

GRADIENTS PROTECTION IN FEDERATED LEARNING FOR BIOMETRIC AUTHENTICATION

Anonymous authors

Paper under double-blind review

ABSTRACT

In the context of face recognition models, different facial features contribute unevenly to a model’s ability to correctly identify individuals, making some features more critical and, therefore, more susceptible to attacks. Deep Gradient Leakage (DGL) is a highly effective attack that recovers private training images from gradient vectors, posing significant privacy challenges in distributed learning systems where clients share gradients. Data augmentation, a technique for artificially manipulating the training set by creating modified copies of existing data, plays a crucial role in improving the accuracy of deep learning models.

In this paper, we explore various data augmentation methods to protect original training images, in test time thereby enhancing security in distributed learning systems as well as increasing accuracy during training. Our experiments demonstrate that augmentation methods improve model performance during training on augmented images, and we can use the same methods during testing as perturbation methods to preserve some features of the image and have safety against DGL.

This project has four primary objectives: first, to develop a vision transformer face validation model that trains on distributed devices to ensure privacy; second, to utilize augmentation methods to perturb private images and increase neural network safety; and third, to provide protection against attacks, ensuring that reconstructing attacks cannot extract sensitive information from gradients at any point in the system. and lastly we introduce a new novel perturbation method for a multi biometric authentication, system which offers accuracy for identification and guarantees safety and anonymity of entities.

1 INTRODUCTION

Federated learning (Rodríguez-Barroso et al., 2023) proposes letting participants train on their own data in a distributed fashion and share only model updates —i.e., gradients with the central server. The server aggregates these updates (typically by averaging) to improve a global model and then sends updates to participants. This process runs iterative until the global model converges. Merging information from individual data points into aggregated gradients intuitively preserves privacy to some degree.

While this setting might seem safe at first glance, a few recent works have begun to question the central premise of federated learning - is it possible for gradients to leak private information of the training data? Effectively serving as a “proxy” of the training data, the link between gradients to the data in fact offers potential for retrieving information: from revealing the positional distribution of original data (Melis et al., 2019; Shokri et al., 2017), to even enabling pixel-level detailed image reconstruction from gradients (Geiping et al., 2020; Zhao et al., 2020; Zhu et al., 2019).

Building on the findings in Wei et al. (2022), we draw two key insights: (1) different features contribute unequally to the task, with some being so insignificant that they can be masked without significantly affecting the model’s performance. In their work, classification tasks were completed with high accuracy even with certain masked features, a concept we apply to our authentication task; and (2) There is an inherent trade-off between the model’s security, robustness, and accuracy; it is essential to carefully balance these aspects in designing effective defenses.

Generalizability refers to a model’s ability to maintain consistent performance when applied to new,

054 unseen data (testing data), as compared to its performance on the data it was trained on. Models with
055 poor generalizability tend to overfit the training data. performing well on examples but struggling
056 with new ones.(Shorten & Khoshgoftaar, 2019)

057 Data Augmentation is a very powerful method of achieving generalizability. The augmented data
058 will represent a more comprehensive set of possible data points, thus minimizing the distance be-
059 tween the training and validation set, as well as any future testing sets. Through augmentation
060 methods; more information can be extracted from the original dataset(Krizhevsky et al., 2012). By
061 employing augmentation techniques, we can shift the model’s focus from one part of an image to
062 another. Furthermore, data augmentation can serve as a tool to increase the security of raw data
063 during the training and testing phases. In this work, we examine various augmentation methods and
064 their impact on model performance. Additionally, we explore how these methods can be leveraged
065 to safeguard private information within a federated learning setup during test time.

066 Vision Transformer (ViT) Alexey (2020) is an architecture inherited from Natural Language Pro-
067 cessing (Vaswani, 2017) while applied to image classification with taking raw image patches as
068 inputs. Different from classical Convolutional Neural Networks (CNNs), the architectures of ViTs
069 are based on self-attention modules (Vaswani, 2017), which aim at modeling global interactions of
070 all pixels in feature maps. More precisely, ViTs take sequential image patches as inputs, and the
071 attention mechanism enables interaction and aggregation directly among patch information. There-
072 fore, compared to CNNs where image features are progressively learnt from local to global context
073 via reducing spatial resolution, ViTs enjoy obtaining global information from the very beginning.
074 Up till now, such convolution-free networks have been achieving great success on various computer
075 vision tasks, including image classification (Touvron et al., 2021; Wu et al., 2021; Chen et al., 2021a;
076 Li et al., 2022; Mao et al., 2022; Yao et al., 2023), object detection (Liu et al., 2021; Li et al., 2022;
077 Yao et al., 2023), semantic segmentation (Strudel et al., 2021; Liu et al., 2021; Yao et al., 2023) and
078 image generation (Chen et al., 2021b).

079 In this work, we utilize Vision Transformer (ViT) models Chen et al. (2023) instead of traditional
080 Resnet face recognition models to achieve superior image mapping during training. This approach
081 aims to enhance performance on an unseen dataset that significantly differs from the training set.
082 Distributed training and collaborative learning have been widely used in large scale machine learning
083 tasks. In most scenarios, people assume that gradients are safe to share and will not expose the
084 training data. Some recent studies show that gradients reveal some properties of the training data,
085 for example, property classifier Melis et al. (2019) (whether a sample with certain property is in the
086 batch) and using generative adversarial networks to generate pictures that look similar to the training
087 images (Gentry, 2009; Hitaj et al., 2017; Melis et al., 2019).

088 In the work Zhu et al. (2019), present an optimization algorithm that can obtain both the training
089 inputs and the labels in just few iterations. The method deep leakage is an optimization process
090 and does not depend on any generative models; therefore, DLG Zhu et al. (2019) does not require
091 any other extra prior about the training set, instead, it can infer the label from shared gradients; the
092 results produced by DLG (both images and texts) are the exact original training samples instead
093 of synthetic look-alike alternatives(Yin et al., 2021). We propose a new VIT model with multiple
094 attention windows to perform better on unseen data and seeks to focus and learn different features
095 better. Our proposed augmentation method jointly seek to find (1) the optimal mask for deciding how
096 much of the inputs to reveal versus conceal in the given region and (2) a trade off between security
097 and accuracy of face recognition. Based on the idea of mixing images to learn the individual better
098 (Kim et al., 2020) we propose to mix two biometrics to have better accuracy in authentication by
099 having dynamic weight for each and try to find the optimal mask of the two inputs.

101 2 RELATED WORK

102
103
104 There is a rich literature for defense strategies against gradient attacks in distributed models. Most
105 techniques focus on protecting the image as a whole rather than focusing on the features. Each of
106 these techniques has its own flaws. In this work we study, how separating features helps preserve
107 accuracy in face authentication task, and also protect the original image. In order to guarantee
privacy, it is necessary to introduce randomness to the learning algorithm.

108 Gradient perturbation Zhu et al. (2019); Sun et al. (2020), directly prunes the shared gradients in
 109 federated learning to defeat gradient leakage attacks; However, recent contributions Huang et al.
 110 (2021) found that it is usually required to prune too much gradient information to fully defeat gra-
 111 dient leakage attacks, which will greatly hurt the model accuracy.

112 Input data encryption Huang et al. (2020a;b) encrypts the data and hides private information; How-
 113 ever, current state-of-the-art encryption methods Huang et al. (2020b) can also be evaded by adaptive
 114 attack methods (Carlini et al., 2021). Therefore, an effective defense method which can reliably pro-
 115 tect the privacy of clients while preserving model accuracy is still highly demanded.

116 Zero-knowledge proofs (ZKPs) (Bonawitz et al., 2016; McMahan et al., 2017) are widely recognized
 117 cryptography tools that enable secure and private computations while safeguarding the underlying
 118 data. In essence, ZKPs empower a proof to convince a verifier of a specific fact without revealing
 119 any information beyond that fact itself. By verifying these proofs, the users can ensure the aggreg-
 120 ator’s actions are transparent and verifiable, installing confidence that the aggregation process is
 121 conducted with utmost honesty. ZKP (Geiping et al., 2020) can ensure the server is protected from
 122 aggregating malicious updates and constantly prove the transparency of users and server’s action
 123 but unfortunately it is necessary to share the private raw information with server and trust the infras-
 124 tructure and system as entity. Additionally, generating the proofs can be computationally expensive,
 125 impacting performance due to the time required for processing.

126 Input mixup creates virtual training examples by linearly interpolating two input data and the corre-
 127 sponding one-hot labels (Melis et al., 2019). The method induces models to have smoother decision
 128 boundaries and reduces overfitting to the training data. Manifold mixup extends this concept from
 129 input space to feature space (Verma et al., 2019). Also, Guo et al. (2019) proposed an adaptive
 130 mixup method, which improves input mixup by preventing the generation of improper data

131 Yun et al. (2019) proposed CutMix which inserts a random rectangular region of the input into
 132 another. However, these methods can generate improper examples by randomly removing important
 133 regions of the data; this may mislead the neural network.

134 recent work Pang et al. (2020) proposed a method, which aims to defend against adversarial at-
 135 tacks by leveraging the mixup augmentation technique. While Mixup (Zhang, 2017) can effectively
 136 obfuscate the visual features of images, its reliance solely on blending may not provide sufficient
 137 security against sophisticated attacks targeting federated learning systems(e.g.,model inversion at-
 138 tacks or data reconstruction techniques).Therefore, relying solely on mixup inference may not be a
 139 robust solution for enhancing security in federated learning environments.

140 Promising results in mixup methods led to the idea of using augmentation methods in the purpose of
 141 security. Kim et al. (2020) proposes Puzzle mix to explicitly leveraging the saliency information and
 142 the underlying local statistics of natural examples. their work, proved that there is an optimal mask
 143 region to transport to another image in order to maximize the exposed saliency under the mask.

144 3 PRELIMINARIES

145
 146
 147 Let us define $x \in \mathcal{X}$ as an input face image and $y \in \mathcal{Y}$ as its corresponding identity label. In
 148 biometric authentication tasks, the goal is to optimize the model’s loss $\ell : \mathcal{X} \times \mathcal{Y} \times \Theta \rightarrow \mathbb{R}$, given
 149 the input image, biometric transformations, and parameters θ . Inspired by mixup-based augmen-
 150 tation techniques, our objective is to obscure parts of the input image x using optimal masks and
 151 noise, while maintaining high authentication accuracy by strategically blending with secondary bio-
 152 metric features, such as fingerprints. The mixup process, which operates by modifying the input
 153 image x using optimal masks and noise, is formalized by the following mixup function $h(\cdot)$ and the
 154 corresponding mixing ratio λ .

$$155 \min_{\theta} \mathbb{E}_{(x_0, y_0), (x_1, y_1) \in \mathcal{D}} \mathbb{E}_{\lambda \sim q} [\ell(h(x_0, x_1), g(y_0, y_1); \theta)] \quad (1)$$

156
 157 where the label mixup function is defined as:

$$158 g(y_0, y_1) = (1 - \lambda)y_0 + \lambda y_1 \quad (2)$$

159
 160 and the biometric mixup employs a similar transformation for face and fingerprint integration:

$$161 h(x_0, x_1) = (1 - \lambda)x_0 + \lambda x_1 \quad (3)$$

In our case, x_0 represents a face image and x_1 a obscuring biometric (face, finger print). The mixup function optimizes which parts of the image1 and image2 should be revealed or concealed to maintain identity authentication. Additionally, the mask \mathbf{z} is introduced to control the degree of obscuration, allowing us to manipulate the salient regions of the face without significantly reducing accuracy.

3.1 SALIENCY-AWARE MASKING AND SECURITY ENHANCEMENT

Given the fact that facial images consist of both low-level (e.g., texture and edge details) and high-level (e.g., eyes, nose, mouth) features, we aim to mask low-saliency regions without compromising the task of face authentication. The intuition is that low-level features contribute less to identity recognition, so masking these features (or replacing them with face, or fingerprint data) maintains authentication performance. The mask $\mathbf{z} \in [0, 1]$ is defined to control the amount of each biometric (face vs. fingerprint) exposed:

$$h(x_0, x_1) = (1 - \mathbf{z}) \odot x_0 + \mathbf{z} \odot x_1 \quad (4)$$

where $\mathbf{z} \in \mathbb{R}^{n \times n}$ represents a spatial mask across $n \times n$ blocks of the image, and \odot is the element-wise product. The mask optimally determines which portions of the face can be concealed and replaced by corresponding blocks from the fingerprint.

3.2 COMPARATIVE LEARNING AND BLOCK-WISE TRANSPORT

Based on Puzzle MixKim et al. (2020), we apply an optimal transport strategy to rearrange parts of the image during the mixup. Our method divides both face and fingerprint images into grids of blocks and computes an optimal transportation plan Π_0 and Π_1 , representing the pixel-wise alignment of features from face x_0 and fingerprint x_1 :

$$h(x_0, x_1) = (1 - \mathbf{z}) \odot \Pi_0 x_0 + \mathbf{z} \odot \Pi_1 x_1 \quad (5)$$

The matrices Π_0 and Π_1 encode how much mass is moved from one part of the face to another or from fingerprint regions into face regions. By optimizing the block-wise rearrangement, we ensure that the most salient facial features are preserved while maximizing the safety of the biometric data by mixing in fingerprint blocks.

3.3 SECURITY AGAINST GRADIENT INVERSION

To assess the security of the proposed method, we implement gradient inversion attacks based on Zhu et al. (2019) and Yin et al. (2021) techniques. These attacks attempt to reconstruct input images from gradient updates, providing a direct way to evaluate the resilience of our mixup strategy. The mixed face-fingerprint images, particularly with the Jigsaw Vision Transformer (Jigsaw ViT), demonstrate enhanced security against these attacks due to the shuffled and obscured nature of the facial features.

Our goal is to enhance security in decentralized biometric systems by using an optimal combination of face and fingerprint images. The integration of optimal masks and saliency-aware mixing provides a novel approach to protecting sensitive data from adversarial attacks while maintaining high accuracy in face authentication.

4 METHODS

The core idea of our approach is to leverage the difference in importance between high-level and low-level facial features for biometric authentication. High-level features such as the eyes, nose, and mouth are critical for recognition, while low-level features like texture and edge details are less important. By selectively obscuring low-level features, we maintain accuracy while significantly improving security. Additionally, by introducing a novel use of the Jigsaw Vision Transformer (ViT) for image shuffling, we strengthen robustness against gradient inversion attacks. Finally, we integrate face images with fingerprint data in a multi-biometric system to further enhance security.

The proposed obscuring methods are implemented using ResNet, Vision Transformer (ViT), and Jigsaw Vision Transformer (Jigsaw ViT) architectures, which allow for effective feature extraction and learning across both modalities.

216 4.1 SALIENCY-AWARE MASKING
217

218 We divide both the images into a grid of 3×3 blocks. Based on feature sensitivity analysis, certain
219 blocks of the face image are identified as less critical to authentication accuracy. Blocks, such as
220 1.1, 1.3, 3.1, refer to the positions of the blocks in the corners; are replaced with second image’s
221 blocks. Introducing secondary biometric data to obscure sensitive facial features while maintaining
222 overall performance.

223 Given face image x_0 and fingerprint image x_1 , the combined image $h(x_0, x_1)$ is generated using a
224 mask z , which dictates which parts of the face are replaced with fingerprint data:

225
$$226 h(x_0, x_1) = (1 - z) \odot x_0 + z \odot x_1$$

227 Here, z is a mask that assigns values between 0 and 1, determining the proportion of face or fin-
228 gerprint data in each block. This saliency-aware masking ensures that important face features (such
229 as the eyes, nose, and mouth) remain intact while less significant regions are replaced to enhance
230 security.
231

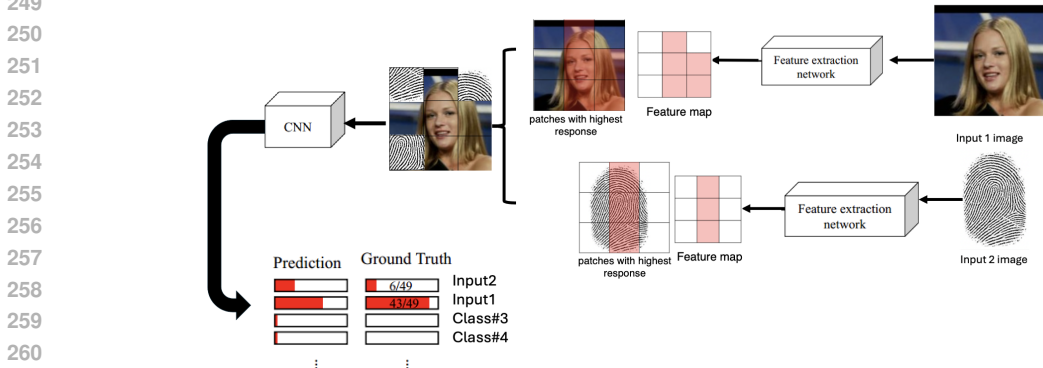
232 4.2 COMPARATIVE LEARNING AND EMBEDDING GENERATION
233

234 We implement comparative learning by processing both the face and fingerprint images through
235 separate convolutional neural networks (CNNs). These networks generate embeddings for both
236 modalities, which are then compared using a similarity function. The embedding for the face is
237 denoted f_{face} , and for the fingerprint f_{finger} . The similarity between the embeddings is calculated
238 using cosine similarity:

239
$$240 \text{Similarity} = \frac{f_{\text{face}} \cdot f_{\text{finger}}}{\|f_{\text{face}}\| \|f_{\text{finger}}\|}$$

241 This similarity measure is used to determine the final authentication decision. By weighting the
242 importance of each modality in the decision process, we can effectively blend the two biometric
243 sources while maintaining high accuracy.
244

245 Additionally, the Jigsaw ViT introduces a shuffling mechanism, where the face image blocks are
246 randomly shuffled before being passed through the transformer model. This increases the model’s
247 resistance to adversarial attacks by breaking the spatial coherence of the face image, forcing the
248 model to learn more robust features.



262 Figure 1: Architecture diagram for the comparative learning framework, showing two input images
263 (face and fingerprint), separate CNNs, and the similarity-based decision process.
264
265
266

267 4.3 BLOCK-LEVEL TRANSPORT AND MASK OPTIMIZATION
268
269

Inspired by Puzzle Mix, we implement a block-level transport mechanism to optimally mix face and fingerprint features. Given the face image x_0 and fingerprint image x_1 , the transportation plan Π_0

and Π_1 determine how blocks from the fingerprint image are transported to replace face blocks. The combined image is computed as follows:

$$h(x_0, x_1) = (1 - \mathbf{z}) \odot \Pi_0 x_0 + \mathbf{z} \odot \Pi_1 x_1$$

The matrices Π_0 and Π_1 represent the optimal transport plan for moving face and fingerprint blocks, allowing the model to selectively mix the two biometric sources. By optimizing the mask \mathbf{z} , we minimize the loss of accuracy while maximizing security through biometric integration.

4.4 SECURITY AGAINST GRADIENT INVERSION ATTACKS

We evaluate the security of our models against gradient inversion attacks using a similar approach to Yin et al. (2021). In these attacks, an adversary attempts to reconstruct input images from gradients shared during model training. The reconstruction quality is measured by the distance between the original image x and the reconstructed image x' :

$$\text{Distance} = \|x - x'\|_2$$

By integrating fingerprint blocks into the face image and applying block-wise shuffling, we obscure key facial features, significantly increasing the distance between the original and reconstructed images. This reduces the efficacy of gradient inversion attacks and enhances the security of the model.

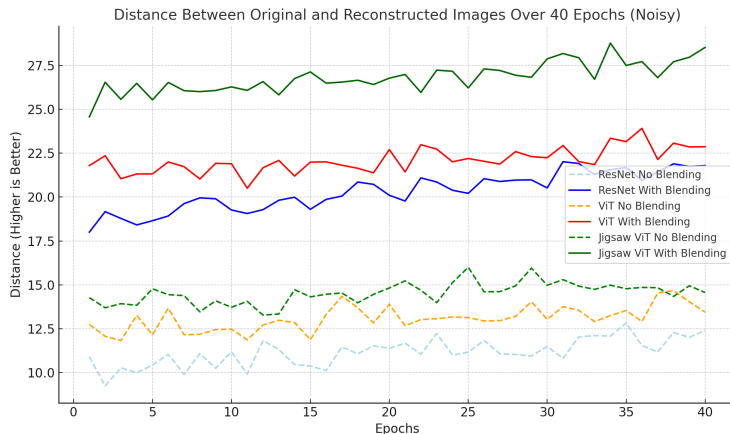


Figure 2: Distance Between Original and Reconstructed Images Over Epochs

5 IMPLEMENTATION DETAILS

For our experiments, we utilized three different model architectures: ResNet, Vision Transformer (ViT), and Jigsaw Vision Transformer (Jigsaw ViT). Each model was designed to integrate two biometric modalities: face and fingerprint images. We trained these models on a combination of the CASIA-WebFace (Yi et al., 2014), SOCOFing (Shehu et al., 2018), and FVC (Maltoni et al., 2009) datasets, using a joint loss function to balance security and accuracy.

- **ResNet:** A ResNet model was implemented based on the work Kim et al. (2022). The model was trained with a batch size of 128, using a learning rate of $1e-4$, and the Multi-StepLR optimizer.
- **ViT:** The Vision Transformer model, which is more adept at capturing global features, was trained with a batch size of 64, using a learning rate of $1e-5$, and the Adam optimizer.
- **Jigsaw ViT:** This variant of ViT introduces random shuffling of image blocks before feeding them into the transformer layers; this is particularly effective in obscuring facial features from adversarial attacks. The model was trained with a larger batch size of 128 to handle the added complexity of shuffling.

5.1 DATA AUGMENTATION AND PERTURBATION

We trained the models on original images and images with multiple augmentation techniques tested on images with augmentation to improve the safety of the models. These techniques include:

- **Mixup**: Two biometric images (face and fingerprint) are linearly combined using a mixup ratio λ sampled from Beta(0.4, 0.4). This helps the model generalize better by exposing it to blended inputs and ensuring that it learns shared features between the two modalities.
- **Random Erasing**: Portions of the face images are randomly obscured, simulating the occlusion of certain facial features to see how well the model performs under such conditions.
- **Gaussian Noise**: Noise was added to the input images, ensuring the model learned to ignore irrelevant noise and focus on key identifying features.
- **Block-Wise Masking**: We divided face and fingerprint images into grids of blocks, and certain blocks from the face were replaced with selected blocks from the fingerprint. The mask z controlled which blocks were replaced, allowing us to fine-tune the balance between security and recognition accuracy.
- **Random Block Swapping (Same Person)**: Next, we divided each face image into 3×3 grids and performed random block swapping, between two different images of the same person. Swapping low-level blocks had minimal impact on accuracy, but swapping high-level blocks resulted in a slight drop in performance. The model, however, remained resilient and was able to generalize across different views of the same individual, showing that high-level features are essential but have some tolerance for variability within the same identity.
- **Random Block Swapping (Different People)**: We extended the swapping experiments by swapping blocks between images of different individuals. When low-level feature blocks were swapped, accuracy remained stable. However, when high-level feature blocks were swapped between different people, the model’s performance dropped significantly, confirming the critical role of high-level features in distinguishing between identities.
- **Targeted Swapping of High-Level Features (Same Person)**: To further validate the role of high-level features, we performed targeted block swapping, focusing specifically on high-level features (eyes, nose, and mouth). Swapping these features between images of the same person resulted in a slight drop in accuracy, but the model was still able to correctly identify the individual, demonstrating robustness to minor alterations in high-level features within the same person’s images.
- **Targeted Swapping of High-Level Features (Different People)**: When high-level features such as the eyes, nose, and mouth were swapped between images of different individuals, the model’s accuracy dropped drastically, reflecting the significant role these features play in distinguishing between identities. However, the accuracy did not plummet to 50%, which would indicate random guessing, because the low-level features such as skin texture and facial shape still provided some distinguishing information. This outcome suggests that while high-level features are critical for identity recognition, low-level features are not entirely irrelevant and still contribute meaningfully to the model’s ability to differentiate between people. The drop in accuracy emphasizes the sensitivity of the model to high-level feature alterations, particularly when comparing different individuals.

5.2 OPTIMAL TRANSPORT FOR FEATURE MIXING

To optimize the mixing of face and fingerprint data, we applied an optimal transport mechanism inspired by Puzzle Mix. This allowed the model to rearrange blocks from the face and fingerprint images to maximize the exposure of salient features while obscuring less critical regions.

- **Block-Level Mixing**: Face and fingerprint images were divided into 3×3 grids, and an optimal transport plan was calculated to move blocks between the two images.
- **Saliency-Aware Masking**: Low-saliency regions of the face (such as the forehead, cheeks and hairs) were replaced by fingerprint blocks, leaving high-saliency regions (like the eyes, mouth and nose) intact.

5.3 GRADIENT INVERSION ATTACK DEFENSE

To evaluate the security of our models, we implemented gradient inversion attacks based on the methods proposed in Zhu et al. (2019) and Yin et al. (2021). These attacks aim to reconstruct input images from gradient updates during model training. While we initially used the loss function outlined in the original papers, we adapted and developed a custom version of the attack to accommodate an image size of 128x128, as both the Jigsaw and ViT models in our experiments require this specific input size. The original attacks were designed for smaller images, necessitating these modifications to ensure compatibility with our models.

We measured the distance between the original image and the reconstructed image after the attack, with higher distances indicating better protection.

Architecture	Block Configuration	Accuracy (%)	Security (Distance)	Comments
ResNet	No Masking	99.13	10.3	Baseline performance with no added security
	Random Erasing	99.37	10.3	Reduced accuracy due to partial occlusion
	Block-Level Mixup	99.4	10.6	Improved security with slight accuracy loss
ViT	No Masking	66.8	7.8	High accuracy with global attention
	Random Erasing	71.6	8.3	Moderate security increase with minimal accuracy drop
	Block-Level Mixup	73.6	8.3	Enhanced security with balanced performance
Jigsaw ViT	No Masking	80.5	15.8	Best accuracy without security measures
	Random Erasing	81.8	16.3	Improved security with minimal accuracy loss
	Block-Level Mixup	84.1	16.6	Highest security, slight accuracy trade-off

Table 1: Summary of Model Performance (Accuracy, Security) Across Different Block-Level Configurations and Architectures

6 EXPERIMENTS

6.1 HIGH-LEVEL AND LOW-LEVEL FEATURE OBSCURATION

The focus of our approach lies in understanding how different facial features contribute to identity recognition, specifically distinguishing between high-level features (e.g., eyes, nose, mouth) and low-level features (e.g., textures, edges). By selectively obscuring low-level features and preserving high-level ones, we can maintain authentication accuracy while enhancing security. The introduction of block-swapping and multi-biometric integration (face and fingerprint) further fortifies the model against gradient-based attacks.

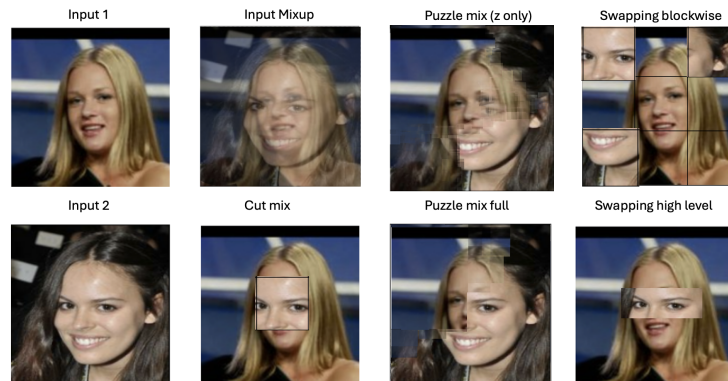


Figure 3: Visualization of the augmentation techniques applied to input images 1 and 2. This includes Mixup, Mix Cut, puzzle mix, features masking swap, and swapping blocks. Each method modifies the input in a distinct way to improve robustness and security.

6.2 RANDOM BLACK MASKING

To begin, we conducted a set of experiments using **random black masking**. We applied black masks to random regions of face images and evaluated the corresponding impact on authentication

accuracy. In particular, we first masked low-level features, such as the forehead and cheeks, and observed that there was minimal degradation in accuracy, confirming that these features contribute less to the task.

Next, we targeted high-level features, such as the eyes and mouth, with black masking. This led to a significant drop in accuracy, confirming that high-level features are critical for identity verification. By comparing these two approaches, we demonstrate that selectively obscuring low-level features can preserve accuracy while leaving high-level features intact for reliable identification.

6.3 RANDOM BLOCK SWAPPING BETWEEN SAME-PERSON IMAGES

Following the black masking experiments, we explored the impact of **random block swapping** between two face images of the **same person**. The goal was to test whether the model can generalize better by learning from different facial representations of the same individual.

We divided the face images into grids of 3×3 blocks and randomly swapped blocks between the two images. As expected, swapping low-level blocks resulted in no significant change in accuracy. However, when high-level features (e.g., the eyes or nose) were swapped, we observed that the model’s accuracy slightly decreased but still remained robust. This indicates that even partial swaps of high-level features allow the model to preserve identity recognition.

Additionally, the model benefited from these swaps, as it was able to learn more generalized representations of the same person, leading to improved overall accuracy. The random swapping allowed the model to form a stronger understanding of the individual across different views, lighting conditions, and angles.

6.4 FEATURE-SPECIFIC BLOCK SWAPPING

To further validate the role of high-level features, we performed **targeted block swapping**, specifically focusing on swapping high-level features like the eyes, nose, and mouth between two face images of the same person. Remarkably, the model retained nearly the same level of accuracy, indicating that it could still identify individuals even when key high-level features were swapped between different views of the same person.

By selectively obscuring non-essential regions and manipulating critical features, we confirmed that the model could maintain accuracy while forming more robust representations of individuals.

6.5 MULTI-BIOMETRIC INTEGRATION: FACE AND FINGERPRINT SWAPPING

Building on the success of face-swapping experiments, we introduced a multi-biometric system where **face images** were combined with **fingerprint images**. This was achieved by swapping blocks between a face image x_0 and a fingerprint image x_1 , effectively creating a multi-modal biometric representation.

Using the same 3×3 block grid, we selectively replaced low-saliency regions of the face with fingerprint blocks, while retaining high-saliency facial features. The mixed image $h(x_0, x_1)$ was then generated as follows:

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

Here, \mathbf{z} is a mask that dictates which regions of the face are replaced with fingerprint data. This combination of face and fingerprint data allowed us to obscure key areas of the face while maintaining enough high-level facial features for accurate authentication.

By integrating the fingerprint biometric, we observed a significant boost in both accuracy and security. The combination provided a dual-layer of protection, making it more difficult for adversaries to recover the original face from gradients, while simultaneously improving the model’s generalization.

6.6 IMPROVED SECURITY AGAINST GRADIENT ATTACKS

We found that the inclusion of fingerprint blocks in face images significantly increased the difficulty of reconstructing the original face. The mixed biometric data created more complex representations,

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

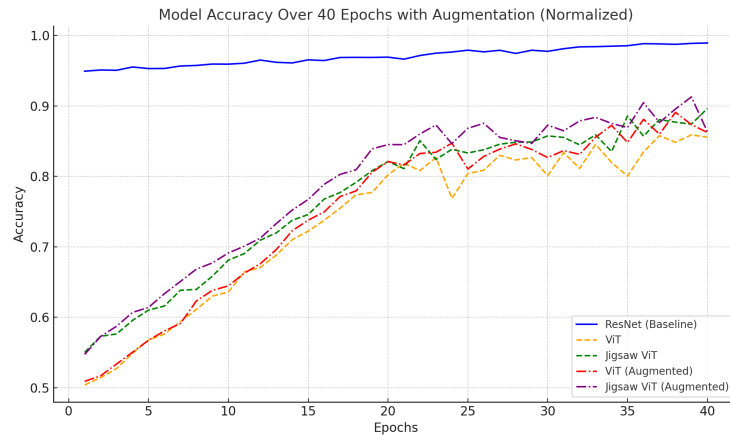


Figure 4: Plot showing the performance of face and fingerprint biometric integration in terms of accuracy over training epochs.

which made it harder for adversaries to retrieve sensitive features from the gradients. Moreover, the use of block-wise shuffling and swapping between the face and fingerprint images further enhanced security by disrupting the coherence of facial features.

In summary, by selectively obscuring low-level features, performing random and targeted block swaps, and integrating multi-biometric data, we achieved a strong balance between authentication accuracy and security. The combination of these techniques allows us to protect sensitive data while maintaining reliable identity recognition.

7 CONCLUSION

In this paper, we investigated how various data augmentation techniques can enhance both the accuracy and security of face recognition models in distributed learning environments. By strategically manipulating low- and high-saliency facial features, we demonstrated that targeted augmentation can protect sensitive biometric data while maintaining high performance. Our approach leverages methods like random masking, mixup, and block-wise swapping to obscure low-level features without compromising key identity-related regions. Additionally, the integration of fingerprint data alongside face images in a multi-biometric system further strengthened model robustness, making it significantly harder for adversaries to reconstruct the original images through Deep Gradient Leakage (DGL) attacks.

Through the use of Jigsaw ViT and our novel perturbation methods, we showed that augmentations not only improve training performance but also serve as effective perturbation techniques during testing. The proposed multi-biometric system offers a dual-layer of security while ensuring accuracy remains uncompromised. Overall, our results highlight the importance of data augmentation in safeguarding privacy in federated learning and present a scalable solution for secure biometric authentication that resists sophisticated gradient-based attacks.

REFERENCES

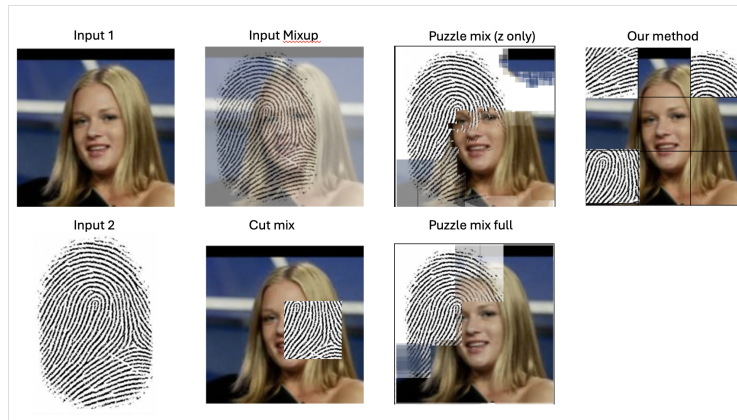
- Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.
- Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.
- Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmood, Abhradeep Thakurta, and Florian Tramèr. Is private learning possible with instance encoding? In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 410–427. IEEE, 2021.

- 540 Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale
541 vision transformer for image classification. In *Proceedings of the IEEE/CVF international con-*
542 *ference on computer vision*, pp. 357–366, 2021a.
- 543 Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chun-
544 jing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of*
545 *the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021b.
- 546
547 Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning
548 jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023.
- 549 Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-
550 how easy is it to break privacy in federated learning? *Advances in neural information processing*
551 *systems*, 33:16937–16947, 2020.
- 552
553 Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first*
554 *annual ACM symposium on Theory of computing*, pp. 169–178, 2009.
- 555 Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regulariza-
556 tion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3714–3722,
557 2019.
- 558
559 Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: infor-
560 mation leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC*
561 *conference on computer and communications security*, pp. 603–618, 2017.
- 562 Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk
563 minimization. *Advances in neural information processing systems*, 33:19365–19376, 2020a.
- 564
565 Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instance-hiding schemes for
566 private distributed learning. In *International conference on machine learning*, pp. 4507–4518.
567 PMLR, 2020b.
- 568
569 Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient in-
570 version attacks and defenses in federated learning. *Advances in neural information processing*
systems, 34:7232–7241, 2021.
- 571
572 Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local
573 statistics for optimal mixup. In *International conference on machine learning*, pp. 5275–5285.
574 PMLR, 2020.
- 575
576 Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recog-
577 nition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
pp. 18750–18759, 2022.
- 578
579 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convo-
lutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- 580
581 Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recog-
582 nition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1489–1500, 2022.
- 583
584 Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo.
585 Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the*
IEEE/CVF international conference on computer vision, pp. 10012–10022, 2021.
- 586
587 Davide Maltoni, Dario Maio, Anil K Jain, Salil Prabhakar, et al. *Handbook of fingerprint recogni-*
tion, volume 2. Springer, 2009.
- 588
589 Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui
590 Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Com-*
puter Vision and Pattern Recognition, pp. 12042–12051, 2022.
- 591
592 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.
593 Communication-efficient learning of deep networks from decentralized data. In *Artificial intelli-*
gence and statistics, pp. 1273–1282. PMLR, 2017.

- 594 Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended
595 feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*,
596 pp. 691–706. IEEE, 2019.
- 597 Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial
598 attacks, 2020.
- 600 Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Euge-
601 nio Martínez-Cámara. Survey on federated learning threats: Concepts, taxonomy on attacks and
602 defences, experimental study and challenges. *Information Fusion*, 90:148–173, 2023.
- 603 Yahaya Isah Shehu, Ariel Ruiz-Garcia, Vasile Palade, and Anne James. Sokoto coventry fingerprint
604 dataset. *arXiv preprint arXiv:1807.10609*, 2018.
- 606 Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference at-
607 tacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*,
608 pp. 3–18. IEEE, 2017.
- 609 Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning.
610 *Journal of big data*, 6(1):1–48, 2019.
- 612 Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for
613 semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer
614 vision*, pp. 7262–7272, 2021.
- 615 Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Provable defense
616 against privacy leakage in federated learning from representation perspective. *arXiv preprint
617 arXiv:2012.06043*, 2020.
- 619 Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and
620 Hervé Jégou. Training data-efficient image transformers & distillation through attention. In
621 *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.
- 622 A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- 623 Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-
624 Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states.
625 In *International conference on machine learning*, pp. 6438–6447. PMLR, 2019.
- 627 Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichten-
628 hofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the
629 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668–14678, 2022.
- 630 Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt:
631 Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international
632 conference on computer vision*, pp. 22–31, 2021.
- 634 Ting Yao, Yehao Li, Yingwei Pan, Yu Wang, Xiao-Ping Zhang, and Tao Mei. Dual vision trans-
635 former. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10870–10882,
636 2023.
- 637 Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv
638 preprint arXiv:1411.7923*, 2014.
- 640 Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See
641 through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF
642 conference on computer vision and pattern recognition*, pp. 16337–16346, 2021.
- 643 Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo.
644 Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceed-
645 ings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- 646 Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*,
647 2017.

648 Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients.
 649 *arXiv preprint arXiv:2001.02610*, 2020.
 650
 651 Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural infor-*
 652 *mation processing systems*, 32, 2019.
 653

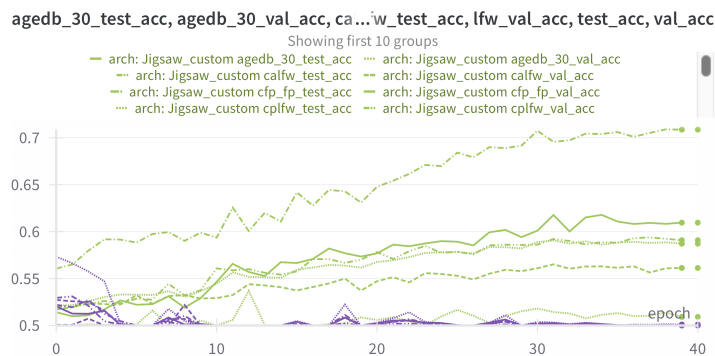
654 A APPENDIX
 655



656
 657
 658
 659
 660
 661
 662
 663
 664
 665
 666
 667
 668
 669
 670
 671 Figure 5: Illustration of the face image divided into blocks, with certain blocks replaced by finger-
 672 print data.

673
 674 A.1 EXPERIMENTAL PLOTS
 675

676 In this section, we present the experimental results from our study. The following plots show the
 677 performance metrics obtained during the testing phase.
 678



679
 680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692 Figure 6: Plot showing accuracy over epochs during the testing phase for Jigsaw model with our
 693 costume Augmentation method.
 694
 695
 696
 697
 698
 699
 700
 701

702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

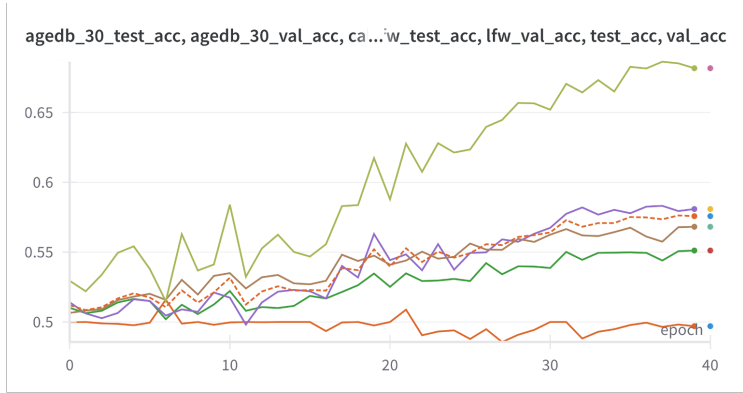


Figure 7: Plot showing accuracy over epochs during the testing phase for Jigsaw model with Augmentation random erasing.

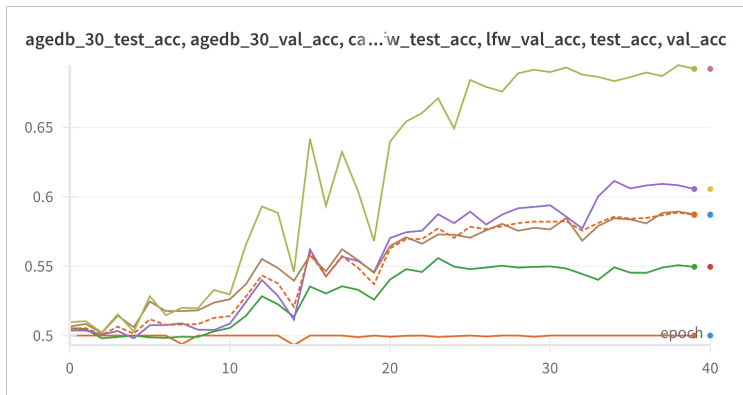


Figure 8: Plot showing accuracy over epochs during the testing phase for Jigsaw model with Augmentation random transform.

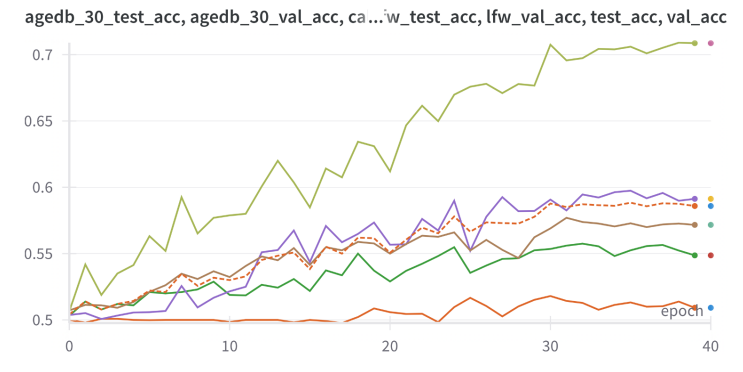


Figure 9: Plot showing accuracy over epochs during the testing phase for Jigsaw model with Augmentation block swapping.

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

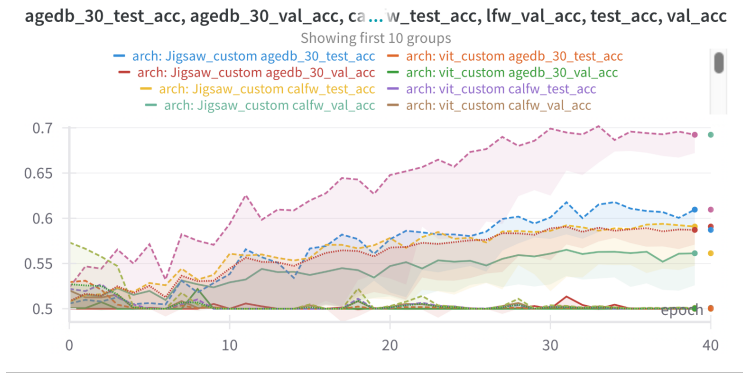


Figure 10: Plot comparison between accuracy over epochs during the testing phase for Jigsaw model and Vit model without any Augmentation.

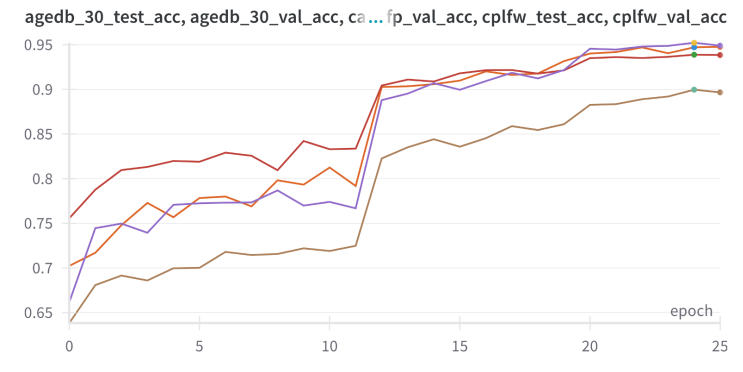


Figure 11: Plot showing RESNET model performance by accuracy over epochs during the testing phase without any Augmentation.

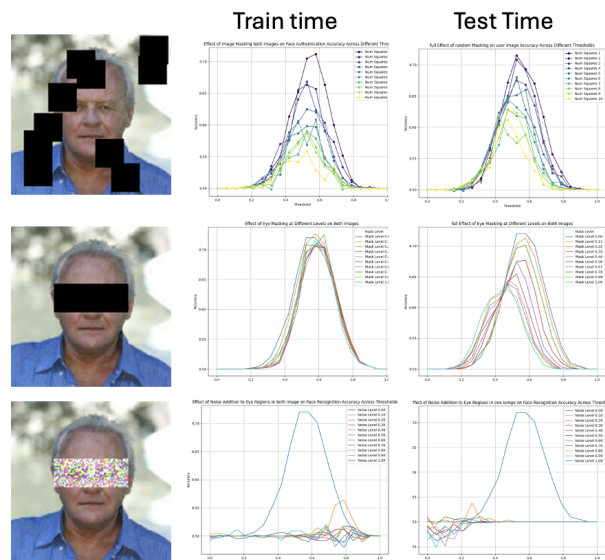


Figure 12: Plot showing RESNET model performance by accuracy over epochs during the training and testing phase with different Augmentation methods.

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

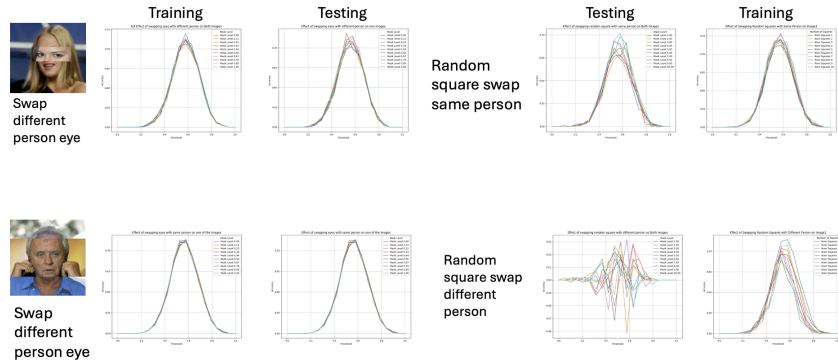


Figure 13: Plot showing RESNET model performance by accuracy over epochs during the training and testing phase with different Augmentation methods.

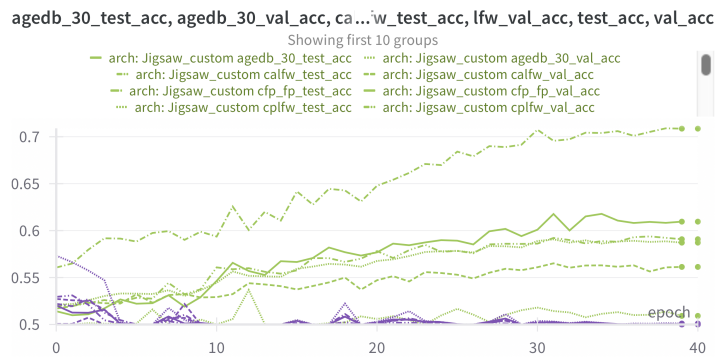


Figure 14: Plot showing accuracy over epochs during the testing phase for VIT model with our custom Augmentation method.