Journal of Signal Processing Systems (2024) 96:809–820
https://doi.org/10.1007/s11265-025-01942-1

Discriminative Frontal Face Synthesis by Using Attention and Metric Learning

Hakan Cevikalp[1] · **Kaya Turgut**[2] · **Cihan Topal**[3]

Received: 24 June 2024 / Revised: 19 November 2024 / Accepted: 2 January 2025 / Published online: 4 February 2025

Abstract
This paper introduces a novel approach for obtaining distinctive frontal facial representations from collections of multiple facial images. The primary objective is to ensure that the profound features extracted through a deep Convolutional Neural Network (CNN) from these learned facial representations exhibit notable separability within the feature space. The acquisition of frontal facial representations capable of effectively representing entire sets of images holds significant value as it considerably reduces the number of image samples requiring processing. This acceleration proves especially advantageous during the classification testing phase. The proposed method combines three fundamental components: attention mechanisms, adversarial methodologies, and metric learning strategies. We adopt a U-Net architecture enhanced by attention modules for the facial aggregation network that generates frontal faces that approximate multiple face images within image sets. Furthermore, we employ both a discriminator network and a pre-trained facial classification network to successfully achieve the goals of adversarial and metric learning. The experimental studies on different face recognition datasets demonstrate that using only attention mechanisms and metric learning strategy is sufficient to synthesize discriminative frontal face images yielding high classification accuracies.

Keywords Frontal face synthesis · Attention module · Face recognition · Image sets · Deep learning

1 Introduction

Face recognition techniques utilizing collections of images are gaining increased popularity due to their various significant merits in contrast to the utilization of single images. Set based approaches to face recognition involve presenting both the gallery and query entities in the form of image sets, as opposed to singular images. These images can be obtained from diverse sources, such as video frames or multiple disordered observations. The categorization framework assigns the query set to the individual within the gallery whose assortment of images exhibits the highest degree of similarity with the provided query set.

The attainment of effectiveness in set based facial recognition methodologies relies on the consideration of two pivotal factors: the selection of models tasked with approximating facial image sets, and the discernment of an appropriate distance metric designated for the quantification of similarity among these models. In this vein, a multitude of distinct models for image sets have been introduced, encompassing linear and affine subspaces [1–4], convex hulls [1, 5, 6], Gaussian mixture models [7], Grassmannian manifolds [8, 9], as well as manifolds comprised of symmetric positive definite (SPD) matrices [10, 11]. These model formulations have been advanced to approximate image sets while concurrently establishing congruous similarity metrics tailored to each respective model instantiation. Most recent set based face recognition methods focused on obtaining representative prototypes for approximation of the image sets in both image and feature spaces. In this paper, we also follow the same principle and generate discriminative frontal faces that represent the images in the sets. Therefore, our proposed method is closely related to the set based face recognition

✉ Hakan Cevikalp
hakan.cevikalp@gmail.com

[1] Electrical and Electronics Engineering, Machine Learning and Computer Vision Laboratory, Eskisehir Osmagazi University, Eskişehir, Turkey

[2] AIE Department, Huawei Turkey R&D Center, Istanbul, Turkey

[3] Department of AI and Data Engineering, Istanbul Technical University, Istanbul, Turkey

methods using representative prototypes for recognition and face frontalization techniques.

1.1 Related Work

Here, our attention will be directed exclusively towards methodologies for set based facial recognition, employing prototype exemplars for representation of images in the sets. These approaches leverage aggregation techniques that utilize either images or deep neural network features to approximate the characteristics of the image sets.

Among the techniques that operate on images, Hassner et al. [12] present a straightforward approach centered around the computation of representative images through clustering. Images within the sets undergo subdivision into subclusters, and the mean image of each subcluster is harnessed as a prototype. The methods most closely aligned with our own are proposed in [13–15]: Rao et al. [13] introduce a methodology that generates discriminative synthesized images via Generative Adversarial Networks (GANs), subsequently employing an aggregation network to create one or a few images that encapsulate the entirety of a video sequence, achieved through the integration of metric learning principles. These representative images, serving as approximations of the video content, are subsequently deployed for the comparative analysis of image sets. A similar approach grounded in the same theoretical foundation is introduced in [14], differing only in its utilization of the U-Net network as the encoding component. The Disentangled Representation Learning-Generative Adversarial Network (DR-GAN) [15] is introduced with the objective of synthesizing identity-preserving faces at specified target poses. This method accommodates both single and multiple images as input sources for the purpose of image synthesis.

In contradistinction to methodologies that enact aggregation at the level of images, there exists a category of approaches that engage feature-level aggregations for the purpose of set-oriented facial recognition. In general, these methods are simpler compared to the methods that aggregate images since they do not focus on synthesizing a realistic face image that approximates the image sets. In this regard, Cao et al. [16] undertook the training of a deep Convolutional Neural Network (CNN) on the VGGFace2 dataset. In the context of set based recognition, the researchers computed the arithmetic mean of CNN-derived features associated with the constituent faces within each set. Subsequently, the resultant mean vector underwent L2 normalization. These normalized mean vectors were then employed to model the image sets, with set similarity being quantified through the application of Cosine distances between these mean vectors. Neural Aggregation Networks (NAN), as presented in references [17, 18], employ attention mechanisms to generate a solitary and succinct feature representation that approximates the collective images within a given set. Likewise, Gong et al. [19] introduced the C-FAN (Componentwise Feature Aggregation Network) approach, which aggregates deep facial representations from images within a set, culminating in the computation of a singular feature vector that encapsulates the entire set. Xie and Zisserman [20] proposed an elegant deep neural network paradigm, wherein the model learns to derive a solitary feature descriptor that embodies all images within a set. This is achieved through the weighted averaging of facial descriptors pertaining to the set's constituent images. Clustering and metric learning are used in [21, 22] for obtaining prototypes representing image sets in the feature space. Once the prototype features are determined, the face samples are classified based on the shortest distances between the face image features and these prototypes.

Face frontalization refers to the process of transforming a facial image obtained from different angles into a frontal view. For this purpose, numerous methods have been proposed. More recent face frontalization methods employed data-driven GAN models for frontal face synthesis. Noticeably, the Disentangled Representation Learning-Generative Adversarial Network (DR-GAN) [15] is introduced with the objective of synthesizing identity preserving faces at specified target poses. Huang et al. [23] proposed Two-Pathway Generative Adversarial Network (TP-GAN) for frontal view synthesis by utilizing both global structures and local details. The Couple Agent Pose Guided Generative Adversarial Network (CAPG-GAN) is proposed in [24] to synthesize arbitrary pose images where landmark heatmaps of faces are used to incorporate pose information in the learning process. Tian et al. [25] proposed CR-GAN that can learn "complete" representations, using a two-pathway learning scheme that utilizes self-supervised learning for frontal face synthesis. There are also hybrid methods that employ 3D face models and GANs together. FF-GAN [26] is proposed to incorporate 3D face model into GANs. The learned 3D model is used for global pose and low frequency information whereas the input images provided high frequency local information. Liu et al. [27] introduced 3D-FM GAN method that solves the image-to-image translation/editing problem by using a conditional Style GAN. Zhou et al. [28] proposed the Rotate and Render method, which is a novel unsupervised framework that can synthesize photorealistic rotated faces using only single-view image collections in the wild. However, we would like to point out that all these face frontalization methods with the exception of DR-GAN take only a single face image as input and transform it to a frontal face. They do not work on image sets as in our proposed method, therefore these methods are different than our proposed method.

1.2 Contributions

In this paper, a novel approach is introduced, which involves acquiring distinctive frontal facial representations from sets of multiple facial images. The primary aim is to ensure that the profound features extracted by a deep CNN (Convolutional Neural Network) from these learned facial representations can be readily distinguished within the feature space. The acquisition of frontal facial representations that can effectively stand for entire sets holds significance as it substantially curtails the volume of image samples necessitating processing, thereby accelerating the testing phase of the classifier. This novel technique integrates three pivotal elements: attention, adversarial, and metric learning methodologies. The proposed method employs a U-Net structure complemented by attention modules for the facial aggregation network. Moreover, it leverages a discriminator network and a pre-trained facial classification network to accomplish the objectives of adversarial and metric learning.

The methodologies employing feature aggregation in the feature domain for the purpose of face recognition incorporate attention mechanisms; however, they refrain from generating a frontal prototype image. In contrast, our approach not only extends this paradigm but also yields distinctive frontal images, the convolutional neural network (CNN) features of which can be harnessed for face recognition tasks.

In comparison to the approaches centered on image-level aggregation, our method deviates by virtue of its employment of attention techniques. To the best of our knowledge, none of the existing methodologies have leveraged attention modules (both the self and cross attention modules together) as in our proposed method to produce a representative facial image, thereby approximating sets of facial images. Moreover, both [13] and [14] opt for the adoption of the contrastive loss function in lieu of the triplet loss function used in our method, and neither [13] nor [15] integrates the U-Net network as the encoding component.

In a broader context, our proposed methodology bridges the gap between methodologies employing aggregation within the image and feature spaces. As such, it furnishes a sophisticated approach for approximating sets of facial images with discerning frontal countenances. Briefly, our contributions can be summarized as follows:

- We propose a novel method utilizes attention, adversarial and triplet losses to create discriminative frontal faces for approximation of image sets.
- We employ U-Net architecture supported with attention modules as encoders and show that it is better suited for creating high resolution frontal faces.
- Our proposed method produces much better frontal faces compared to existing methods and the face recognition accuracies returned by the face classification methods using synthesized frontal faces are superior to accuracies of related methods that use aggregation at the level of images.
- Lastly, our proposed method learns frontal faces that represent the entire sets, and this process significantly reduces the number of image samples to be processed and speeds up the classification in the testing stage.

2 Method

Our proposed method includes three important components including attention, adversarial and metric learning techniques. We utilize attention module in the U-Net architecture, which is used as Frontal Face Aggregation network. In addition to these, we employ a discriminator and a pre-trained CNN classifier networks for adversarial and metric learning as illustrated in Figure 1. Now, we explain each component in more details below.

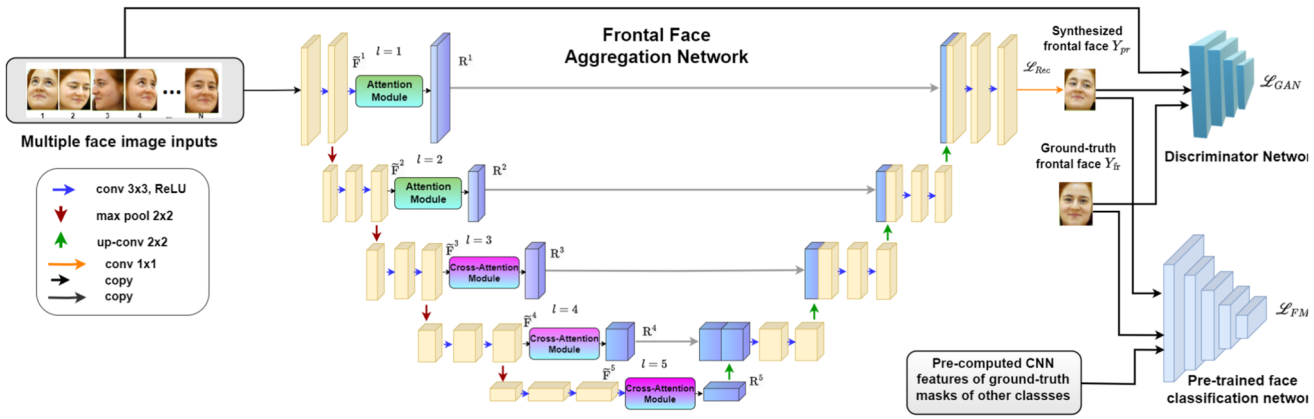

Figure 1 The illustration of the proposed method

2.1 Face Aggregation Network

Our proposed method utilizes U-Net architecture [29] as the face aggregation network since we want to create high-resolution realistic frontal face images. We would like to point out that the aggregation network is similar to generator network in conditional Generative Adversarial Networks (GANs), and the most recent GAN methods such as [30] producing high resolution images employ U-Net architectures as generator networks.

The U-Net architecture comprises both contracting and expansive pathways as illustrated in Figure 1. To implement the conventional U-Net design as introduced in reference [29], we utilized a repetitive sequence involving a pair of 3×3 convolutions, each succeeded by a rectified linear unit (ReLU), and a 2×2 max pooling operation with a stride of 2 for the purpose of down-sampling. With every down-sampling stage (we will call it a *layer* for with a slight abuse of terminology for the sake of notation), the number of feature channels is duplicated. The expansive pathway consists of successive stages, wherein a feature map is first upsampled and then subjected to a 2×2 convolution (referred to as "up-convolution") that reduces the feature channel count by half. This result is then concatenated with the correspondingly cropped feature map from the contracting pathway, followed by two 3×3 convolutions, each succeeded by a ReLU activation function.

In our proposed method, the network takes N multiple RGB face images for each set as input and returns a single frontal face approximating those face images. We used an attention module in each layer of the contracting pathway and transmitted the following output to the expansive path through skip connections as seen in Figure 1. As a reconstruction loss at the end of the U-Net architecture, we employ the pixel-wise mean squared error (MSE) between the ground-truth frontal face and predicted face as,

$$\mathcal{L}_{Rec} = \frac{1}{p} \left\| Y_{fr} - Y_{pr} \right\|_{\mathcal{F}}^2, \tag{1}$$

where Y_{fr} denotes the ground-truth frontal face, Y_{pr} is the U-Net's predicted face image, p is the total number of pixels, and $\|.\|_{\mathcal{F}}$ represents the Frobenius norm of a matrix.

Attention Module Attention module is responsible for finding the proper aggregation weights for each individual for returning a discriminative prototype feature map representing multiple images, and it is visualized in Figure 2. This module takes feature maps of the face images in a specified set and it first creates a single representative feature map by using weighted sum of existing feature maps as in the feature aggregation methods [17, 18]. Now, let $\tilde{\mathbf{F}}^l \in \mathbb{R}^{N \times C \times H \times W}$ feature maps of a single set including N face images, where C is the number of channels in the specified layer, H is the height, W is the width of the feature maps, and l represents the feature layer. We first reshape $\tilde{\mathbf{F}}^l$ to $\mathbf{F}^l \in \mathbb{R}^{N \times d}$, where $d = C \times H \times W$. Let us assume that $\mathbf{f}_n^l \in \mathbb{R}^d$, $n = 1, \ldots, N$, is the vectors forming the matrix $\mathbf{F}^l$. The representative feature vector, $\mathbf{r}^l \in \mathbb{R}^d$, is constructed by using the weighted sum of $\mathbf{f}_n^l$ as

$$\mathbf{r}^l = \sum_{n=1}^{N} a_n^l \mathbf{f}_n^l, \tag{2}$$

where a_n^l, $n = 1, ..., N$, is the linear weight computed by using the following formula,

$$a_n^l = \frac{\exp(e_n^l)}{\sum_{i=1}^{N} \exp(e_i^l)}. \tag{3}$$

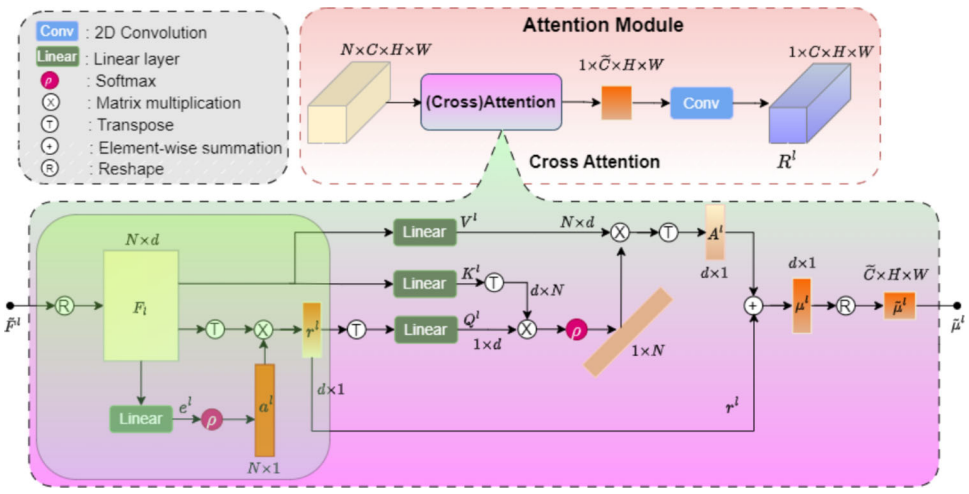

Figure 2 The illustration of the proposed attention module. We used full cross-attention in the last lower 3 layers of the network whereas the attention module returning only $\mathbf{r}^l$ is used in the first two upper layers

Here, e_n^l is the corresponding significance computed via dot product with a kernel filter $\mathbf{w}^l$ as,

$$e_n^l = (\mathbf{w}^l)^\top \mathbf{f}_n^l. \tag{4}$$

It is reported in [20] that higher values of e_n^l represent higher visual quality of the corresponding faces, therefore the resulting representative feature vector is mostly composed of features extracted from high quality face images. Methods using feature aggregation such as [17, 18, 20] stop at this point and do not use the cross-attention mechanism that computes relationships between the feature vectors of images and computed prototype vector. Inspired by the success of attention mechanisms in vision transformers [31], we also compute the correlations (cross-attention) between the image feature vectors and the computed prototype vector, $\mathbf{r}^l$. By using a similar notation in vision transformers, let us set $\mathbf{Q}^l = \mathbf{W}_Q^l \mathbf{r}^l$, $\mathbf{K}^l = \mathbf{W}_K^l \mathbf{F}^l$, and $\mathbf{V}^l = \mathbf{W}_V^l \mathbf{F}^l$. We compute an attention vector, $\mathbf{A}^l$, by using,

$$\mathbf{A}^l = \text{softmax}\left(\frac{\mathbf{Q}^l (\mathbf{K}^l)^\top}{\sqrt{d}}\right)\mathbf{V}^l. \tag{5}$$

We sum the resulting attention vector with the prototype feature vector, and obtain the final feature vector that represents the entire set of features by using,

$$\boldsymbol{\mu}^l = \mathbf{A}^l + \mathbf{r}^l. \tag{6}$$

Finally, we reshape $\boldsymbol{\mu}^l \in \mathbb{R}^d$ to $\tilde{\boldsymbol{\mu}}^l \in \mathbb{R}^{H\times W}$ and apply a 2D convolution using C filters to compute the feature map $\mathbf{R}^l \in \mathbb{R}^{C\times H\times W}$ with the original size by using the formulation,

$$\mathbf{R}^l = \text{Conv2D}(\tilde{\boldsymbol{\mu}}^l). \tag{7}$$

These new features computed by using attention are passed to the expansive pathway through skip connections at the upper layers of the U-Net architecture. Using cross-attention in the proposed method increased the number of network parameters significantly, and caused GPU memory problems. Therefore, we used cross-attention in the last three layers of the U-Net as shown in Figure 1.

2.2 Discriminator Network and Adversarial Loss

In order to use adversarial loss, we need a discriminator network, D, and we employed a patch-based fully convolutional network as in [30, 32]. The U-Net architecture explained above generates realistic frontal faces and the discriminator network aims to distinguish real images from the generated ones. Now let, $\big((I_1, ..., I_N), Y_{fr}\big)$ represent the input images in the set and their corresponding ground-truth frontal face image. Considering the U-Net's predicted frontal face image as $G(I_1, I_2, ..., I_N) = Y_{pr}$, the conditional GAN used in our approach aims to model the conditional distribution of frontal faces given the various multiple images with different poses through the following minimax game,

$$\min_G \max_D \mathcal{L}_{GAN}(G, D), \tag{8}$$

where the objective function of $\mathcal{L}_{GAN}(G, D)$ is equivalent to

$$\mathbb{E}_{((I_1,...,I_N),Y_{fr})}\left[\log(\frac{1}{N}\sum_{i=1}^{N} D(I_i, Y_{fr}))\right] + \mathbb{E}_{(I_1,...,I_N)}\left[\log(1-\frac{1}{N}\sum_{i=1}^{N} D((I_1, ..., I_N), G(I_1, ..., I_N)))\right].$$

It should be noted that this conditional GAN loss function is very similar to the loss function of pix2pix [32] method with the exception that we have multiple images and one corresponding ground-truth. Therefore, we calculate the loss for each input image in the set and then take the average as the final loss.

2.3 Feature Matching and Triplet Losses for Distance Metric Learning

Our objective is to generate frontal face images that exhibit clear separability within the feature space. The utilization of reconstruction and adversarial losses in face generation does not guarantee inherent separability. This deficiency arises due to the absence of a discriminative loss aimed at maximizing differentiation between different classes. To address this limitation, we introduce two additional loss terms. These terms are designed to minimize the distances in CNN features between actual and synthesized frontal faces, while also maximizing the differences in CNN features among synthesized faces of diverse classes. It is worth noting that the concept of feature matching loss has been embraced in various GAN and face frontalization methodologies. However, these approaches typically employ the discriminator network for feature matching purposes. In contrast, our approach involves a distinct and deeper network, specifically trained for precise face classification through a comprehensive dataset. The Rotate and Render method, as proposed in [28], also adopts an alternative network as in our proposed method. However, their chosen network is trained on the ImageNet dataset by using the softmax loss, which might not necessarily yield discriminative CNN features suitable for accurate face classification. Additionally, the prevalent use of Euclidean or L1 norm distances for feature matching further complicates matters. As a result, opting for a classification network trained with the traditional softmax loss function might not be ideal. This is because the softmax loss function tends to produce radially distributed CNN features, which are more suitable for the cosine distances.

To avoid such limitations, we first trained a VGG face classification network by using ESOGU Face Videos dataset [33] including approximately 764K face images of 285 person. In the VGG network, we replaced the softmax loss function with the Deep Simplex classifier loss function [34] since this loss function is better suited for the Euclidean distances. Then, we used this pre-trained network for feature matching and triplet losses. It should be noted that the weights of this pre-trained network are frozen and not updated during the training stage of the frontal face generation.

By following the methods [28, 30, 32], the feature matching loss is realized by extracting features from multiple layers of the pre-trained VGG face classification network and minimizing the distance between synthesized frontal face and its ground-truth frontal face image. Now, assume that $\phi_i(Y_{pr})$ and $\phi_i(Y_{fr})$ respectively denote the CNN features (feature maps) of synthesized and ground-truth frontal faces extracted from the i-th layer of the classifier network. For total nl layers, the feature matching loss can be written as,

$$\mathcal{L}_{FM} = \frac{1}{nl}\sum_{i=1}^{nl}\left\|\phi_i(Y_{fr}) - \phi_i(Y_{pr})\right\|_{\mathcal{F}}^2. \tag{9}$$

The objective of the feature matching loss given above is to reduce the Euclidean distances between the CNN features extracted from both the actual and generated frontal faces. Subsequently, we introduce an additional triplet loss component that aims to maximize the distinction between different facial classes, i.e., inter-class separation. Now let, $\Phi(Y_{pr}^k)$ and $\Phi(Y_{fr}^k)$ respectively denote the CNN feature vectors of the k-th face class just before the classification layer. In this case, the final triplet loss function can be written as,

$$\mathcal{L}_{triplet} = \sum_{k=1}^{K}\sum_{j=1, j\neq k}^{K}\max\left(0, m + \left\|\Phi(Y_{pr}^k) - \Phi(Y_{fr}^k)\right\|^2 - \left\|\Phi(Y_{pr}^k) - \Phi(Y_{fr}^j)\right\|^2\right), \tag{10}$$

where K is the number of total classes in the training set. The triplet loss function ensures that the distance between the CNN features of the synthesized frontal face and its corresponding ground-truth face class image is smaller than the distance between the CNN features of the synthesized frontal face and ground-truth frontal faces of other classes by at least a selected margin, m.

2.4 Overall Loss Function

Our final loss function is the sum of the loss functions described above and it is given as,

$$\mathcal{L} = \mathcal{L}_{FM} + \eta\mathcal{L}_{triplet} + \lambda\mathcal{L}_{Rec} + \kappa\mathcal{L}_{GAN}, \tag{11}$$

where λ, κ, and η are the weight parameters that must be set by the user. To determine the weights, we used a small dataset and applied the grid search methodology. The accuracies are computed based on the pixel-wise mean squared error (MSE) between the ground-truth frontal faces and predicted faces. Similar to the existing studies such as [32] in the literature, setting λ parameter to higher values, e.g., $\lambda = 100$ worked best. Also, we set the κ parameter to very low values since the GAN loss has a degrading effect on the performance.

3 Experiments

3.1 Datasets

In our experiments, we utilized ESOGU-285 Video [33] and CMU Multi-PIE Face [35] datasets for training our proposed network. For evaluation, we used Honda/UCSD [36] and IJB-A datasets [37]. The details of each dataset are given below:

ESOGU-285 Video Dataset The ESOGU-285 database is a video dataset comprising 285 individuals, each represented by eight distinct video recordings. These videos were captured in an indoor setting under different lighting conditions during two separate sessions, with a minimum three-week interval between them. The video lengths vary, with the shortest video comprising 100 frames and the longest extending to 1360 frames. In total, the dataset encompasses 764,006 frames distributed across 2,280 videos.

CMU Multi-PIE Face Dataset The CMU Multi-PIE face database encompasses over 750,000 images featuring 337 individuals, captured across a timeframe of five months during up to four distinct sessions. During these sessions, subjects were photographed from 15 different angles and under 19 distinct lighting conditions, all while displaying various facial expressions. Furthermore, the database includes high-resolution frontal images for each individual.

Honda/UCSD Dataset The Honda/UCSD dataset consists of 20 individuals and 59 video sequences with each sequence including approximately 300-500 frames. During the testing, 20 sequences set aside for training are used as the gallery image sets and the remaining 39 sequences are used as probe sets.

IJB-A Dataset The IJB-A dataset is a collection of face data organized around templates, comprising a total of 1,845 individuals and encompassing 5,712 images, 2,085 videos, with an average of 11.4 images and 4.2 videos per individual. A template in this dataset consists of a varying number of static images and video frames obtained from diverse sources. These images and videos have been sourced from the Inter-

net and exhibit complete unconstraint, featuring significant variations in factors such as pose, lighting, and image quality.

For IJB-A dataset, there are specific pre-defined protocols for conducting various types of facial recognition tasks, including 1-to-1 template-based face verification, 1-to-N template-based face identification, and 1-to-N open-set video face identification. For reporting accuracies, we follow the standard benchmark procedure for IJB-A dataset to evaluate the proposed method on "search" protocol for 1:N face identification.

3.2 Implementation Details and Results

For training we used all images coming from both ESOGU-285 Video and CMU Multi-PIE Face datasets. The total number of training images is 1,413,321. All images are aligned based on the 5 points landmarks returned by the Retina Face detector [38]. We selected a frontal face image from each image set as ground-truth frontal face mask. The total number of masks (and hence the total number of image sets in the training) is 3201 and this value is larger than the total number of people since each individual typically has several image sets, e.g., each person in ESOGU-285 Video dataset has 8 image sets created by using the videos collected in each scenario. As a classification network used for feature matching and triplet losses, we trained a VGG-19 network on ESOGU-285 Video dataset. In addition, we also utilized another VGG-19 network trained on ImageNet object dataset. We trained the network with a batch size of 8, where each input consists of 16 face images fed into the model. Incorporating cross-attention in every layer of the UNet architecture caused memory issues, exceeding the 24 GB GPU capacity. To address this, we applied cross-attention only in the last three layers of the UNet. However, after training, the network can run with a batch size of 8 using just 8 GB of RAM.

We set the number of multiple face image input size to $N = 16$, i.e., we synthesize a frontal face by using 16 images. The training process for this network spanned approximately 10 days. To assess the resulting model, we employed image sets from the HONDA/UCSD dataset. Initially, we identified the frontal face images that would serve as ground-truth references. Subsequently, we excluded these reference images from the sets and randomly selected 32 images from each set. These selected images were used as input for the trained model, from which we obtained the predicted frontal face images. To assess the reconstruction performance, we employed three metrics describe below:

Mean Squared Error (MSE) This metric is used to measure the average squared difference between the predicted values and the actual values. In our tests, we used the pixel-wise mean squared error (MSE) between the ground-truth frontal face and predicted face as,

$$MSE(Y_{fr}, Y_{pr}) = \frac{1}{p} \left\| Y_{fr} - Y_{pr} \right\|_{\mathcal{F}}^2, \tag{12}$$

where Y_{fr} denotes the ground-truth frontal face, Y_{pr} is the U-Net's predicted face image, p is the total number of pixels, and $\|.\|_{\mathcal{F}}$ represents the Frobenius norm of a matrix.

Structural Similarity Index Measure (SSIM) SSIM [39] is a perceptual metric that quantifies image quality degradation by comparing the structural information in two images. It is a quality metric that assesses the visual impact of three key features in an image: brightness, contrast, and structure. It yields values between 0 and 1, where higher values signify superior performance. The formulation of SSIM is given as:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{y}} + C_1)(2\sigma_{xy} + C_2)}{(\boldsymbol{\mu}_{\mathbf{x}}^2 + \boldsymbol{\mu}_{\mathbf{y}}^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{13}$$

Here, $\mathbf{x}$ and $\mathbf{y}$ respectively represent the image patches coming from the ground-truth frontal face and the predicted face returned by the network, $\boldsymbol{\mu}_{\mathbf{x}}$ represents the mean intensity of $\mathbf{x}$, $\boldsymbol{\mu}_{\mathbf{y}}$ represents the mean intensity of $\mathbf{y}$, σ_x^2 denotes the variance of $\mathbf{x}$, σ_y^2 denotes the variance of $\mathbf{y}$, and σ_{xy} is the covariance between $\mathbf{x}$ and $\mathbf{y}$. The pre-selected constants C_1 and C_2 are used stabilize the division with weak denominator.

Fréchet Inception Distance (FID) FID [40] captures the similarity of generated images to real ones. Higher distances indicate a poorer generated image whereas a score of 0 indicates a perfect match. The formulation of FID is given as:

$$FID(Y_{fr}, Y_{pr}) = \left\| \mu_{fr} - \mu_{pr} \right\|_2^2 + \mathrm{Tr}(\Sigma_{fr} + \Sigma_{pr} - (2\Sigma_{fr}\Sigma_{pr})^{1/2}), \tag{14}$$

where μ_{fr} denotes the mean of the feature embeddings of the ground-truth frontal images, μ_{pr} denotes the mean of the feature embeddings of the synthesized (predicted) images by the network, Σ_{fr} represents the covariance matrix of the feature embeddings of the ground-truth frontal images, and Σ_{pr} represents the covariance matrix of the feature embeddings of the synthesized images. Here Tr(.) denotes the trace of a matrix (sum of the diagonal entries).

In addition, we also reported the classification accuracies (CAs) obtained for the Honda/UCSD dataset by using standard testing protocol defined above.

The results are presented in Table 1. Our comparisons were exclusively made with DR-GAN [15] as the authors of [13] and [1] did not provide access to their source codes or trained

Table 1 Frontal face synthesis performances on Honda/UCSD dataset.

Loss Terms	MSE↓	SSIM↑	FID↓	CA(%)↑
CR-GAN	0.0681	0.647	3,223,295	81.02
DR-GAN	0.0350	0.624	2,335,097	**100**
Proposed Method	**0.0097**	**0.764**	**2,321,217**	**100**

The bold results in the tables represent the best accuracy in the tables and they must remain

models for evaluation purposes. We conducted a comparative assessment of DR-GAN and our proposed method, subjecting both to identical settings. Both methods were tested using the same set of images, and the same number of image size, N. In addition, we also used a face frontalization method CR-GAN [25] for comparison. As indicated before, CR-GAN takes only a single face image as an input and synthesizes a frontal face image. Therefore, it is not possible to generate a single frontal face for N images in the sets. For reporting results, we synthesized frontal faces for N images within sets independently and computed the metrics by using ground-truth masks and averaged the results. As seen in the table, our proposed method significantly outperforms DR-GAN and CR-GAN in terms of MSE, SSIM, and FID metrics. DR-GAN and proposed method achieve 100% test classification accuracy, whereas the classification accuracy of CR-GAN is quite low, 82.01%. Figure 3 provides a visual representation of the ground-truth masks and the resultant frontal faces produced by DR-GAN, CR-GAN and our proposed approach. For CR-GAN, we selected the best synthesized frontal face image within a set. As depicted in the figure, DR-GAN and our proposed method demonstrate competence in generating frontal images. The synthesized frontal images by CR-GAN are far from ideal and identities of the individuals are completely lost. When our proposed method is compared to DR-GAN, a notable distinction emerges: our proposed method yields images with a higher degree of realism, while those generated by DR-GAN exhibit a more artificial quality. Additionally, the images produced by our method exhibit a closer alignment with the ground-truth masks. In contrast, certain images generated by DR-GAN noticeably diverge from the ground-truth masks, as evidenced, for instance, in the outputs of DR-GAN displayed in the 5th, 8th, and last columns. These outcomes were anticipated given that DR-GAN relies solely on adversarial losses and does not incorporate the use of feature matching and triplet losses derived from a separate classification network as in our proposed method.

3.3 IJB-A Results

In IJB-A dataset, there are 10 random training and testing splits. Each split provides gallery and probe sets. The gallery set consists of 112 or 113 subjects and probe set has 167 subjects (55 subjects different from gallery). For template based face identification, we used "search" protocol for 1:N face identification. Here the Rank-N classification accuracies are reported for identification, and the classification rate is the percentage of probe searches, which correctly finds the

Figure 3 The synthesized frontal images returned by DR-GAN, CR-GAN and the proposed method.

Table 2 Identification accuracies (%) on the IJB-A benchmark.

Method	IJB-A Dataset True Positive Identification Rate (TPIR) (%)		Rank-N Accuracy (%)		
	@FPIR=0.01	@FPIR=0.1	Rank-1	Rank-5	Rank-10
NAN [17]	81.7 ± 4.1	91.7 ± 0.9	95.8 ± 0.5	$\mathbf{98.0 \pm 0.5}$	$\mathbf{98.6 \pm 0.3}$
C-FAN [19]	86.9 ± 4.7	92.9 ± 1.0	94.6 ± 0.8	96.2 ± 0.5	– –
DREAM [41]	– –	– –	94.6 ± 1.1	96.8 ± 1.0	– –
GFA [42]	$\mathbf{96.4 \pm 3.9}$	$\mathbf{97.4 \pm 0.8}$	$\mathbf{97.1 \pm 1.3}$	– –	98.5 ± 0.4
ArcFace [43]	93.5 ± 1.5	94.1 ± 1.5	94.7 ± 1.3	95.3 ± 1.2	95.6 ± 1.2
Pooling Faces [12]	– –	– –	$84.6\pm$ – –	$93.3\pm$ – –	– –
DR-GAN [44]	– –	– –	85.5 ± 1.5	94.7 ± 1.1	– –
Proposed Method (Various N)	80.3 ± 1.7	85.1 ± 1.5	88.6 ± 1.1	91.3 ± 0.6	92.8 ± 0.6
Proposed Method (Fixed $N = 16$)	85.2 ± 1.9	89.2 ± 1.5	91.0 ± 1.3	93.4 ± 1.0	94.0 ± 1.0

The bold results in the tables represent the best accuracy in the tables and they must remain

probe's gallery mate in the gallery set within top N rank-ordered results. In addition, we also report the TPIR (True Positive Identification Rate) accuracies obtained for different FPIR (False Positive Identification Rate) values. It should be noted that many gallery and probe sets include different number of images, and the number of images can be smaller than 16. Therefore, to use the proposed method with different number of input images, we revised the network so that it can accept any desired number of images and re-trained the resulting network by randomly selecting images between 7-12 instead of fixed 16 images from the same training dataset described before. The resulting network is capable of generating frontal images from different number of images.

During the evaluation stage, we first obtained frontal images in the gallery and probe sets. We synthesized the images coming from different medias separately. For the method using fixed input image size, we created copies of images if the number of images is less than 16. After synthesizing frontal images, we extracted the CNN features of the resulting images by using the ArcFace method (ResNet-100 architecture) [45] trained on MS1MV3 dataset [46]. The results are given in Table 2. Proposed Method (Fixed $N = 16$) represents the network accepting fixed 16 images as input, whereas Proposed Method (Various N) is the network that can synthesize frontal images from any desired number of images. As a baseline, we also report the accuracies obtained by the ArcFace method using all images in the gallery and probe sets. The "– –" symbol in the table stands for not reported results.

The last four techniques delineated in the table's lower section are the methods that aggregate images within the image domain. Within this subset of approaches, both of our proposed methods demonstrate a marked superiority over its counterparts, yielding superior performance in terms of Rank-1 accuracy. Among our proposed methods, Proposed Method (Fixed $N = 16$) significantly outperforms Proposed Method (Various N) accepting various number of input images. There is not a big performance difference between Rank-1 and Rank-5 accuracies as in other compared methods. ArcFace method also follows a similar pattern. Nevertheless, it is important to note that our accuracy metrics fall short when compared with methodologies focused on the aggregation of convolutional neural network (CNN) features. Such a discrepancy is anticipated, given that these latter approaches leverage more intricate classification networks and optimize CNN features for classification during the training stage. In contrast, our approach employs a comparatively shallower classification network during the training process, and our proposed methodology does not undergo end-to-end optimization. Rather, it involves a two-step process in which

Table 3 Performance scores for various loss terms

Loss Terms	MSE	SSIM	CA(%)
$\mathcal{L}_{Rec}$ (No Attention)	0.0130	0.712	70.2
$\mathcal{L}_{Rec}$	0.0101	0.753	96.4
$\mathcal{L}_{Rec} + \kappa\mathcal{L}_{GAN}$	0.0110	0.753	93.8
$\lambda\mathcal{L}_{Rec} + \mathcal{L}_{FM} + \eta\mathcal{L}_{triplet}$ (VGG Face)	0.0105	0.743	98.4
$\lambda\mathcal{L}_{Rec} + \mathcal{L}_{FM} + \eta\mathcal{L}_{triplet}$ (VGG Object)	0.0110	0.756	**100**
$\lambda\mathcal{L}_{Rec} + \kappa\mathcal{L}_{GAN} + \mathcal{L}_{FM} + \eta\mathcal{L}_{triplet}$ (VGG Object)	0.0110	0.750	98.4
$\lambda\mathcal{L}_{Rec} + \kappa\mathcal{L}_{GAN} + \mathcal{L}_{FM} + \eta\mathcal{L}_{triplet}$ (VGG Face)	**0.0100**	**0.761**	**100**

The bold results in the tables represent the best accuracy in the tables and they must remain

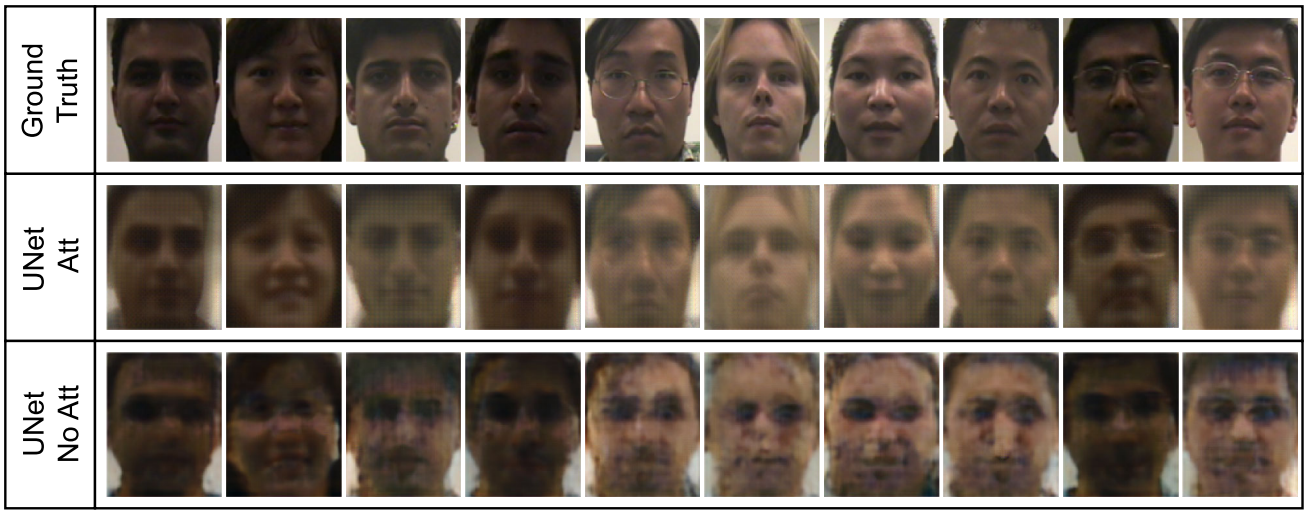

Figure 4 The synthesized frontal images returned by the U-Nets with and without attention modules

frontal images are initially synthesized, followed by their classification using a separate, distinct classification network. However, our proposed method offers the distinct advantage of synthesizing frontal images from a multitude of input images. This speeds up the testing stage as the classification network operates on a reduced set of images.

3.4 Ablation Studies

We conducted ablation experiments to show the impact of individual loss terms on the frontal face synthesis. To this end, we randomly selected 10 individuals from the ESOGU-285 video dataset for training. Each individual has 8 different face image sets per class, yielding 80 different frontal face masks. The tests are evaluated on 20 individuals selected from a completely different dataset, Honda/UCSD as before. The results are given in Table 3.

We first conducted tests to verify the importance of attention mechanism in the proposed method. To this end, we removed the attention modules from the layers and used the direct filter outputs. The results obtained for the U-Net architecture without attention modules are denoted as $\mathcal{L}_{Rec}$ (No Attention) in the table. Figure 4 shows the synthesized frontal faces returned by the proposed network with and without attention modules. As seen in the figure, the synthesized images returned by the network using attention modules are significantly better compared to the one without attention module. Also, the performance scores for the network using attention module demonstrate that using attention modules is crucial for the success.

In our experiments, we also assessed the effects of two distinct neural networks: VGG-19 object and VGG-19 face classification networks, which were trained on the ImageNet and ESOGU Face Videos datasets, respectively. VGG-19 object network employed the conventional softmax loss function, while we utilized the deep simplex classification loss for the VGG face classification network. The best accuracy is achieved by the network using VGG-19 face classification network, but there is not a significant difference between the scores.

The most surprising result is the negative effect of adversarial loss on the performance. Using GAN loss term with higher coefficients decrease the performance as seen in the fourth column of the table. This may be attributed to the fact that there are images closer to frontal positions in our trained and tested sets, thus a simple aggregation is sufficient for frontal face synthesis rather than trying to generate frontal faces through adversarial loss.

3.5 Comparison of testing times

We conducted experiments for comparison of the testing times of methods to demonstrate the proposed method is more efficient in terms of testing times. As we stated before, the acquisition of frontal facial representations that can effectively stand for entire sets holds significance as it substantially curtails the volume of image samples necessitating processing, thereby accelerating the testing speed. It is because, once we synthesize the frontal image representing the entire set, we simply extract CNN feature of this frontal image and use it for classification. In contrast, the classical methods using CNN features of entire set must extract feature representation of each image independently then take the average and use the resulting CNN feature for classification. We compared the testing time of our proposed method to the testing times of classical ArcFace and DR-GAN methods. The experiments were performed within a

Table 4 Comparison of testing times on Honda/UCSD dataset

Methods	Testing Time (s)
ArcFace	17.3 s
DR-GAN	7.9 s
Proposed Method	6.0 s

Python environment utilizing a PC equipped with an Intel Xeon processor, 128 GB of RAM and an NVIDIA Tesla V100 GPU with 16 GB of RAM. Table 4 presents the recorded testing times in seconds (s), representing the GPU computation time required to generate all frontalized images for a set of 16 randomly selected face images drawn from the video sequences of the HONDA/UCSD dataset. There are 59 face image sets belonging to 20 individuals in Honda/UCSD dataset and we randomly selected 32 images from 20 sets. Therefore, the total number of tested image sequences is 40, and each test sequence includes 16 face images. For our proposed method and DR-GAN, we first synthesize the frontal face image and then use ArcFace to extract its CNN feature. For the ArcFace method, we extract CNN features of all 16 images in a set and then take their average. As seen in the table, our proposed method is the fastest method in terms of testing time. More precisely, the proposed method is 2.9 times faster than the ArcFace method and 1.3 times faster than DR-GAN despite we use a more complex network compared to DR-GAN. These results are expected since we feed all 16 face images at the same time to our network and then run ArcFace network only once to extract CNN feature of the resulting face image. In contrast, DR-GAN synthesizes a frontal face image for each face input separately and then run another aggregation module to weight each output to obtain a unique frontal face image. Therefore, our method runs faster compared to DR-GAN although we use a more complicated UNet architecture.

4 Conclusions and Future Work

We have introduced an innovative approach aimed at generating frontal images from a collection of multiple facial images. Our methodology involves employing a U-Net architecture enriched with attention modules for facial feature aggregation, integrating a discriminator network for adversarial learning, and utilizing a VGG-19 classification network to ensure that the CNN features of the synthesized images can be readily classified within the CNN feature space. Our experimental findings reveal that the incorporation of adversarial loss adversely affects performance, whereas relying solely on attention mechanisms and distance metric learning losses proves capable of generating realistic and distinctive frontal facial images. It is important to note that our proposed method yielded lower accuracies when compared to approaches that aggregate features in the CNN domain. In future investigations, we intend to enhance accuracy by employing a single deeper classification network during both the training and testing phases.

Acknowledgements This work was supported by the Scientific and Technological Research Council of Turkey (TUBİTAK) under grant number EEEAG-118E294.

References

1. Cevikalp, H., & Triggs, B. (2010). Face recognition based on image sets. In IEEE Conference on Computer Vision and Pattern Recognition.
2. Cevikalp, H., Yavuz, H. S., & Triggs, B. (2020). Face recognition based on videos by using convex hulls. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(12), 4481–4495.
3. Hu, Y., Mian, A. S., & Owens, R. (2012). Face recognition using sparse approximated nearest points between image sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(3), 1992–2004.
4. Yang, M., Zhu, P., Van Gool, L., & Zhang, L. (2013). Face recognition based on regularized nearest points between image sets. In IEEE International Conference on Automatic Face and Gesture Recognition.
5. Zhu, P., Zuo, W., Zhang, L., Shiu, S.C.-K., & Zhang, D. (2014). Image set-based collaborative representation for face recognition. *IEEE Transactions on Information Forensics and Security, 9*, 1120–1132.
6. Cevikalp, H., & Dordinejad, G. G. (2020). Video based face recognition by using discriminatively learned convex models. *International Journal of Computer Vision, 128*, 3000–3014.
7. Wang, W., Wang, R., Huang, Z., Shan, S., & Chen, X. (2018). Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. *IEEE Transactions on Image Processing, 27*, 151–163.
8. Hamm, J., & Lee, D. (2008). Grassmann discriminant analysis: a unifying view on subspace-based learning. In International Conference on Machine Learning.
9. Wang, T., & Shi, P. (2009). Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters, 30*, 1161–1165.
10. Huang, Z., Wang, R., Shan, S., Li, X., & Chen, X. (2015). Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In ICML.
11. Huang, Z., Wang, R., Shan, S., Van Gool, L., & Chen, X. (2018). Cross euclidean-to-riemannian metric learning with application to face recognition from video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*, 2827–2840.
12. Hassner, T., Masi, I., Kim, J., Choi, J., & Harel, S. (2016). Pooling faces: Template based face recognition with pooled face images. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
13. Rao, Y., Lin, J., Lu, J., & Zhou, J. (2017). Learning discriminative aggregation network for video-based face recognition. In IEEE Conference on Computer Vision.
14. Hormann, S., Cao, Z., Knoche, M., Herzog, F., & Rigoll, G. (2021). Face aggregation network for video face recognition. In International Conference on Image Processing.

15. Tran, L., Yin, X., & Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR).
16. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognizing faces across pose and age. In IEEE International Conference on Automatic Face & Gesture Recognition.
17. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., & Hua, G. (2017). Neural aggregation network for video face recognition. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR).
18. Liu, Z., Hu, H., & Bai, J. (2019). Feature aggregation network for video face recognition. In IEEE International Conference on Computer Vision (ICCV) Workhops.
19. Gong, S., Shi, Y., Kalka, N. D., & Jain, A. K. (2019). Video face recognition: component-wise feature aggregation network (c-fan). In International Conference on Biometrics.
20. Xie, W., & Zisserman, A. (2018). Multicolumn networks for face recognition.
21. Kim, M., Liu, F., Jain, A., & Liu, X. (2022). Cluster and aggregate: Face recognition with large probe set. In Neural Information Processing Sytems (NeurIPS).
22. Uzun, B., Cevikalp, H., & Saribas, H. (2023). Deep discriminative feature models (ddfms) for set based face recognition and distance metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 5594–5608.
23. Huang, R., Zhang, S., Li, T., & He, R. (2017). Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2458–2467.
24. Hu, Y., Wu, X., Yu, B., He, R., & Sun, Z. (2018). Pose-guided photorealistic face rotation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8398–8406.
25. Tian, Y., Peng, X., Zhao, L., Zhang, S., & Metaxas, D. N. (2018). Cr-gan: Learning complete representations for multi-view generation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence.
26. Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. (2017). Towards large-pose face frontalization in the wild. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 4010–4019.
27. Liu, Y., Shu, Z., Li, Y., Lin, Z., Zhang, R., & Kung, S. Y. (2022). 3d-fm gan: Towards 3d-controllable face manipulation. In European Conference on Computer Vision.
28. Zhou, H., Liu, J., Liu, Z., Liu, Y., & Wang, X. (2020). Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR).
29. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351 of LNCS, Springer, pages 234–241.
30. Wang, T. C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
31. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.
32. Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
33. Yalcin, M., Cevikalp, H., & Yavuz, H.S. (2015). Towards large-scale face recognition based on videos. In IEEE International Conference on Computer Vision Workshop
34. Cevikalp, H., & Saribas, H. (2023). Deep simplex classifier for maximizing the margin in both euclidean and angular spaces. In Scandinavian Conference on Image Analysis.
35. Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. In IVC.
36. Lee, K. C., Mo, J., Yang, M. H., & Kriegman, D. (2003). Video-based face recognition using probabilistic appearance manifolds. In IEEE Society Conference on Computer Vision and Pattern Recognition.
37. Klare, B., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., & Jain, A. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In IEEE Society Conference on Computer Vision and Pattern Recognition.
38. Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
39. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.
40. Heusel, M., Ramsauer, H., Unterthiner, T., & Nessler, B. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Neural Information Processing Sytems (NeurIPS).
41. Cao, K., Rong, Y., Li, C., Tang, X., & Loy, C. C. (2018). Pose-robust face recognition via deep residual equivariant mapping. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR).
42. Peng, B., Jin, X., Wu, Y., & Li, D. (2019). Geometry guided feature aggregation in video face recognition. In International Conference on Computer Vision Workshops.
43. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR).
44. Wei, X., Wang, H., Scotney, B., & Wan, H. (2020). Minimum margin loss for deep face recognition. *Pattern Recognition, 97*, 1–9.
45. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In IEEE Society Conference on Computer Vision and Pattern Recognition (CVPR).
46. Deng, J., Guo, J., Liu, T., Gong, M., & Zafeiriou, S. (2020). Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In European Conference on Computer Vision.