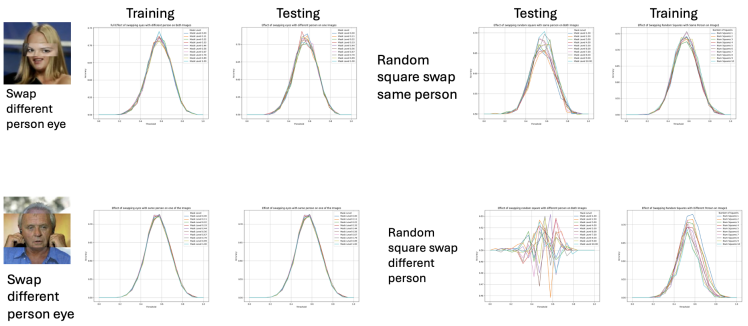

Figure 6: Impact of semantic feature swapping (eyes and patches) across same and different identities.

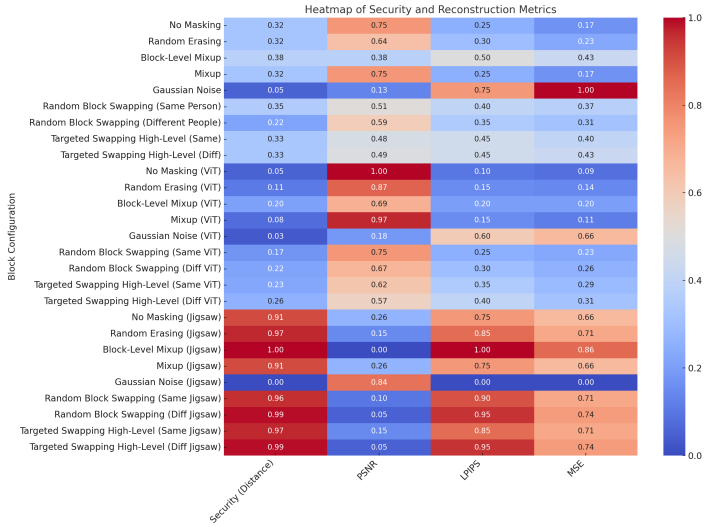

Figure 7: Normalized heatmap of Security(Distance of reconstructed image from original image) and Reconstruction Quality Metrics across different augmentation strategies and model types.

Table 3: Summary of Privacy-Preserving Techniques in Distributed Learning

Technique	**Strengths**	**Weaknesses**
Gradient Perturbation Zhu et al. (2019); Huang et al. (2021); Yang et al. (2022)	Simple to implement; reduces direct data leakage	Requires heavy noise or gradient truncation; lowers accuracy
Differential Privacy Bonawitz et al. (2016); McMahan et al. (2017)	Formal privacy guarantees; widely studied	Adds noise impacting model performance, especially in biometrics
Homomorphic Encryption Cheon et al. (2017)	Allows computations on encrypted data; strong data confidentiality	High computational and communication overhead; latency issues
InstaHide Huang et al. (2020); Carlini et al. (2021)	Obfuscates raw inputs; easy to apply pre-training	Vulnerable to adaptive reconstruction attacks; limited FL suitability
Mixup, PuzzleMix Zhang (2017); Kim et al. (2020)	Improves robustness and generalization	Not designed to prevent gradient leakage; may remove key biometric features

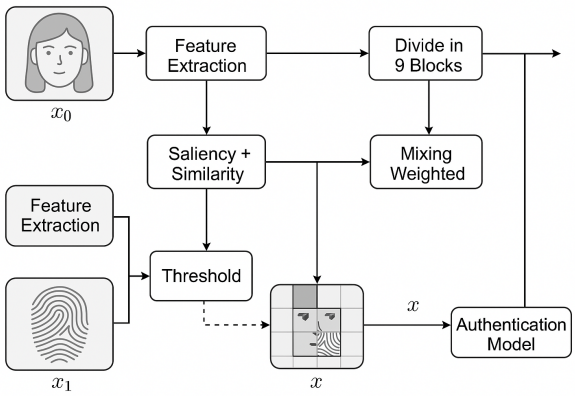

Figure 8: Detailed pipeline illustrating feature extraction, saliency and similarity assessment, block-wise division, and weighted mixing for biometric fusion. The output is processed by the authentication model.

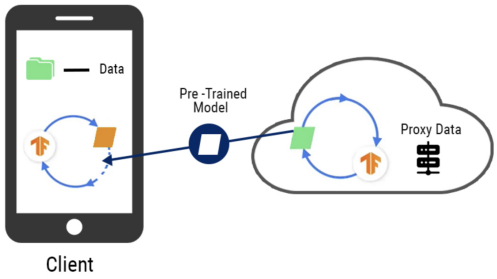

Figure 9: Local model architecture on the client device. The local model directly processes raw biometric data and applies perturbations to obscure sensitive features while preserving the critical identity information required for authentication. Google Cloud Skills Boost (2024)

For enhanced clarity and comprehension of our methodology, we have added additional visual aids illustrating our suggested pipeline along with an array of augmentation techniques.

Figure 8 illustrates the sequential phases of our suggested biometric authentication system. First, the input biometric data, comprising the face (x_0) and fingerprint (x_1), undergo a feature extraction process to identify unique and identity-specific attributes. The resulting features are divided into nine spatial blocks and combined using a saliency-focused, weighted blending method with established thresholds. This generates a fused biometric profile used for authentication.

In Figure 11, we display a range of augmentation techniques applied to biometric data. This includes traditional methods such as Mixup, CutMix, and Puzzle Mix, alongside our novel block-level fusion technique. Unlike standard methods, our approach integrates segments of fingerprints into facial images, striking a balance between enhanced privacy and maintained recognition accuracy.

A.1 Limitations and Discussion

While our proposed framework demonstrates promising results in preserving authentication accuracy and enhancing privacy through feature-level obfuscation, some limitations remain. Our study largely relied on specific datasets, which may not fully represent the diverse variations present in real-world scenarios. Therefore, further assessment is necessary to ascertain if our approach can be successfully applied to datasets with differing characteristics. Although the framework is designed to be efficient, employing localized block-

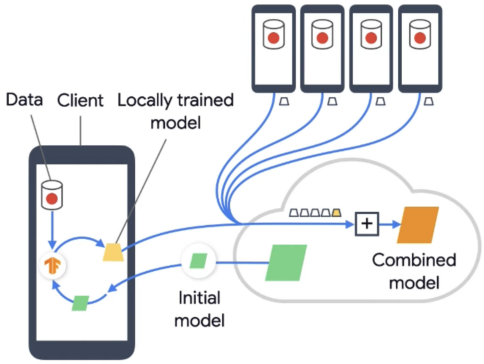

Figure 10: Global model architecture on the server. The global model aggregates gradients from client devices and leverages a separate dataset composed of original and perturbed biometric images. It learns to authenticate users based solely on perturbed biometric inputs by understanding feature importance from training, thereby enhancing privacy during authentication Google Cloud Skills Boost (2024).

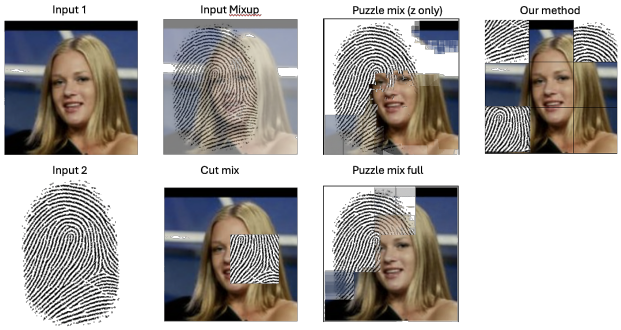

Figure 11: Examples of different augmentation strategies applied to biometric data: Mixup, CutMix, Puzzle Mix (early and full), and our block-based fusion method. Our method strategically combines fingerprint and facial regions, balancing security and authentication accuracy.

wise transformations introduces additional computational demands on client devices. This added burden may be significant for devices with limited processing capabilities, possibly affecting the user experience.

We have tested our defense mechanism against several recognized gradient-based attacks. However, given the rapid development of adversarial techniques, our approach might not be immune to every type of complex attack, especially those exploiting new vulnerabilities. The framework assumes that client devices are free from malware and secure. Should this assumption be false, the privacy assurances provided by our approach could be jeopardized.

Overcoming these constraints can open new avenues for future research. Expanding evaluations to encompass a wider range of datasets will help assess the method's generalizability. Improving the computational efficiency of transformations might make the framework more suitable for resource-constrained devices. Additionally, deploying adaptive strategies to counteract new adversarial threats will enhance system robustness. Exploring methods to protect the integrity and security of client devices will strengthen privacy assurances. Finally, using a broader set of evaluation metrics will offer deeper insights into the practical implications of our framework.