FRA: a Novel Face Representation Augmentation Algorithm for Face Recognition

Soroush Hashemifar, Abdolreza Marefat,
Javad Hassannataj Joloudari and Hamid Hassanpour

January 29, 2023

FRA: A novel Face Representation Augmentation algorithm for face recognition

Soroush Hashemifar[1], Abdolreza Marefat[2], Javad Hassannataj Joloudari[3,*], Hamid Hassanpour[4]

[1]School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[2]Department of Artificial Intelligence, Technical and Engineering Faculty, South Tehran Branch, Islamic Azad University, Tehran, Irans
[3]Department of Computer Engineering, Faculty of Engineering, University of Birjand, Birjand 9717434765, Iran
[4]Faculty of Computer Engineering & Information Technology, Shahrood University of Technology, P.O. Box 316, Shahrood, Iran

Corresponding author[*]: javad.hassannataj@birjand.ac.ir

Abstract
A low amount of training data for many state-of-the-art deep learning-based Face Recognition (FR) systems causes a marked deterioration in their performance. Although a considerable amount of research has addressed this issue by inventing new data augmentation techniques, using either input space transformations or Generative Adversarial Networks (GAN) for feature space augmentations, these techniques have yet to satisfy expectations. In this paper, we propose a novel method, named the Face Representation Augmentation (FRA) algorithm, for augmenting face datasets. To the best of our knowledge, FRA is the first method that shifts its focus towards manipulating the face embeddings generated by any face representation learning algorithm in order to generate new embeddings representing the same identity and facial emotion but with an altered posture. Extensive experiments conducted in this study convince the efficacy of our methodology and its power to provide noiseless, completely new facial representations to improve the training procedure of any FR algorithm. Therefore, FRA is able to help the recent state-of-the-art FR methods by providing more data for training FR systems. The proposed method, using experiments conducted on the Karolinska Directed Emotional Faces (KDEF) dataset, improves the identity classification accuracies by 9.52 %, 10.04 %, and 16.60 %, in comparison with the base models of MagFace, ArcFace, and CosFace, respectively.

Keywords:
Face Recognition, Face Embeddings, Face Representation Learning, Autoencoder, Vision Transformers, Latent Space Data Augmentation, Facial Pose Reconstruction

1. Introduction

Face images are one of the most popular biometric modalities which have been continuously utilized in Face Recognition (FR) systems [1]. It is used in a wide range of contexts with the aim of identity authentication and its applications vary from daily life and finance to military and public security [2]. In fact, in comparison with other biometrics, such as the fingerprint, iris, or retina which are ubiquitously used for authorizing individuals, FR can provide us with the most convenient way to capture visual information without the need for any extra activity from the subject. In recent years, FR has been one of the most proactively studied areas in Computer Vision [3]. Particularly, with the advent of deep learning and architectures like Convolutional Neural Networks (CNNs) [4], a large number of efficient facial recognition methods with outstanding performance have been invented to address this challenge [5-12]. These successful algorithms depend heavily on the performance of neural networks which use a cascade of layers comprised of neurons that are able to learn different levels of abstractions and representations from the input data [2]. These representations are more powerful substitutions for hand-crafted features from facial attributes such as Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) [13, 14]. Their principal advantage is that they obviate the need for manually and exhaustively searching for the best features representing one's face. Moreover, the process of learning representations via deep learning-based algorithms makes the generated features surprisingly discriminative in that the inter-class diversity and intra-class compactness within the training data are all taken into account by the network itself [15].

However, there are still problematic scenarios in which FR systems fail to realize the expectations. For instance, in real-life situations, the imagery of a person's face has a high chance of being in a variety of facial expressions, occlusions, poor illumination, low resolution, etc. [16-18], and all these factors cause substantial degradation of the overall performance of the current FR algorithms. Thus, different approaches have been adopted to rectify the negative impact of such barriers in FR systems [19-21]. Some have opted for experimenting and devising new loss functions whose capability to better feedback to their neural network in the backpropagation step, enables the extracted deep features to be more discriminative and clearly separable [2, 6, 9, 22-26]. In addition to these works, different architectures have been implemented to extract feature maps which are more useful in terms of facial representations.

Moreover, developing larger and more variant datasets has been one of the main stimuli which have been pushing the boundaries in recent FR systems [27]. Nevertheless, although some of these benchmark datasets can be found in large volumes, we often lack such a training set of images when it comes to real use cases. A typical case would be a situation in which the goal is to train a deep learning-based method on a private, in-house set of identities that have been chosen by a multimedia organization for video indexing purposes. The data-gathering phase can be very time and labor-consuming and sometimes even impossible, and it acts as an impediment in the way of achieving a tailored amount of training datasets. These have motivated researchers to pave the way by introducing different data augmentation techniques.

Data augmentation refers to a set of techniques that are used to increase the number of training datasets without the loss of previously annotated data. The benefit of such methods is that it equips the trained model with more generalizability and acts as a regularizer in the case of overfitting which is one of the most frequent complications when dealing with a small amount of training data [28, 29]. Overall, there are two mainstream categories of methods for augmenting data. The first set of methods has the aim of manipulating the data in the input space in that they simply take the input image and apply different geometric transformations such as translations, cropping, vertical and horizontal flipping, rotation, etc [30]. Even though these methods are proven to be extremely useful in some other challenges like image classification, object detection, and image captioning in computer vision, in the case of FR they cannot be as helpful as they expected. The main reason is that in order for any FR system to capture a reliable visual representation of a face crop image, the content should be aligned in terms of facial landmarks. This means that any geometric alteration on these which conspicuously happens when one uses these classical methods, can perturb the overall performance of FR pipeline. These challenges have motivated the researchers to shift their studies' direction toward more modern and domain-specific solutions [31-33], leading to the second set of methods, which are known to be Generative Adversarial Networks (GANs) [34]. These methods are the well-known type of generative models which are used with the objective of transforming the input data in feature space with the aim of generating new augmented image data. This group of models is capable of adjusting the facial attributes existent in a face image such as hair style, expression, posture, skin color, etc. to a target style. However, in most cases, these generative models cannot create realistic outputs and these models deal with the high complexities of reconstructing the feature space to input space, without having any considerable improvement on the downstream task, which in our case, is classification on the identity of the samples.

In order to address these difficulties, in this paper we propose the Face Representation Augmentation (FRA) algorithm. This algorithm augments the posture of a given face image in the latent space. This means that, given a set of embeddings representing a specific person, the proposed approach alters the embedding to sustain the identity-related features with a transformed pose feature. The FRA algorithm can help the existing facial recognition systems especially when the number of training samples is imbalanced or less than expected. Our main contributions in this paper are itemized in the following:

1. A novel algorithm for facial posture augmentation inside the latent space to reduce the complexity of the image augmentation problem.
2. Generating noiseless, non-duplicated embeddings which are proved to be linearly separable.
3. Extensive experiments were conducted on the Karolinska Directed Emotional Faces (KDEF) [35] dataset and improved the identity classification accuracies in comparison with the base models of MagFace, ArcFace, and CosFace, respectively.

The rest of the paper is organized as follows. In Section 2, we briefly review the related works on face-specific data augmentation and representation learning. Then, in Section 4, we present the details of our proposed methodology. In Section 4, we demonstrate the results of our experiments in comparison with other related state-of-the-art approaches. Finally, the conclusion will be drawn in Section 5.

2. Related works

In this section, we present an overview of face-specific data augmentation techniques. These are categorized into two groups classical and generative-based methods in 2.1. Additionally, we review the related literature of FR algorithms in 2.2.

2.1. Face-Specific Data Augmentation

To begin with, five data augmentation techniques for face photos were reported by Lv et al. [29]. These techniques were landmark perturbation, hairdo synthesis, glasses synthesis, postures synthesis, and lighting synthesis. Vincent et al. [36] tried to synthesize more data by applying different types of noise such as Gaussian and Salt-and-pepper with the objective of training Stacked Denoising Autoencoders on more complicated samples. Wang et al. [37] addressed the issue of data augmentation in picture classification using conventional transformation techniques and GANs. They also suggested a technique for learning network-based augmentations that better enhance the classifier in the setting of generic photos rather than face images.

Moreover, although the hair is not an intrinsic part of the human face, it interferes with facial recognition since it obscures the face and changes its appearance. Using DiscoGAN, which was developed to find cross-domain relationships using unpaired data, Kim et al. altered hair color. In addition to the color, Kim et al. in [38], suggested changing the bang by transferring an unsupervised visual characteristic using a reconfigurable GAN. An online compositing technique was used in the face synthesis system proposed by Kemelmacher-Shlizerman et al. [39]. The system might produce a series of fresh photographs with the input person's identification and the questioned look using one or more photos of their face and a text query like curly hair. Jiang et al., in [40], proposed Pose and expression resilient Spatial aware GAN (PSGAN). It starts by using Makeup Distill Network to separate the reference image's makeup into two spatially aware makeup matrices. After that, a module called Attentive Makeup Morphing is developed to let users describe how a pixel's appearance in the source picture is altered based on the reference image. In order to ease applications in the real-world setting, PSGAN is the first to concurrently accomplish partial, shade tunable, and pose/expression robust makeup transfer. In order to separate the makeup from the reference picture as two makeup matrices, an MDNet is also included. The flexible partial and shade adjustable transfer is made possible by the spatially aware makeup matrices. To learn all cosmetics attributes [41], including color, form, texture, and position, it comprises an enhanced color transfer branch and a new pattern transfer branch. They present makeup in this work as a combination of color transformation and pattern addition, and they create a thorough makeup transfer technique that works for both delicate and dramatic looks. They suggest using warped faces in the Ultraviolet (UV) space while training two network branches to eliminate the disagreement between input faces in terms of form, head posture, and expression. They also create a new architecture with two branches for color and pattern transfer. They present brand-new cosmetics transfer datasets with extreme fashions that were not taken into account in the earlier datasets.

2.2. Representation Learning for Face Recognition

Representation learning refers to a set of algorithms that are designed to solve a variety of challenges like image retrieval [42-44], the person [45, 46] and vehicle [47, 48] re-identification, landmark detection, and fine-grained object recognition [49, 50]. The task of face recognition in computer vision is heavily dependent on learning representations that have fine intra-class and large inter-class distances [51]. Previous works [6, 22, 25, 52, 53] have mainly adopted different, more robust loss functions with the aim of learning representations that satisfy the aforementioned requirements.

In [52], a deep convolutional neural network, named FaceNet, was proposed which learns facial representations with the help of triplet loss. The main objective of this work is to achieve an embedding $f(x)$ from an image x into a d-dimensional Euclidean space R^d. The obtained embedding is generated in a way that the squared distance among the embeddings from one class is small and that of the embeddings from different classes is large. This algorithm achieves 99.63% and 95.12% accuracy in LFW [54] and YouTube Faces Database [55] respectively. Liu et al. [53] have proposed a new look at the loss functions which are based on the Euclidean margin between the produced embeddings. For CNNs to learn discriminative facial characteristics with clear and innovative geometric interpretation, they suggest the A-Softmax loss. The assumption that faces also lie on a manifold is fundamentally compatible with the learnt features' discriminative spread on a hypersphere manifold. In order to approximate the learning problem that minimal inter-class distance is greater than maximum intra-class distance, they develop lower the margin set between such classes.

In [22], the authors have proposed ArcFace, a major modification of the Softmax loss to further improve the robustness of the learned deep features. By utilizing the arc-cosine function to calculate the angle between the current feature and the target weight and adding an additive angular margin to the target angle, the target logit can be obtained. Then, these logits are rescaled by a fixed feature norm followed by exactly the same steps in the Softmax loss function. Their approach has the following advantages over the others. (1) Directly optimizing the geodesic distance margin (2) State-of-the-art performance in several benchmark datasets: achieving 99.53% accuracy (3) Easiness in terms of implementation (4) Efficiency in terms of computational complexity.

In [25], the authors reformulated the Softmax loss as a cosine loss with the aim of introducing a novel loss function, named Large Margin Cosine Loss (LMCL). Their improvement is to further maximize the decision margin in the angular space by introducing and training a deep model called CosFace. In this deep model, LMCL guides the convolutional layers to learn features with huge cosine margins. Their results demonstrate that they have achieved 97.96% accuracy in face verification on the MegaFace benchmark, which has been a major improvement in comparison to previous works.
Meng et al. [6] proposed a new set of losses that enable the network to learn embeddings whose magnitude represents the quality of the given face. By extending ArcFace [22] and introducing the MagFace loss function, they demonstrate that the more likely the subject is to be recognized, the bigger the magnitude of the generated embedding becomes. MagFace learns to generate these universal embeddings by pulling the easier samples within a class of identities to the class center and pushing them away from the origin. This makes the embeddings robust to ambiguity and the absence of high discriminative features which prevalently exist in unconstrained face images in real scenarios. They have achieved 99.83% verification accuracy in the LFW benchmark dataset. In Table 1, a comparison of these works is depicted.

Table 1. Verification accuracy of MagFace, CosFace, ArcFace, and SphereFace. These models are evaluated on CALFW, CPLFW, AgeDB, LFW, and CFP-FP datasets.

Method	CALFW [56]	CPLFW [57]	AgeDB [58]	LFW [54]	CFP-FP [59]
MagFace [6]	96.15	92.87	98.17	99.83	98.46
CosFace [25]	96.18	92.18	98.17	99.78	98.26
ArcFace [22]	95.96	92.72	98.05	9981	98.40
SphereFace [53]	95.58	91.27	97.05	99.67	96.84

Moreover, although these approaches have significant performance, directly applying GAN approaches appears to have a few disadvantages. Models collapse, difficulty in training and convergence problems, and poor image generation effect, along with the unreliable results of the generator for unconstrained input images, cause the generated image examples to be incapable of being utilized for industrial data augmentation tasks [61 ,60].

3. Proposed approach

3.1. Overview

This section presents the proposed FRA algorithm. As can be inferred from Figure 1, our method includes four steps. These are as follows: face detection and alignment, input preparation: facial landmark and representation extraction, pose feature extraction, and representation augmentation. Steps 1 and 2 comprise our data preprocessing pipeline which is explained in Section 3.2. Steps 3 and 4 represent our main contribution to this paper and are explained in Section 3.4.

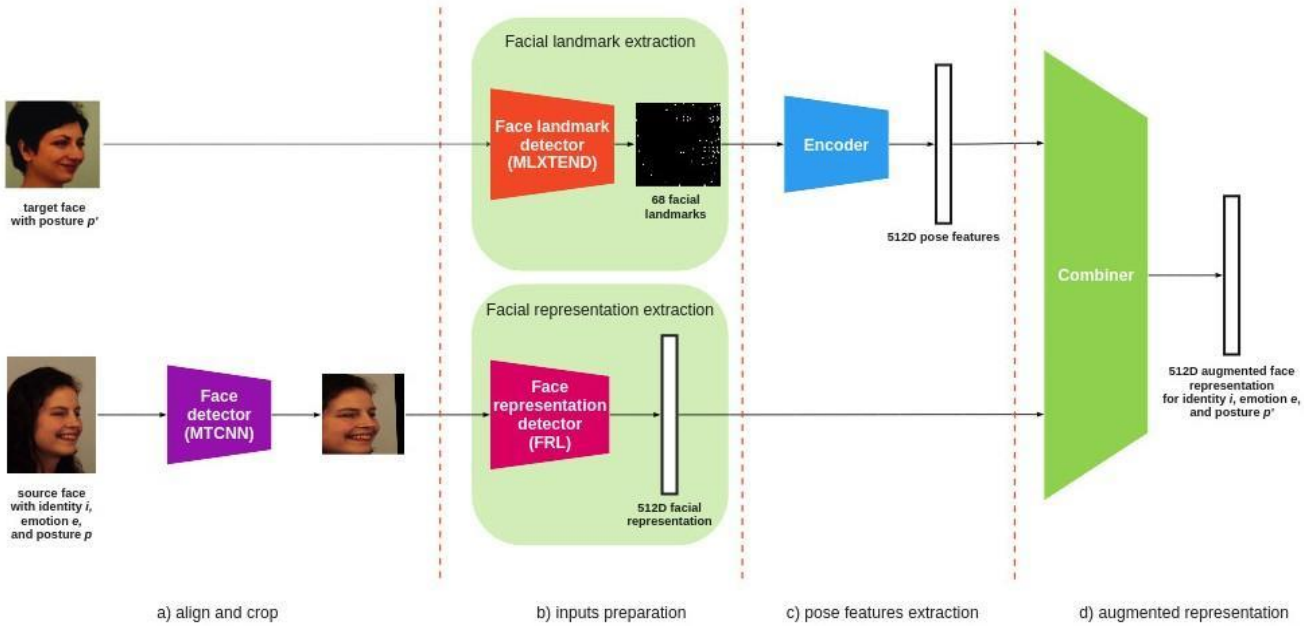


Figure 1. The overall procedure of FRA. FRA is composed of four steps to generate a new representation vector with identity *i*, emotion *e*, and posture *p*, by applying a target posture *p* on a base image with identity *i* and emotion *e*.

3.2. Dataset preprocessing and preparation

Our data preprocessing step includes three main phases. These three phases are depicted in Figure 1. As is seen in the first phase, we feed the raw face images to the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm [53] which is a robust face and landmark detector. MTCNN provides us with 5 landmark points, including the center of both eyes, the tip of the nose, and the left and right corners of the lips, and a bounding box that perfectly encloses the face area within the image without any padding. In this phase, we also align the face images by feeding the acquired facial landmarks along with the face image itself to the method of warp affine which exists in OpenCV [62], a famous library with ready-to-use computer vision-related algorithms.

In the second phase, we feed the aligned face images to MLXTEND[1] so as to determine more facial landmarks. As is shown in Figure 1, MLXTEND outputs 68 facial key points which we use to construct binarized images with pixel value 0 (completely black) for the background and 1 (completely white) for facial landmarks. On the other hand, we need to have fixed-size embeddings for each sample within the dataset. These embeddings are in fact the training data for the combiner module which will be explained in Section 3.4. In our case, we use two of the most reliable and robust face representation learning algorithms, namely MagFace [6] and FaceNet [52], for obtaining embeddings for each image. MagFace's learning procedure is for a universal embedding that is quality aware, meaning that the easier the sample is for the recognition task, the closer its feature vector becomes to the center of the class. Furthermore, FaceNet is an algorithm that directly learns a mapping from the samples to a compact Euclidean space and the distances correlate to the similarity degree of a given pair of face images. In Phase 3, the binarized images generated in Phase 2 are fed to the AE model in order to generate an embedding vector representing posture features.

[1]http://rasbt.github.io/mlxtend/

Finally, in Phase 4, pose and face representation vectors are fed into the combiner module to generate an augmented face representation vector.

3.3. Facial Landmark Restoration using Autoencoders

Autoencoders (AE) are a particular type of neural networks whose main functionality is to encode the input into a meaningfully compacted representation and decode this into the input space afterwards [63, 64]. Following this paradigm, in this paper, we have been inspired by the work done by Meng et al. [65] and decided to use an AE-based model for encoding our input space (binarized images of landmarks explained in Section 3.2) into the latent space (embeddings), as shown in Figure 2. Given *Si* as a sample of facial landmarks image, the output of $F(S_i)$ is a reconstructed image S'_i, where $F(S_i)$ is, $B * A$. After the AE model's convergence, we can discard *B (decoder)* and take only *A* which has learned to encode the input into an optimized and meaningful latent space representation denoted by V_i. It is worth mentioning that V_i plays a vital role in our proposed method which is the latent representation of the posture of the face. Figure 2 illustrates the proposed AE-based model and its architecture.

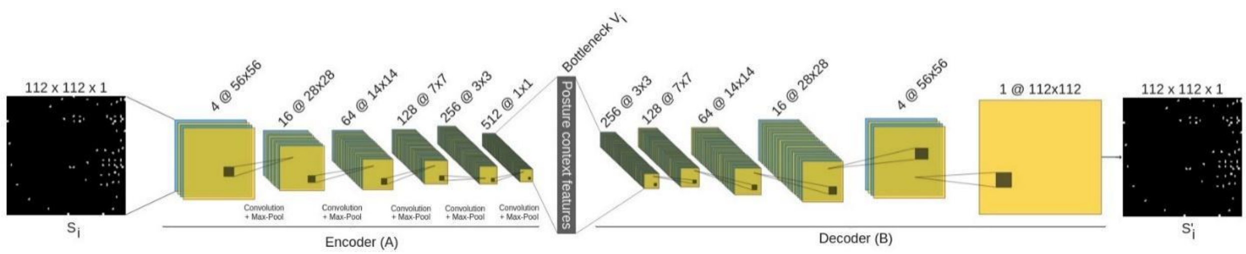


Figure 2. A general architecture of an autoencoder-based model. FRA utilizes a typical convolutional autoencoder with a bottleneck of size 512. This bottleneck vector is used in further steps.

3.4. Combining Feature Vectors and Feature Extraction using Vision Transformers

Vision Transformers (ViT) are deep learning models whose versatility in various fields such as natural language processing, speech recognition and computer vision has made them a prominent choice for researchers [66]. In comparison with the conventional CNNs, ViT models have achieved competitive superior results in vision tasks like object detection [67], image recognition [68], image super-resolution [69], and segmentation [70, 71].

At the core of ViT models, there is a mechanism of attention that has been probably one of the most significant concepts in the domain of deep learning. Its inspiration is the biological attributes of human beings in that, to recognize an object, we tend to focus on the most distinctive parts of that entity instead of paying attention to all parts of it as a whole [72]. In terms of deep neural networks, this can be interpreted as assigning importance scores for a given set of features where the higher scores are for more relevant features and the lower ones for the features with less saliency [73]. As can be observed from Figure 3, the model learns to have more focus on the parts which represent the target object in the image.

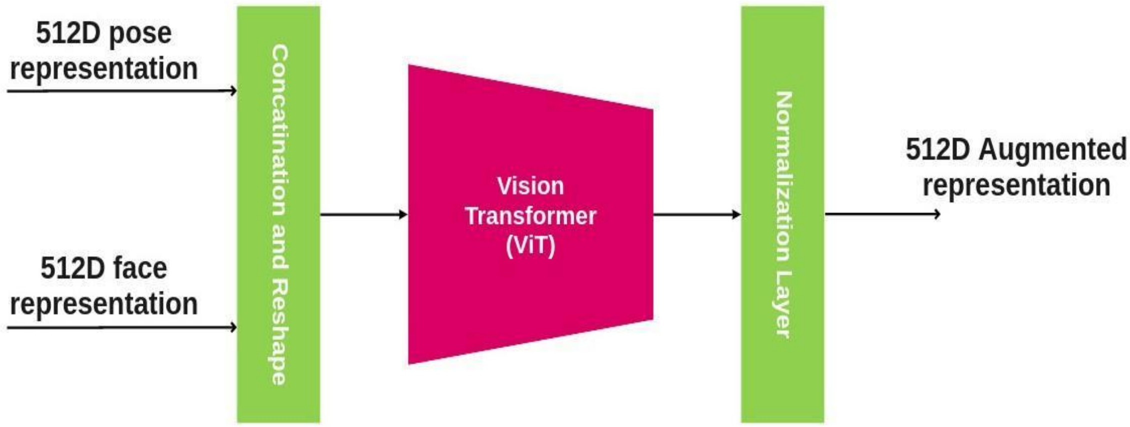


Figure 3. The paradigm of combining two representation vectors using ViT. The combiner takes two representation vectors with a size of 512 and combines them into a 32x32 matrix to be processed by a vision-transformer component.

Moreover, transformers [74] refer to a set of neural networks which use the mechanism of attention. These models consist of multiple encoders and decoders whose architectures are identical to each other. In these models, a multi-head self-attention (MSA) mechanism is used for encoding the input, followed by decoders which include an extra attention layer in order to process the encoder's output. Self-attention is a function denoted in Equation (1).

$$Attention(Q,K,V) = softmax\left(\frac{Q.K^T}{\sqrt{d_k}}\right).V \quad \text{s.t. } Q = W^Q x, K = W^K x, V = W^V x \tag{1}$$

where W^Q, W^K, and W^V are weight matrices used in linear transformations on inputs *x* to produce *Q*, *K*, and *V*. The attention score is then calculated by $Q.K^T$ as the dot product of the query and each key, scaled by the dimension d_k of the key *K*. Put *x = "x1, x2, x3, ... , xn"* to calculate an answer based on a collection of queries *Q*, keys *K*, and values *V*. In MSA, *Q*, *K*, and *V* are projected linearly and this is done for *h* consecutive times with different learned weights. Then, by applying the self-attention mechanism on each of the outputs in the previous step simultaneously, we obtain *h* outputs which are heads. Then, these heads are concatenated to achieve the final output. The following demonstrates these computations in mathematical terms.

$$MultiHeadAttention(Q,K,V) = Concat(head_1, head_2, \dots, head_h).W^O \quad \text{s.t. } head_i = Attention(QW^Q, KW^K, VW^V) \tag{2}$$

MSAs, compared with CNNs, transform feature maps with huge data-specific kernels and this makes them as expressive as the CNN-based architectures [75]. The key difference exists where convolutions diversify feature maps whereas MSAs combine them. According to [76], the Fourier analysis of feature maps demonstrates that convolutions boost high-frequency components whereas MSAs, on the other hand, attenuate them.

Furthermore, finding elements that are more pertinent for the depiction of the altered posture is made easier by the multi-head attention layer. In order to do this, the scaled dot product attention gives greater weight to the characteristics of the input facial representation and encoded posture that is more pertinent while providing less weight to the features that are less relevant [77]. The procedure chooses features from various input regions and aids in improving representation performance since there are several heads in the attention layer.

In this paper, we have opted for using a ViT-based architecture for extracting features. As stated before, this policy ensures that the model is trained to attend to the most salient feature values within the identity and posture-related feature vectors simultaneously. Considering E of size R^{e} as the embedding obtained from a pre-trained facial representation learning algorithm and Vi of shape R^{v} as the bottleneck vector generated by the autoencoder part of FRA, we concatenate and reshape the produced feature vector to make it of shape $R^{(e+v)/32 \times (e+v)/32}$.

3.5. Generating Pose-Aware Face Embeddings

After the ViT module in our proposed model, the output is normalized, making it into the range of 0-1, after which it is fed to a fully connected layer. This layer helps us produce an output of a 1-d shape which is also considered the embedding generated by our model.

3.6. Multi-Task Loss Function

In the training procedure, we utilized a Multi-part Loss Function (MLF) as the learning objective. This MLF comprises a Binary Cross-Entropy (BCE) loss function, which is used to train the autoencoder, so it can reconstruct the posed style better. Since the activation function of the last layer of our autoencoder is *Sigmoid*, it can lead to loss of saturation (plateau) [78]. This saturation could prevent gradient-based learning algorithms from convergence. In order to avoid this issue, it would be better to have a *logarithm function* in the objective function to undo the *exponential function* within the *Sigmoid*. This is why *BCE* is preferred, because it uses a *logarithm function*, unlike Mean Squared Error (MSE).

The second part of our loss function is a type of *N-pair Loss* [79]. *N-pair loss* generalizes triplet loss [52] to include comparison with multiple negative samples. The objective of this function is to keep the distance between the anchor and positive smaller than the distance between the anchor and negative representations, as shown in Figure 4.

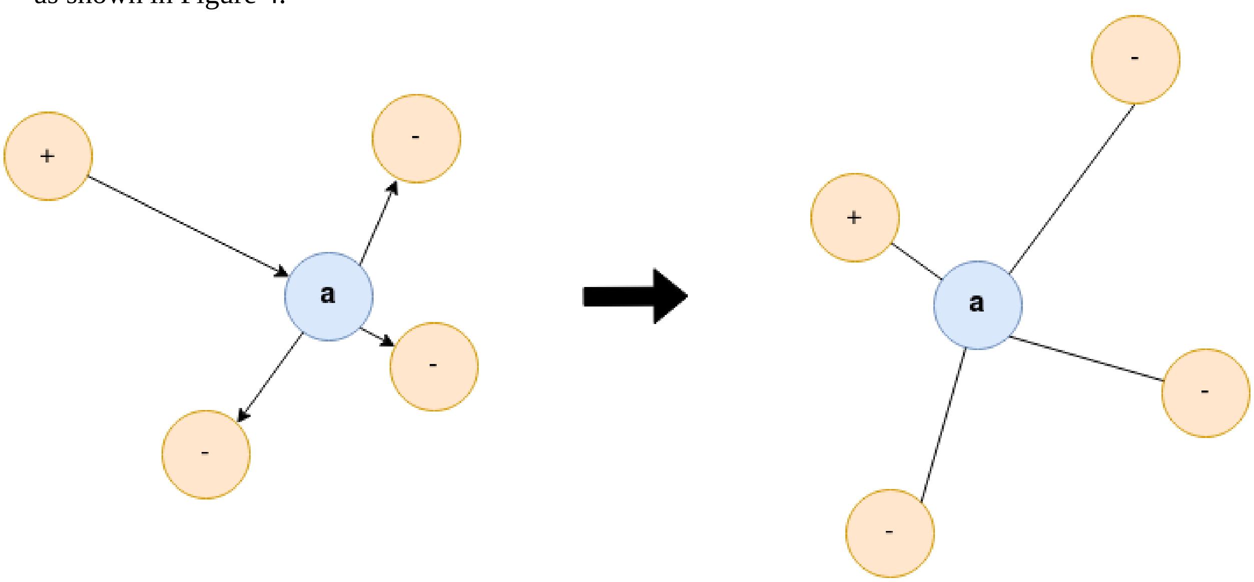


Figure 4. Effect of the proposed loss function during the learning process. The N-pair loss allows the model to distinguish between pose-variant representation vectors with the same identity and emotion, as well as possible.

The proposed multi-task loss function is defined as follows,

$$L_{total} = -\sum_{i}^{N} \left(y_i \log\log(p_i) + (1-y_i)\log\log(1-p_i)\right) + L_t\left(X^a, X^p, X^n_{pose}\right) + L_t(X^a, X^p, X^n_{identity}) + L_t(X^a, X^p, X^n_{emotion}) \quad (3)$$

in which, y_i and p_i denote the reconstructed pose style and the original pose style, respectively, and also $L_t(X^a, X^p, X^n)$ is defined as

$$L_t = \frac{1}{N}\sum_{i}^{N} max(d\left(f(X_i^a), f(X_i^p)\right) - d(f(X_i^a), f(X_i^n)) + m, 0) \quad (4)$$

where *m* is a margin applied to impose the separability between genuine and imposter pairs, and *f* denotes the proposed architecture. *d* is the euclidean distance applied on normalized features and it is given by Equation (5).

$$d(x_i, y_i) = \|x_i - y_i\|_2^2 \quad (5)$$

In Equation (6), *a* and *p* denote the anchor (generated) representation and the positive (real) representation, respectively. Additionally, X^n_{pose}, $X^n_{identity}$, and $X^n_{emotion}$ denote the negative representation w.r.t pose, identity, and emotion of the anchor face, respectively. Specifically, negative pose representations have the same identity as the anchor, but with different poses. The same holds for negative emotion representation. But, for negative identity representation, the representation of another person is chosen randomly, regardless of what pose or emotion it has. The goal of the triplet loss is to achieve,

$$d(f(X_i^a), f(X_i^n)) > d\left(f(X_i^a), f(X_i^p)\right) + m \quad (6)$$

The optimal state for each single triplet loss is achieved when $d\left(f(X_i^a), f(X_i^p)\right)$ is equal to zero and $d(f(X_i^a), f(X_i^n))$ is greater than the predefined margin.

4. Results and Discussion

In this section we first introduce the benchmark dataset that we have used for evaluating our proposed method. Then, we elaborate the details of our implementation and introduce the metrics used in this paper. Finally, we demonstrate our experimental results and discussion.

4.1. Datasets

With the object of benchmarking our results, we have used the KDEF dataset. It is a publicly available dataset of 4900 face images, covering 140 unique identities. The images demonstrate face images with varying pose and emotion styles. Some samples of these datasets are shown in Figure 5.

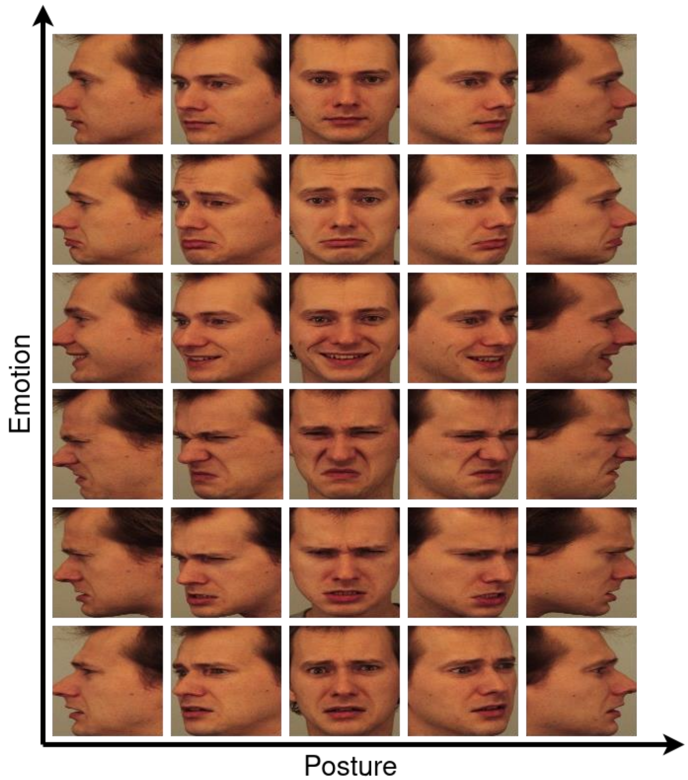


Figure 5. A few samples of the KDEF dataset. The KDEF dataset provides face images of 140 different people in various postures and emotions.

4.2. Implementation details

We carried out our experiments on a machine with a Core i7-1165G7 @ 2.80GHz CPU with 64 Gigabytes of RAM and a GeForce RTX 2060 12 GB GPU. All models were implemented and trained using the Pytorch framework. Table 2 shows the hyperparameter setting.

Table 2. Details of the training procedure and the utilized FRLs. The hyperparameter settings are shown.

<table>
<tr><th rowspan="2">FRL arch.</th><th rowspan="2"># epochs</th><th rowspan="2">Init. learning rate</th><th rowspan="2">Dropout rate</th><th rowspan="2">Triplet margin</th><th colspan="5">ViT</th></tr>
<tr><th>Embedding dim</th><th>FC dim</th><th># Heads</th><th># Layers</th><th>Patch size</th></tr>
<tr><td>MagFace</td><td>255</td><td>0.001</td><td>0.4</td><td>10.0</td><td rowspan="3">256</td><td rowspan="3">256</td><td rowspan="3">4</td><td rowspan="3">4</td><td rowspan="3">8</td></tr>
<tr><td>ArcFace</td><td>320</td><td>0.001</td><td>0.4</td><td>10.0</td></tr>
<tr><td>CosFace</td><td>157</td><td>0.001</td><td>0.05</td><td>10.0</td></tr>
</table>

Furthermore, with the object of fairly evaluating the proposed FRA, we divided KDEF datasets based on identities with the following distributions:

- We randomly selected all samples from 99 identities which nearly comprise 70.7 % of all identities in KDEF as our training data.
- We randomly selected all samples from 11 identities which nearly comprise 7.8 % of all identities in KDEF as our validation data.
- We randomly selected all samples from 30 identities which nearly comprise 21.5 % of all identities in KDEF as our testing data.

4.3. Experimental Results

This section details our comprehensive experimental results. Table 3 shows the achieved accuracy of the Support Vector Machine (SVM) [80] classifier on the embeddings generated in three different experiments. These experiments are:

(1) Pre-augmentation accuracy: In this experiment, the training happens on the embeddings extracted using three different FRLs, namely MagFace, ArcFace, and CosFace, and the testing accuracy is achieved on the testing partition of these embeddings (Train/Test split ratio is set 80/20). In this experiment, we used no augmentation technique at all and this is done to find a baseline for the quality of the original data in the chosen benchmark dataset.
(2) Generated embeddings' accuracy: In this experiment, we first augmented the original embeddings to obtain the transformed embeddings. Then, we trained the SVM on the original data and tested its performance on the generated embeddings by the proposed algorithm. This is done to demonstrate how much the proposed model is able to sustain the identity, posture, and emotion-related features without any degradation.
(3) Post-augmentation accuracy: In this experiment, we have augmented the embeddings of the training split using FRA, where the test split is the same as (1). Then, we trained the SVM classifier on the training part and tested it on the testing one.

Table3. Evaluation results of FRA. Pre-augmentation and post-augmentation accuracies show the effectiveness of FRA. Generated embeddings' accuracy denotes the sustainability of FRA.

FRL	**Target**	**(1) Pre-augmentation Accuracy (%)**	**(2) Generated embeddings' Accuracy (%)**	**(3) Post-augmentation Accuracy (%)**
MagFace	Posture	82.38	98.12	96.66
	Identity	86.19	93.61	95.71
	Emotion	44.76	92.43	99.04
	Posture	89.12	99.3	97.9

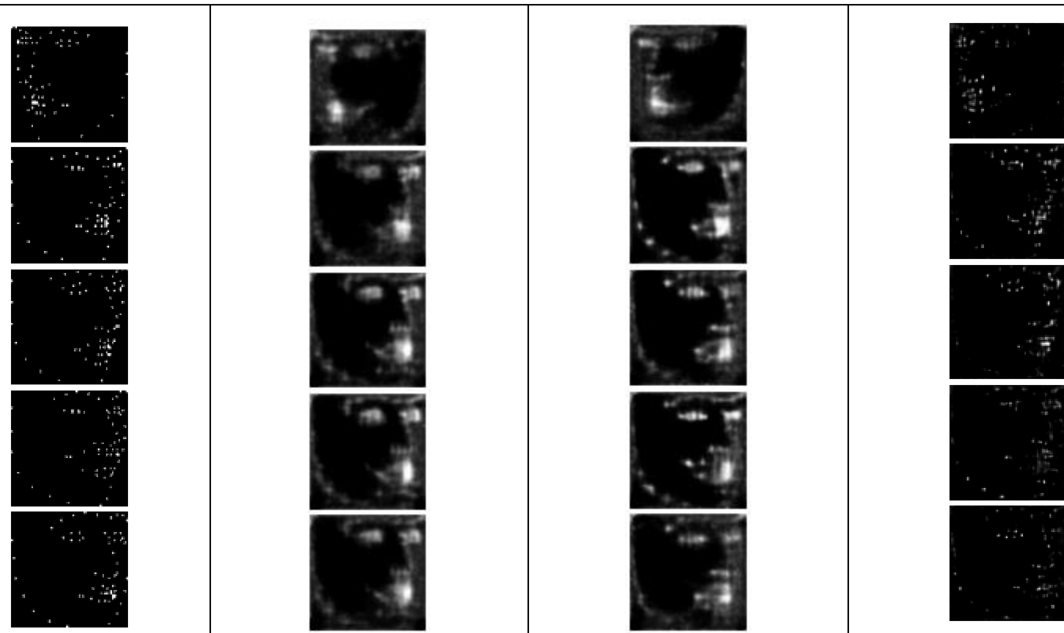
ArcFace	Identity	86.61	91.6	96.65
	Emotion	53.55	92.98	100
CosFace	Posture	99.12	99.91	99.12
	Identity	79.91	88.11	96.50
	Emotion	54.14	87.61	97.37

Based on Table 3, it is observed that our proposed algorithm improves the classification accuracy, not only identity-wise but also in terms of emotion and posture. For instance, SVM outputs 86.19% accuracy on the MagFace embeddings but after the augmentation, this score goes up to 95.71% in Post-augmentation. For the same data, the generated embeddings are much more representative which improves the classification accuracy on the identity of the SVM outputs by 93.61%. This enhancement can also be validated for ArcFace and CosFace since our algorithm increases the accuracy in all three experiments. In addition, FRA can improve the accuracy of SVM embeddings remarkably. In addition to improving the classification accuracy with respect to the identities of the embeddings, FRA improves that pose and emotion-wise. Based on Table 5, the accuracy of the SVM classifier is increased from 86.19% for MagFace embeddings to 95.71% after augmentation. This improvement for ArcFace and Cosface is from 86.61% to 96.65% and from 79.91% to 97.37%, respectively.

Furthermore, our generated embeddings should ensure the fact that they are linearly separable. This means that the embeddings can be classified using a linear classifier such as SVM with a linear kernel. In our experiments we used an SVM classifier with a linear kernel and based on Table 5, we can deduce that FRA is able to improve the accuracy in Phase 2 and Phase 3 by a large margin, effectively enhancing the performance of SVM with a linear kernel.

Moreover, in order to show the independence of FRA from any FR algorithms, we adopted three such approaches, namely MagFace, ArcFace, and CosFace. Based on Table 5, after augmenting the embeddings generated by each of these algorithms, the classification accuracy of the SVM classifier is increased significantly and this proves the fact that FRA is not dependent on any specific FR algorithm as its requirements.

In addition, for our algorithms' performance to be verified thoroughly, the reconstructed binary images which are created by the AE part of the proposed approach are presented. These images are illustrated in Figure 6, showing the original image, and the AE's output when dealing with MagFace embeddings, ArcFace embeddings, and CosFace embeddings. Also, the training and validation loss in the training procedure of our proposed pipeline is shown in Figure 7.



Original landmarks	AE-generated landmarks for Magface FRL	AE-generated landmarks for Arcface FRL	AE-generated landmarks for Cosface FRL

Figure 6. Some instances of reconstructed binarized facial landmark images. These reconstructed images denote how good the AE is performing for each FRL model. The CosFace model reconstructs the most landmarks more precisely.

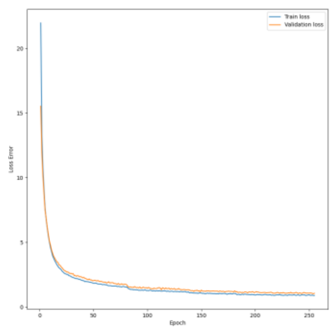
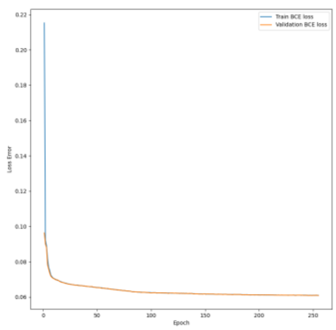
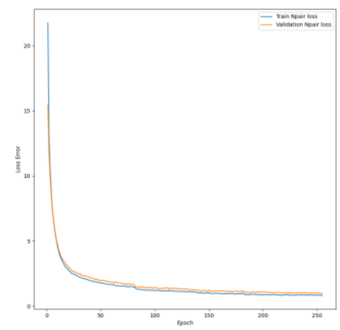
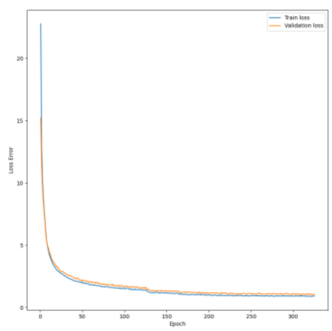
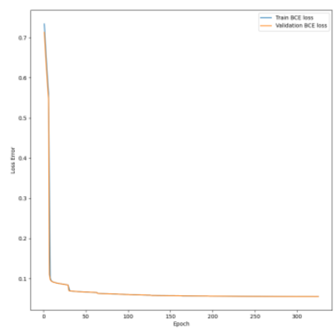
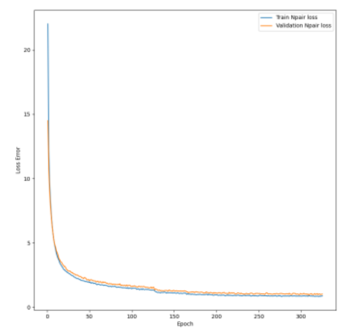
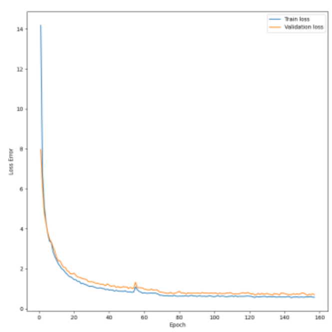
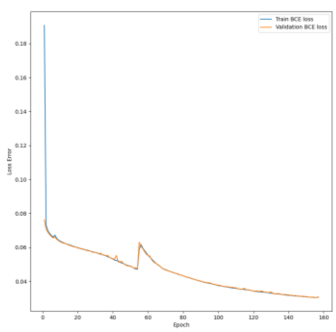
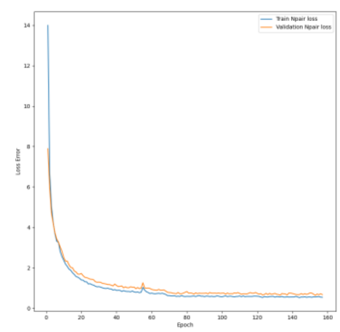
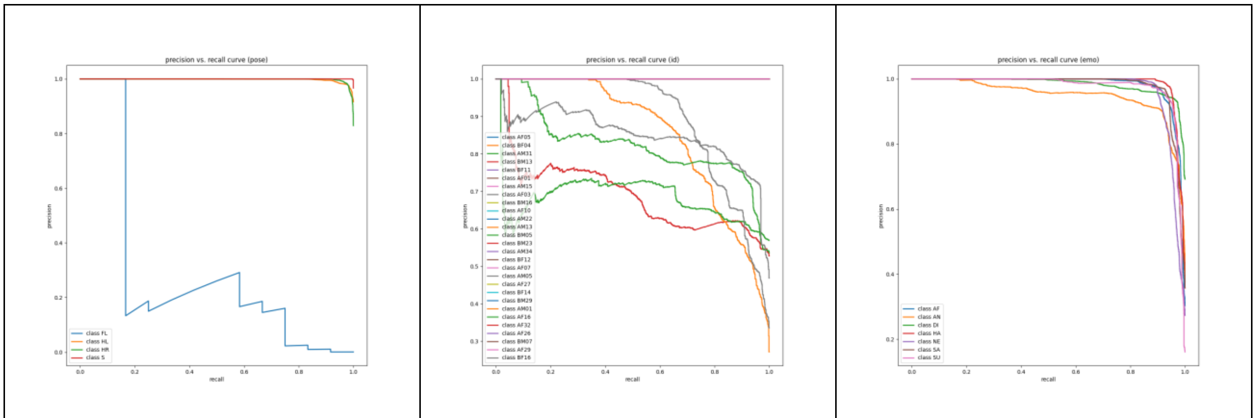
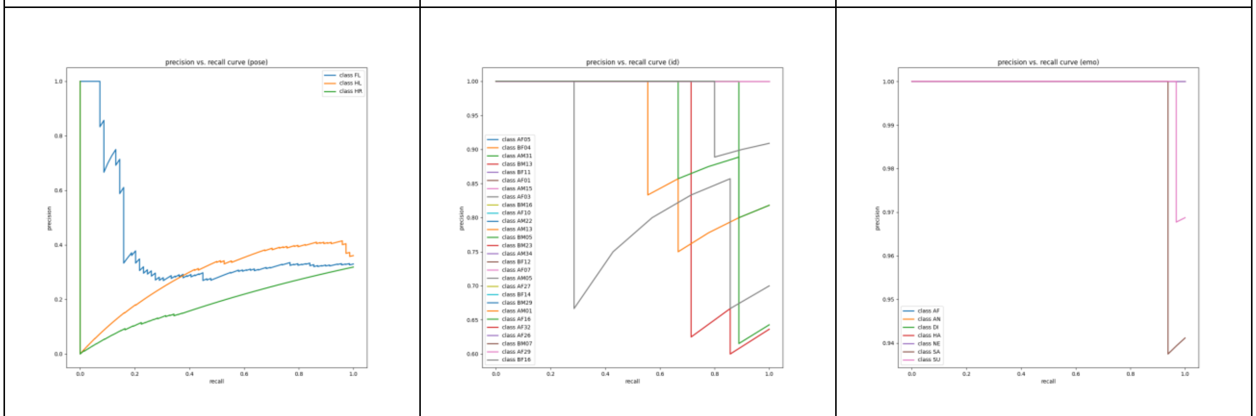
		
Total Loss Magface	BCE Loss Magface	Npair Loss Magface
		
Total Loss Arcface	BCE Loss Arcface	Npair Loss Arcface
		
Total Loss Cosface	BCE Loss Cosface	Npair Loss Cosface

Figure 7. Total loss, BCE loss and Npair loss curves achieved by various FRL methods.

The curves of Precision-Recall (PRC), Receiver Operating Characteristic (ROC), and confusion matrices of all experiments for MagFace, ArcFace, and CosFace are shown in Figures 8-16.



a) Pre-augmentation PRC (pose) | b) Pre-augmentation PRC (id) | c) Pre-augmentation PRC (emotion)



d) Post-augmentation PRC (pose) | e) Post-augmentation PRC (id) | f) Post-augmentation PRC (emotion)

Figure 8. PRC curves for MagFace FRL. As the curves indicate, post-augmentation PRC curves (d, e, and f) have significant improvements in comparison to pre-augmentation PRC curves (a, b, and c) for pose, identity, and emotion, respectively.

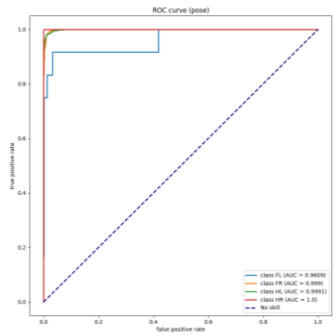
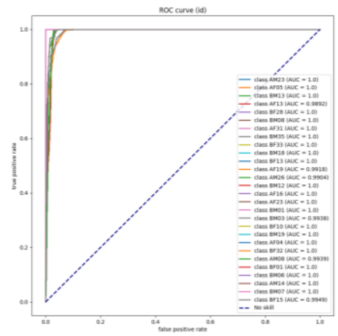
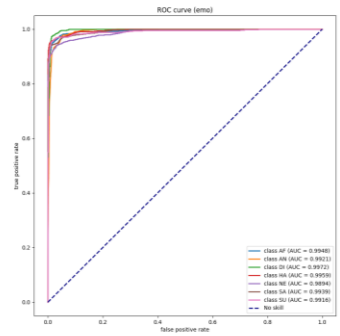
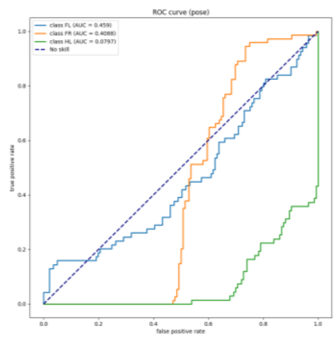
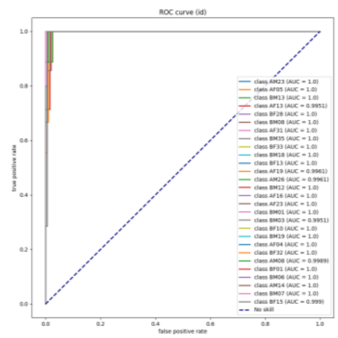
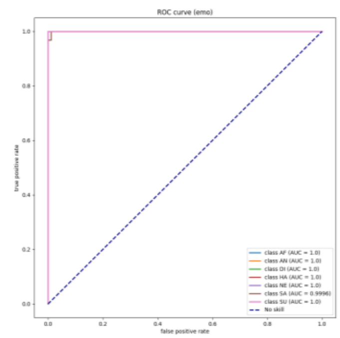
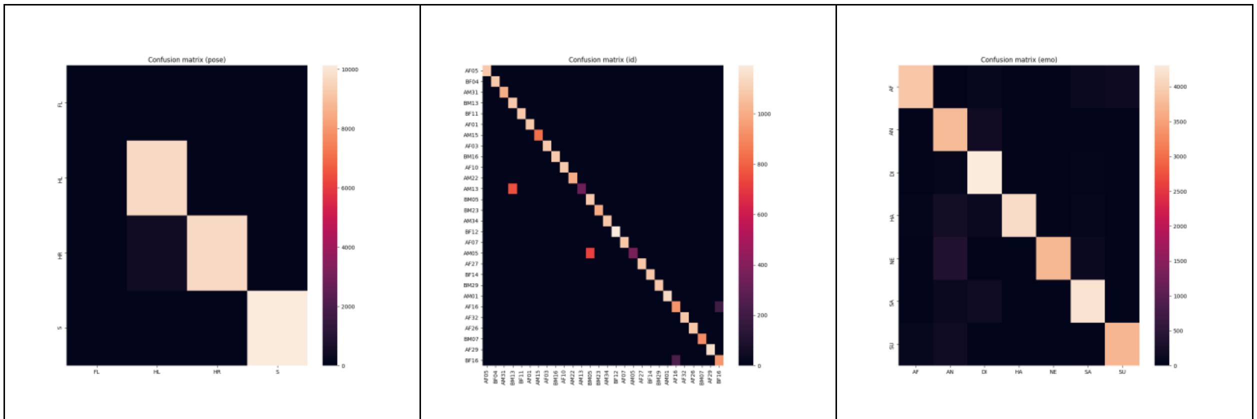
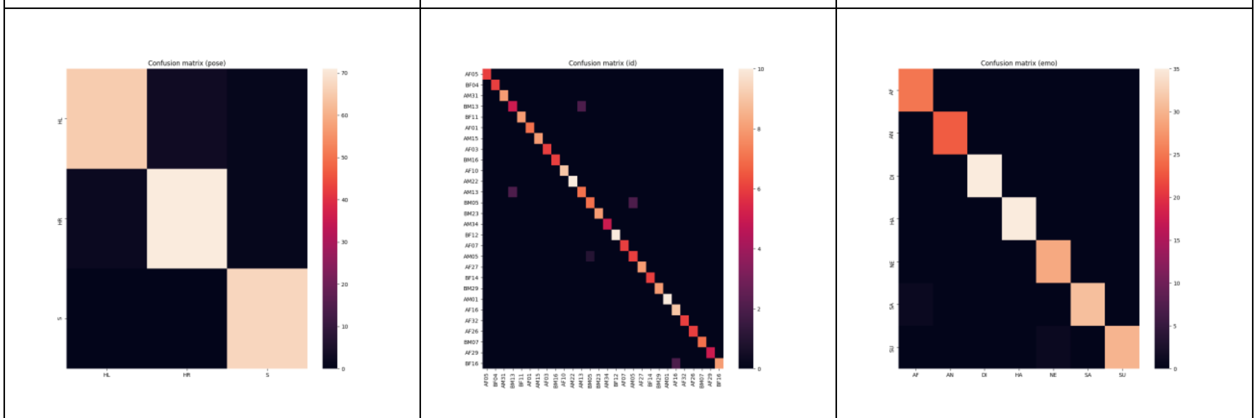
		
a) Pre-augmentation ROC (pose)	b) Pre-augmentation ROC (id)	c) Pre-augmentation ROC (emotion)
		
d) Post-augmentation ROC (pose)	e) Post-augmentation ROC (id)	f) Post-augmentation ROC (emotion)

Figure 9. ROC curves for MagFace FRL. As the curves indicate, post-augmentation ROC curves (d, e, and f) have significant improvements in comparison to pre-augmentation ROC curves (a, b, and c) for pose, identity, and emotion, respectively.

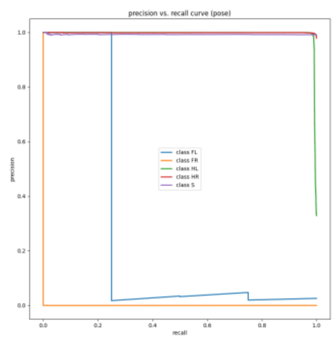


a) Pre-augmentation Confusion (pose)	b) Pre-augmentation Confusion (id)	c) Pre-augmentation Confusion (emotion)

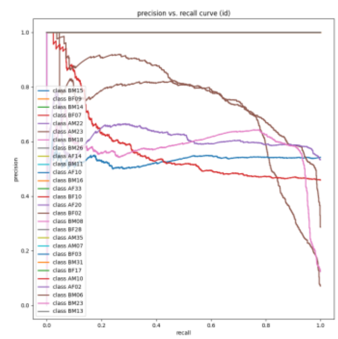


d) Post-augmentation Confusion (pose)	e) Post-augmentation Confusion (id)	f) Post-augmentation Confusion (emotion)

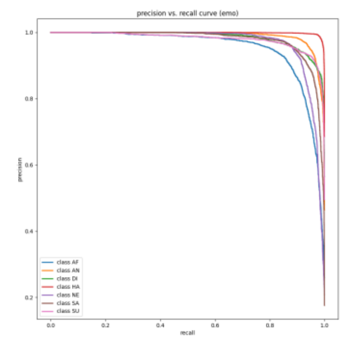
Figure 10. Confusion matrices for MagFace FRL. As the confusion matrices indicate, post-augmentation matrices (d, e, and f) have significant improvements in comparison to pre-augmentation matrices (a, b, and c) for pose, identity, and emotion, respectively.



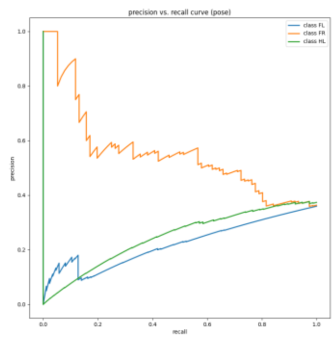
a) Pre-augmentation PRC (pose)



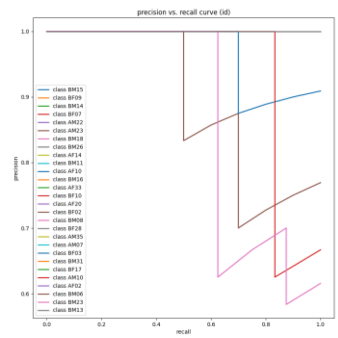
b) Pre-augmentation PRC (id)



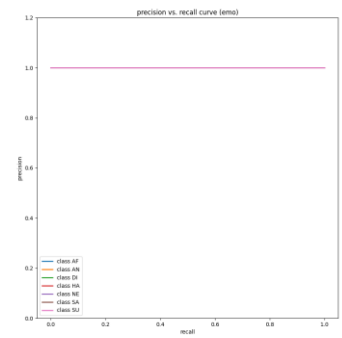
c) Pre-augmentation PRC (emotion)



d) Post-augmentation PRC (pose)

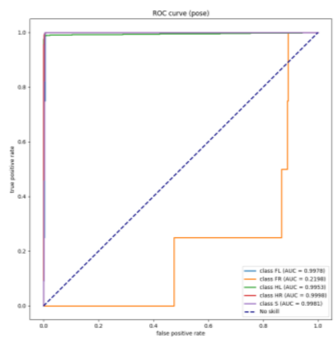


e) Post-augmentation PRC (id)

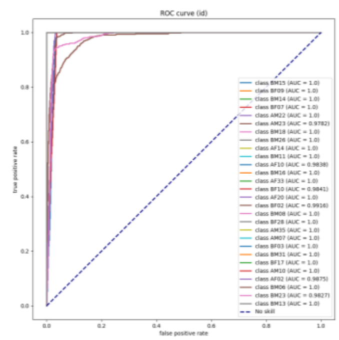


f) Post-augmentation PRC (emotion)

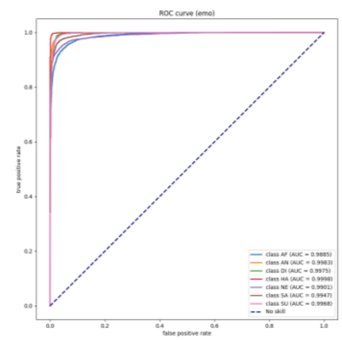
Figure 11. PRC curves for ArcFace FRL. As the curves indicate, post-augmentation PRC curves (d, e, and f) have significant improvements in comparison to pre-augmentation PRC curves (a, b, and c) for pose, identity, and emotion, respectively.



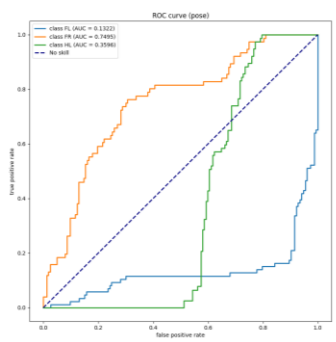
a) Pre-augmentation ROC (pose)



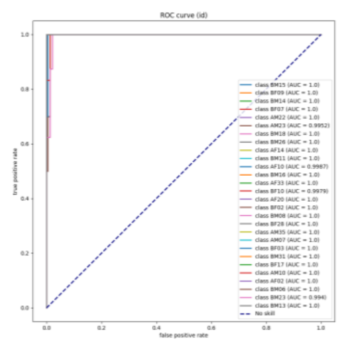
b) Pre-augmentation ROC (id)



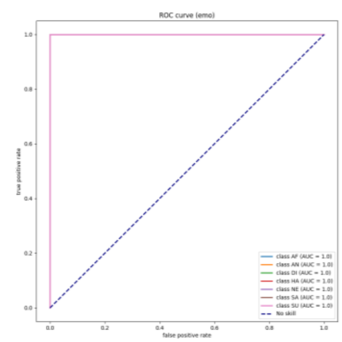
c) Pre-augmentation ROC (emotion)



d) Post-augmentation ROC (pose)

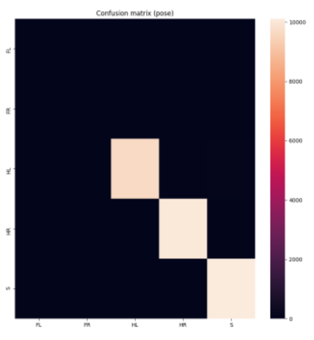


e) Post-augmentation ROC (id)

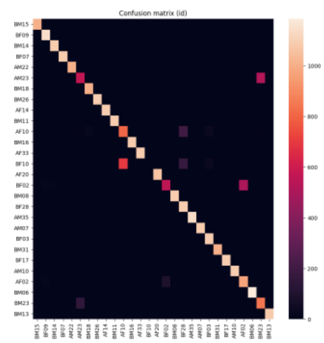


f) Post-augmentation ROC (emotion)

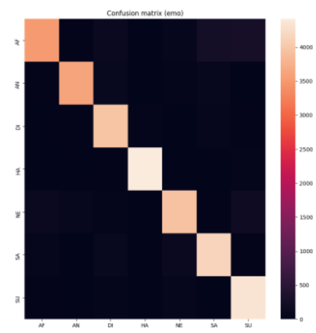
Figure 12. ROC curves for ArcFace FRL. As the curves indicate, post-augmentation ROC curves (d, e, and f) have significant improvements in comparison to pre-augmentation ROC curves (a, b, and c) for pose, identity, and emotion, respectively.



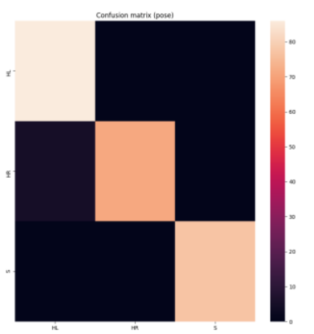
a) Pre-augmentation Confusion (pose)



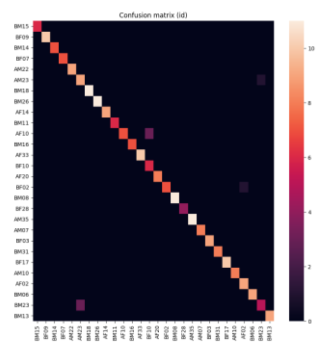
b) Pre-augmentation Confusion (id)



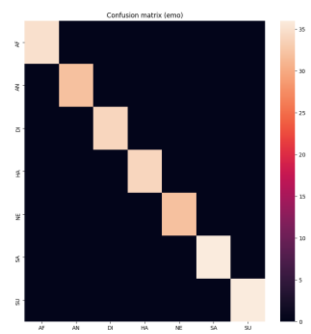
c) Pre-augmentation Confusion (emotion)



d) Post-augmentation Confusion (pose)

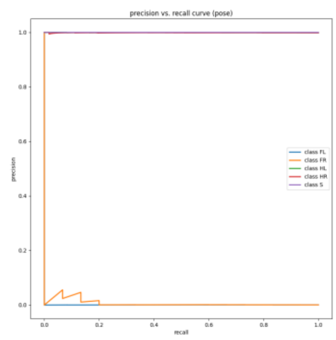


e) Post-augmentation Confusion (id)

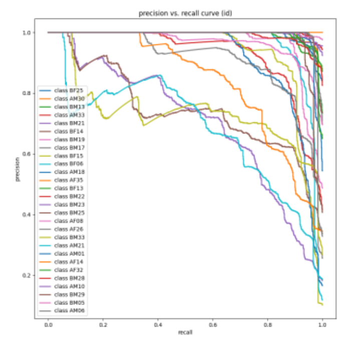


f) Post-augmentation Confusion (emotion)

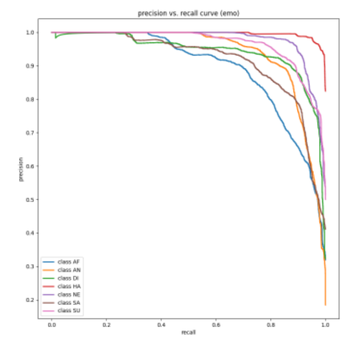
Figure 13. Confusion matrices for ArcFace FRL. As the confusion matrices indicate, post-augmentation matrices (d, e, and f) have significant improvements in comparison to pre-augmentation matrices (a, b, and c) for pose, identity, and emotion, respectively.



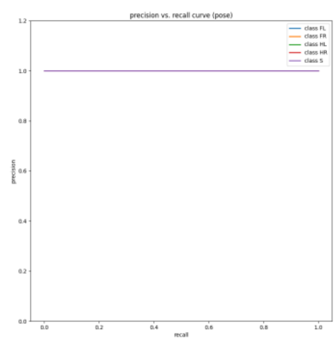
a) Pre-augmentation PRC (pose)



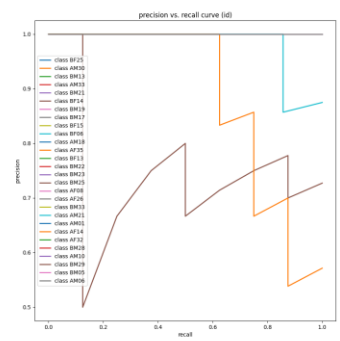
b) Pre-augmentation PRC (id)



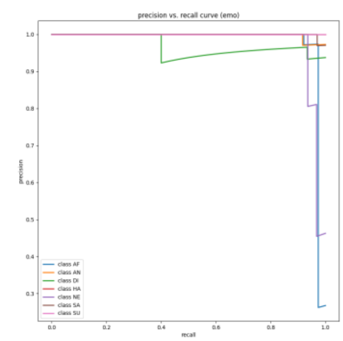
c) Pre-augmentation PRC (emotion)



d) Post-augmentation PRC (pose)



e) Post-augmentation PRC (id)



f) Post-augmentation PRC (emotion)

Figure 14. PRC curves for CosFace FRL. As the curves indicate, post-augmentation PRC curves (d, e, and f) have significant improvements in comparison to pre-augmentation PRC curves (a, b, and c) for pose, identity, and emotion, respectively.

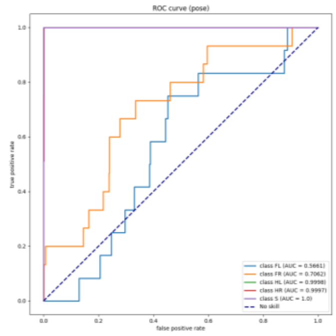
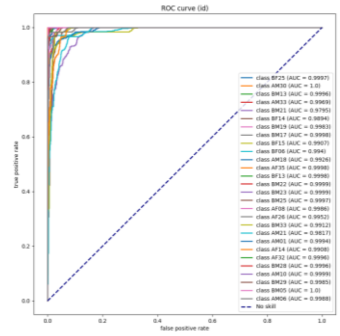
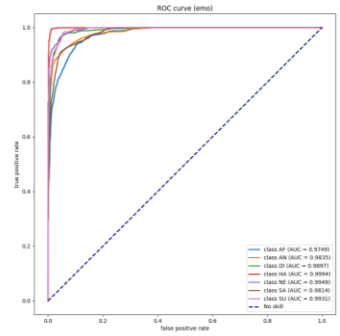
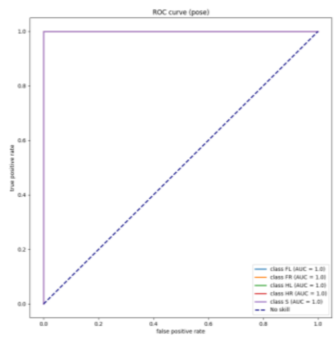
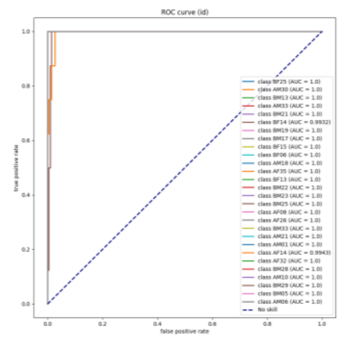
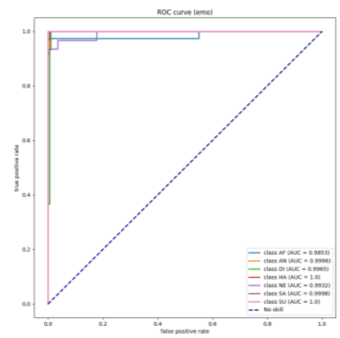
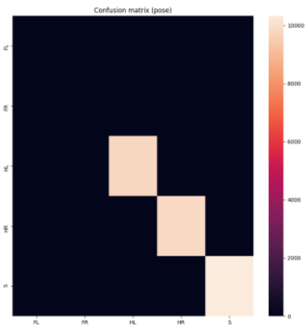
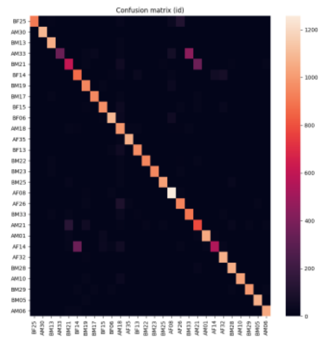
		
a) Pre-augmentation ROC (pose)	b) Pre-augmentation ROC (id)	c) Pre-augmentation ROC (emotion)
		
d) Post-augmentation ROC (pose)	e) Post-augmentation ROC (id)	f) Post-augmentation ROC (emotion)

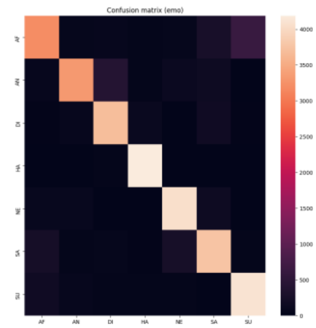
Figure 15. ROC curves for CosFace FRL. As the curves indicate, post-augmentation ROC curves (d, e, and f) have significant improvements in comparison to pre-augmentation ROC curves (a, b, and c) for pose, identity, and emotion, respectively.



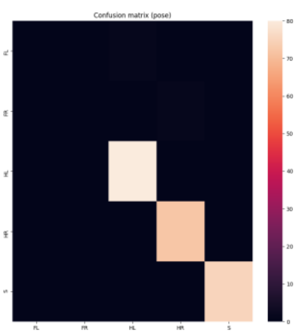
a) Pre-augmentation Confusion (pose)



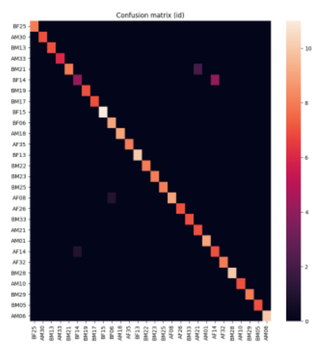
b) Pre-augmentation Confusion (id)



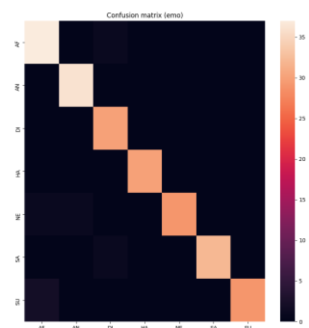
c) Pre-augmentation Confusion (emotion)



d) Post-augmentation Confusion (pose)



e) Post-augmentation Confusion (id)



f) Post-augmentation Confusion (emotion)

Figure 16. Confusion matrices for CosFace FRL. As the confusion matrices indicate, post-augmentation matrices (d, e, and f) have significant improvements in comparison to pre-augmentation matrices (a, b, and c) for the pose, identity, and emotion, respectively.

4.4. Discussion

FR has long been a popular field of study among specialists and academics in the field of biometric recognition and it has the advantages of being non-contact, amiable, and simple to accept. Although remarkable performance has been shown by some state-of-the-art approaches presented in the literature, in real-world scenarios, there still exists the demand to improve such algorithms. For better handling such uncontrolled contexts especially when we face a lack of data, DA techniques are introduced to increase the number of training samples by applying different manipulations. Classical techniques for image transformations such as rotation, skewing, flipping, blurring, etc., and also GAN-based ones which utilize deep generative models and disentangled features to create more realistically transformed face images have well been studied for DA in the domain of FR. However, these techniques have their drawbacks. Classical techniques mostly manipulate face images in a way that distorts their alignment causing FR algorithms' performance when generating distinct representative embeddings dramatically decrease. To prove this, we have experimented with four different transformations, namely, horizontal flip, skewing, blurring, and notifying. We augmented samples in the KDEF dataset and increased the training dataset size to be 4 times more than the original dataset and used different FR algorithms for generating embeddings. Then we classified the embeddings using SVM with respect to their identities. Table 6 details the results achieved by this experiment.

Table 4. Evaluation of MagFace, ArcFace, and CosFace using traditional augmentation techniques on the KDEF dataset. Post-augmentation accuracy scores achieved denote the ineffectiveness of these techniques for FR tasks.

FR Algorithm	**Accuracy Score**	
	Pre-augmentation	**Post-augmentation**
MagFace	20.49%	18.54%
ArcFace	20.12%	17.01%
CosFace	18.19%	11.33%

Table 4 shows that augmenting the face images using the classical approaches does not result in any improvement and they, in fact, degrade the quality of embeddings. For instance, for the MagFace algorithm, the accuracy obtained by SVM is decreased by 2% after applying DA.
Moreover, we conducted another experiment using three state-of-the-art generative-based algorithms namely, CPM [41], AttGAN [81], and PSGAN [40]. Following the previous experiments, we augmented the face images and obtained the classification accuracy before and after augmentation. Table 7 shows the results achieved by this experiment.

Table 5. Evaluation of augmentation techniques using GANs on the KDEF dataset. In the best case, the post-augmentation accuracy has increased a little and in some cases, it has caused a degradation in accuracy.

Algorithm	**Accuracy Score**	
	Pre-augmentation	**Post-augmentation**
CPM [40]	20.49 %	17.82 %
AttGAN	20.49 %	26.70 %
PSGAN [39]	20.49 %	23.40 %

Based on Table 5, it can be claimed that these generative models do not contribute to classification accuracy improvement. Therefore, in order to address this issue, in this paper, we propose a new algorithm, named FRA, which effectively augments the training data for FR algorithms. FRA functions with original embeddings and manipulates them in a way to be representative of the same identity of the embedding and also a differed postural information existent in these representational embeddings. Results achieved by our extensive experiments indicate the efficacy of FRA in augmenting samples in the FR domain.

5. Conclusion

Since data scarcity is a common problem in deep learning-based solutions, it can be very challenging to build up FR systems that are robust enough to recognize face images with extreme diversity. In this paper, we proposed a novel method that augments the face data in latent space. The proposed method utilizes two major components, one of which is an autoencoder and the other is a ViT-based model. The former is used to encode the binarized input images consisting of sparse facial landmarks into a latent space. The latter is used for extracting features from the combined embeddings coming from a pre-trained FRL algorithm and the autoencoder part of our model. Lastly, the output of the proposed model is an embedding representing the main identity with the same emotion but with a different posture. This way, we managed to improve the classification accuracy by 9.52, 10.04, and 16.60, in comparison with the based models of MagFace, ArcFace, and CosFace, respectively.

References

[1] P. Terhorst, J.N. Kolf, N. Damer, F. Kirchbuchner, A. Kuijper, SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2020, pp. 5651-5660.

[2] M. Wang, W. Deng, Deep face recognition: A survey, Neurocomputing, 429 (2021) 215-244.

[3] W. Ali, W. Tian, S.U. Din, D. Iradukunda, A.A. Khan, Classical and modern face recognition approaches: a complete review, Multimedia tools and applications, 80 (2021) 4825-4880.

[4] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM, 60 (2017) 84-90.

[5] Y. Taigman, M. Yang, M.A. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, Proceedings of the IEEE conference on computer vision and pattern recognition2014, pp. 1701-1708.

[6] Q. Meng, S. Zhao, Z. Huang, F. Zhou, Magface: A universal representation for face recognition and quality assessment, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2021, pp. 14225-14234.

[7] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE2018, pp. 67-74.

[8] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, B. Zheng, Age-invariant face recognition by multi-feature fusionand decomposition with self-attention, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18 (2022) 1-18.

[9] Y. Zheng, D.K. Pal, M. Savvides, Ring loss: Convex feature normalization for face recognition, Proceedings of the IEEE conference on computer vision and pattern recognition2018, pp. 5089-5097.

[10] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, Proceedings of the IEEE International Conference on Computer Vision2017, pp. 5409-5418.

[11] C. Peng, N. Wang, J. Li, X. Gao, DLFace: Deep local descriptor for cross-modality face recognition, Pattern Recognition, 90 (2019) 161-171.

[12] A. Bhattacharyya, S. Chatterjee, S. Sen, A. Sinitca, D. Kaplun, R. Sarkar, A deep learning model for classifying human facial expressions from infrared thermal images, Scientific Reports, 11 (2021) 20696.

[13] S. Gupta, K. Thakur, M. Kumar, 2D-human face recognition using SIFT and SURF descriptors of face's feature regions, The Visual Computer, 37 (2021) 447-456.

[14] Y. Kortli, M. Jridi, A. Al Falou, M. Atri, A comparative study of CFs, LBP, HOG, SIFT, SURF, and BRIEF for security and face recognition, Advanced Secure Optical Image Processing for Communications, IOP Publishing2018.

[15] Y. Duan, J. Lu, J. Zhou, Uniformface: Learning deep equidistributed representation for face recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2019, pp. 3415-3424.

[16] B. Lahasan, S.L. Lutfi, R. San-Segundo, A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression, Artificial Intelligence Review, 52 (2019) 949-979.
[17] Y. Li, K. Guo, Y. Lu, L. Liu, Cropping and attention based approach for masked face recognition, Appl. Intell., 51 (2021) 3012-3025.
[18] P. Li, L. Prieto, D. Mery, P.J. Flynn, On low-resolution face recognition in the wild: Comparisons and new techniques, IEEE Transactions on Information Forensics and Security, 14 (2019) 2000-2012.
[19] L. Song, D. Gong, Z. Li, C. Liu, W. Liu, Occlusion robust face recognition based on mask learning with pairwise differential siamese network, Proceedings of the IEEE/CVF International Conference on Computer Vision2019, pp. 773-782.
[20] Z. Lu, X. Jiang, A. Kot, Deep coupled resnet for low-resolution face recognition, IEEE Signal Processing Letters, 25 (2018) 526-530.
[21] J. Deng, S. Zafeririou, Arcface for disguised face recognition, Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops2019, pp. 0-0.
[22] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2019, pp. 4690-4699.
[23] Y. Srivastava, V. Murali, S.R. Dubey, A performance evaluation of loss functions for deep face recognition, Computer Vision, Pattern Recognition, Image Processing, and Graphics: 7th National Conference, NCVPRIPG 2019, Hubballi, India, December 22–24, 2019, Revised Selected Papers 7, Springer2020, pp. 322-332.
[24] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, F. Huang, Curricularface: adaptive curriculum learning loss for deep face recognition, proceedings of the IEEE/CVF conference on computer vision and pattern recognition2020, pp. 5901-5910.
[25] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, Proceedings of the IEEE conference on computer vision and pattern recognition2018, pp. 5265-5274.
[26] X. Wang, S. Wang, J. Wang, H. Shi, T. Mei, Co-mining: Deep face recognition with noisy labels, Proceedings of the IEEE/CVF International Conference on Computer Vision2019, pp. 9358-9367.
[27] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, Proceedings of the IEEE international conference on computer vision2015, pp. 3730-3738.
[28] S.C. Wong, A. Gatt, V. Stamatescu, M.D. McDonnell, Understanding data augmentation for classification: when to warp?, 2016 international conference on digital image computing: techniques and applications (DICTA), IEEE2016, pp. 1-6.
[29] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, X. Zhou, Data augmentation for face recognition, Neurocomputing, 230 (2017) 184-196.
[30] I. Masi, A.T. Trần, T. Hassner, G. Sahin, G. Medioni, Face-specific data augmentation for unconstrained face recognition, International Journal of Computer Vision, 127 (2019) 642-667.
[31] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, W. Gao, Efficient 3D reconstruction for face recognition, Pattern Recognition, 38 (2005) 787-798.
[32] H. Mohammadzade, D. Hatzinakos, Projection into expression subspaces for face recognition from single sample per person, IEEE Transactions on Affective Computing, 4 (2012) 69-82.
[33] Y. Shen, P. Luo, J. Yan, X. Wang, X. Tang, Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis, Proceedings of the IEEE conference on computer vision and pattern recognition2018, pp. 821-830.
[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM, 63 (2020) 139-144.
[35] D. Lundqvist, A. Flykt, A. Öhman, Karolinska directed emotional faces, Cognition and Emotion, (1998).
[36] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, L. Bottou, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of machine learning research, 11 (2010).
[37] J. Wang, L. Perez, The effectiveness of data augmentation in image classification using deep learning, Convolutional Neural Networks Vis. Recognit, 11 (2017) 1-8.
[38] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, International conference on machine learning, PMLR2017, pp. 1857-1865.
[39] I. Kemelmacher-Shlizerman, Transfiguring portraits, ACM Transactions on Graphics (TOG), 35 (2016) 1-8.
[40] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, S. Yan, Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2020, pp. 5194-5202.

[41] T. Nguyen, A.T. Tran, M. Hoai, Lipstick ain't enough: beyond color matching for in-the-wild makeup transfer, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2021, pp. 13305-13314.
[42] C.-Y. Wu, R. Manmatha, A.J. Smola, P. Krahenbuhl, Sampling matters in deep embedding learning, Proceedings of the IEEE international conference on computer vision2017, pp. 2840-2848.
[43] Y. Movshovitz-Attias, A. Toshev, T.K. Leung, S. Ioffe, S. Singh, No fuss distance metric learning using proxies, Proceedings of the IEEE international conference on computer vision2017, pp. 360-368.
[44] E. Ustinova, V. Lempitsky, Learning deep embeddings with histogram loss, Advances in Neural Information Processing Systems, 29 (2016).
[45] Q. Xiao, H. Luo, C. Zhang, Margin sample mining loss: A deep learning based method for person re-identification, arXiv preprint arXiv:1710.00478, (2017).
[46] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, Proceedings of the IEEE conference on computer vision and pattern recognition2017, pp. 403-412.
[47] A. Sanakoyeu, V. Tschernezki, U. Buchler, B. Ommer, Divide and conquer the embedding space for metric learning, Proceedings of the ieee/cvf conference on computer vision and pattern recognition2019, pp. 471-480.
[48] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, L.-Y. Duan, Group-sensitive triplet embedding for vehicle reidentification, IEEE Transactions on Multimedia, 20 (2018) 2385-2399.
[49] Y. Em, F. Gag, Y. Lou, S. Wang, T. Huang, L.-Y. Duan, Incorporating intra-class variance to fine-grained visual recognition, 2017 IEEE International Conference on Multimedia and Expo (ICME), IEEE2017, pp. 1452-1457.
[50] E. Smirnov, A. Oleinik, A. Lavrentev, E. Shulga, V. Galyuk, N. Garaev, M. Zakuanova, A. Melnikov, Face representation learning using composite mini-batches, Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops2019, pp. 0-0.
[51] Y. Shi, X. Yu, K. Sohn, M. Chandraker, A.K. Jain, Towards universal representation learning for deep face recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2020, pp. 6817-6826.
[52] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, Proceedings of the IEEE conference on computer vision and pattern recognition2015, pp. 815-823.
[53] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, Proceedings of the IEEE conference on computer vision and pattern recognition2017, pp. 212-220.
[54] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, Workshop on faces in'Real-Life'Images: detection, alignment, and recognition2008.
[55] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, CVPR 2011, IEEE2011, pp. 529-534.
[56] T. Zheng, W. Deng, J. Hu, Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments, arXiv preprint arXiv:1708.08197, (2017).
[57] T. Zheng, W. Deng, Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments, Beijing University of Posts and Telecommunications, Tech. Rep, 5 (2018).
[58] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou, Agedb: the first manually collected, in-the-wild age database, proceedings of the IEEE conference on computer vision and pattern recognition workshops2017, pp. 51-59.
[59] S. Sengupta, J.-C. Chen, C. Castillo, V.M. Patel, R. Chellappa, D.W. Jacobs, Frontal to profile face verification in the wild, 2016 IEEE winter conference on applications of computer vision (WACV), IEEE2016, pp. 1-9.
[60] K. Cong, M. Zhou, Face Dataset Augmentation with Generative Adversarial Network, Journal of Physics: Conference Series, IOP Publishing2022, pp. 012035.
[61] A. Kammoun, R. Slama, H. Tabia, T. Ouni, M. Abid, Generative Adversarial Networks for face generation: A survey, ACM Computing Surveys, 55 (2022) 1-37.
[62] G. Bradski, The openCV library, Dr. Dobb's Journal: Software Tools for the Professional Programmer, 25 (2000) 120-123.
[63] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, California Univ San Diego La Jolla Inst for Cognitive Science1985.
[64] D. Bank, N. Koenigstein, R. Giryes, Autoencoders, arXiv preprint arXiv:2003.05991, (2020).
[65] Q. Meng, X. Xu, X. Wang, Y. Qian, Y. Qin, Z. Wang, C. Zhao, F. Zhou, Z. Lei, PoseFace: Pose-invariant features and pose-adaptive loss for face recognition, arXiv preprint arXiv:2107.11721, (2021).
[66] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open, (2022).

[67] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer2020, pp. 213-229.
[68] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, International conference on machine learning, PMLR2021, pp. 10347-10357.
[69] F. Yang, H. Yang, J. Fu, H. Lu, B. Guo, Learning texture transformer network for image super-resolution, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2020, pp. 5791-5800.
[70] L. Ye, M. Rochan, Z. Liu, Y. Wang, Cross-modal self-attention network for referring image segmentation, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2019, pp. 10502-10511.
[71] A.M. Hafiz, S.A. Parah, R.U.A. Bhat, Attention mechanisms and deep learning for machine vision: A survey of the state of the art, arXiv preprint arXiv:2106.07550, (2021).
[72] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing, 452 (2021) 48-62.
[73] F. Chollet, Deep learning with Python, Simon and Schuster2021.
[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems, 30 (2017).
[75] J.-B. Cordonnier, A. Loukas, M. Jaggi, On the relationship between self-attention and convolutional layers, arXiv preprint arXiv:1911.03584, (2019).
[76] N. Park, S. Kim, How do vision transformers work?, arXiv preprint arXiv:2202.06709, (2022).
[77] N. Park, S. Kim, Blurs behave like ensembles: Spatial smoothings to improve accuracy, uncertainty, and robustness, International Conference on Machine Learning, PMLR2022, pp. 17390-17419.
[78] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press2016.
[79] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, Advances in neural information processing systems, 29 (2016).
[80] C. Cortes, V. Vapnik, Support vector machine, Machine learning, 20 (1995) 273-297.
[81] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, Attgan: Facial attribute editing by only changing what you want, IEEE transactions on image processing, 28 (2019) 5464-5478.