Normalized Face Image Generation with Perceptron Generative Adversarial Networks

Xiaofeng Liu[1,4], B.V.K Vijaya Kumar[1], Yubin Ge[1], Chao Yang[2], Jane You[3], Ping Jia[4]
[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA
[2]Viterbi School of Engineering, University of Southern California, Los Angeles, CA
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4]University of Chinese Academy of Sciences, Beijing, China
liuxiaofeng@cmu.edu,kumar@ece.cmu.edu,yug37@pitt.edu,harryyang@gmail.com,jiap@ciomp.cn

Abstract

This paper presents a deep neural architecture for synthesizing the frontal and neutral facial expression image of a subject given a query face image with arbitrary expression. This is achieved by introducing a combination of feature space perceptual loss, pixel-level loss, adversarial loss, symmetry loss, and identity-preserving loss. We leverage both the frontal and neutral face distributions and pre-trained discriminative deep perceptron models to guide the identity-preserving inference of the normalized views from expressive profiles. Unlike previous generative methods that utilize their intermediate features for the recognition tasks, the resulting expression- and pose- disentangled face image has potential for several downstream applications, such as facial expression or face recognition, and attribute estimation. We show that our approach produces photorealistic and coherent results, which assist the deep metric learning-based facial expression recognition (FER) to achieve promising results on two well-known FER datasets.

1. Introduction

With the rapid development of deep learning methods and a large amount of publicly available annotated face and facial expression images, unconstrained face recognition (FR) methods [1] and facial expression recognition (FER) technologies [2] have made significant advances in recent years. Although the performance we achieved on several benchmark datasets surpasses humans, the identity, pose and expression variations affecting the appearance of these images degrade the performance in many real-world applications.

Existing methods normally try to adopt hand-crafted or learned purely task-related features or recover a normalized face via synthesis techniques. For the first category, traditional methods often make use of robust local descriptors such as Gabor, Haar and local binary patterns (LBP) to account for local distortions [3]. In contrast, deep learning methods often handle invariance with pooling operations to ensure tolerance to very large intra-class variations [4]. However, due to the tradeoff between invariance and discriminability, they cannot deal with large pose cases effectively.

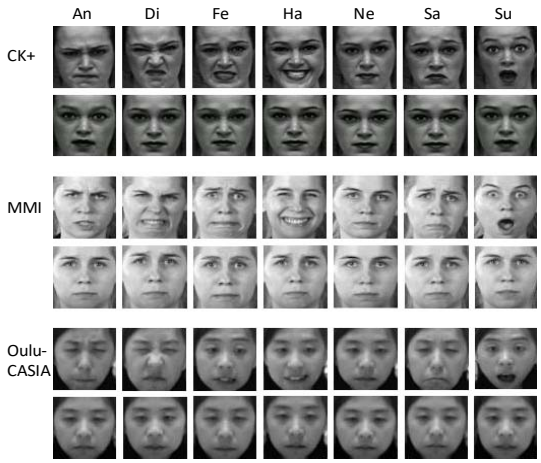

Figure 1. Input-output pairs of the proposed normalized face generative adversarial network (NFGAN) from the CK+, MMI and Oulu-CASIA datasets. Top row in each group: a subject in a database with different expressions (from left to right: angry, disgust, fear, happy, neutral, sad and surprise). Bottom row in each group: generated normalized face image in response to the input of the expressive face images in the top row.

The earlier efforts on face synthesis render a normalized view using the 3D geometrical transformations, which align the 2D plane image with a general [5] or identity specified [6] three-dimensional model. These approaches are good at normalizing small posed and subtle-expression faces, but their performance decreases largely under real-world applications due to severe texture loss and distortions. Recently, some deep learning based methods have been proposed to recover a frontal face image in a data-driven approach. For instance, Zhu et al. [7] propose to disentangle the identity and pose representation while learning to estimate a frontal

2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)
978-1-5386-2248-3/18/$31.00 © 2018 IEEE

view. The results are encouraging, but still lack fine details and tend to be blurry. Thus, only the intermediate features are used for face recognition task. Their synthesized face is still not good enough to be useful for other facial analysis tasks (e.g., attribute estimation and forensics).

The outstanding capability of Generative Adversarial Networks (GAN) in modeling 2D data distributions has significantly advanced many ill-posed low level vision problems, such as super-resolution [8] and inpainting [9]. We try to make the normalized face image generation process well constrained by representing the prior knowledge of the frontal and neutral face image distribution by a GAN [10]. Moreover, inspired by the symmetric structure of a face, a symmetry loss is proposed to fill out the occluded parts. We can recover the lost information inherent in projecting a 3D object to a 2D image space. Moreover, to faithfully preserve the most prominent facial structure, we adopt a perceptual loss [11] in the compact feature space in addition to the pixel-wise L1 loss. Incorporating the identity-preserving loss is helpful for a faithful synthesis and improves its potential to be applied to face analysis tasks. We show in Figure 1 some input samples and their generated results by the proposed normalized face-GAN (NFGAN).

Normally, the GAN schemes are not well-matched to supervised recognition tasks. The GAN-generated results are expected to align with the central part of the data distribution. The generated samples usually cannot support the network to adjust the boundary. In this paper, we utilize those results for the data augmentation of a deep metric learning facial expression recognition (FER) network (i.e., 2B-[N+M] Softmax [2]) to facilitate identity-disentangled FER.

The key contributions of this paper are: 1) We propose an end-to-end neural network architecture, termed NFGAN, to generate high quality identity-preserved normalized face image by combining prior knowledge from data distribution (adversarial training) and domain knowledge of faces (symmetry and identity-preserving loss). 2) The feature level similarity is measured for semantical perceptron and offer photorealistic details to the results. 3) We demonstrate a possible 'recognition via generation' framework and achieve state-of-the-art FER performance. Although several generative adversarial networks have been proposed for face synthesis, our approach is the first attempt to be effective for the FER task with synthesized faces.

2. Related work

As it is a classic topic in machine learning, several approaches have been proposed for generative models. Conventional approaches such as Gaussian Mixture Model (GMM), Principal Component Analysis (PCA), Independent Component Analysis (ICA), *etc*., have difficulty in modeling complex patterns of irregular distributions [12]. Recently, Restricted Boltzmann machines (RBM), Hidden Markov Model (HMM), Markov Random Field (MRF) *etc.,* have been employed for modeling images of digits, texture patches, and well-aligned faces [13]. However, their limited ability of feature representation restricts further development. Since the deep hierarchical architectures of the recent generative models appear to be capable of capturing complex structure of data, the generated images from these deep hierarchical structures are far more realistic.

The auto-encoder is a neural network that is trained to attempt to copy its input to its output. A denoising auto-encoder (DAE) pairs a differentiable encoder and decoder, which encodes an image sample x to a latent representation z and then decodes the z back to another image space $\tilde{x}$, which enables us to construct a transformation network [14]. For the normalized face generation task, the pose and expressions are regarded as the noise to be denoised. The main limitation of it is that the squared pixel-wise reconstruction error would cause the generated samples to look blurry as they lead to the generation of the mean image of the distribution [15]. The stochastic variational auto-encoder is proposed from the view of the directed graphical model [16]. A recognition network is used to approximate the intractable posterior with Gaussian latent variables. Gregor et al. [17] developed a recurrent auto-encoder with attention mechanism (DRAW) to generate images via a trajectory of patches.

Recently, the Generative Adversarial Network (GAN) [10] has been proposed to simultaneously train two networks: a generative network *Gen* to synthesize images (maps latents z to image space), and a discriminative network *Dis* to discriminate between real training images with generated images. *Dis* assigns a probability $\mathrm{y} = Dis(x) \in [0,1]$ when x is real and probability $1 - Dis(\tilde{x})$ when $\tilde{x}$ is fake. The *Gen* and *Dis* play a game to minimize or maximize the following binary cross entropy loss:

$$\mathcal{L}_{GAN} = \mathbb{E}_{x\sim data}[\log Dis(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - Dis(Gen(z)))] \quad (1)$$

with x denoting an input sample and $z\sim p(z)$. In practice, $\log(1 - Dis(Gen(z)))$ is easy to saturate early in learning for the *Dis* can reject the generated images with high confidence when *Gen* is poor. If we only train Gen to maximize $\log Dis(x)$, much stronger gradients could be offered by the *Dis*. With the GAN, an expected image can be generated from a randomly sampled vector z from a certain distribution. There are methods to combine the auto-encoder and GAN to achieve the image style transforms [18], which are related to and which have inspired our proposed architecture. However, it is still hard to utilize the generated results to improve a supervised recognition task for the GAN is always trying to model the central part of the data distribution, while the class boundary in feature space plays the vital role for classification. Limited attempts have been made in this potential area. The semi-GAN [19] adds an extra task for a discriminator network to improve semi-supervised recognition task. The face rotator schemes proposed by Tran [20]

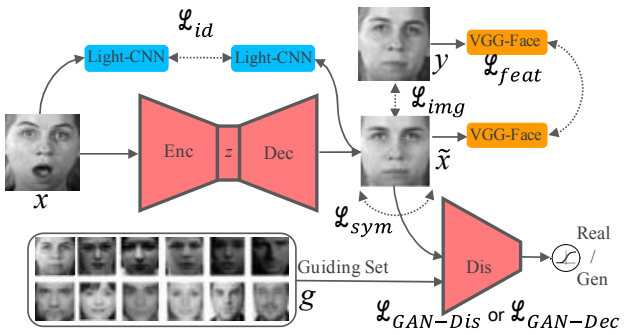

Figure 2. Framework of our normalized face generation scheme.

and Huang [21] generate a frontal face to as the preprocessing for the face recognition network. However, we utilized the results in the perspective of data augmentation and applied it for FER.

The proposed scheme is also related to the technologies for visualizing and exploring the properties of convolutional neural networks (CNNs). Mahendran and Vedaldi [22] proposed a framework to reconstruct an image from the CNN representation.

Compared with these works, our model keeps the whole identity-related content invariant, while disentangling the nuisance factors such as the pose and the expression. Both the prior knowledge from data distribution and domain knowledge of faces are incorporated. Moreover, feature level semantical similarity is further exploited to improve face generation quality.

3. Network architecture

The NFGAN as shown in Figure 2, is composed of 5 parts: 1) the Encoder network, (2) the Decoder network, (3) the Discriminator network, (4) the VGG-face network and (5) the Light CNN network. The function of the *Enc* and *Dec* network is the same as that in auto-encoder [14]. The *Enc* maps the input sample image x to a latent representation z through a learned distribution $P(z|x)$, while the *Dec* generates predicted a facial image $\tilde{x}$ corresponding to z. The function of the *Dec* and *Dis* is the same as that in the GAN [10]. The network of *Dec* tries to learn the real distribution by the *Dis* which trying to distinguish between generated image $\tilde{x}$ and real image in the guiding set g. We expand the contents of each model component in Section 4.

The structure of the *Enc* is shown in Table 1. Batch Normalization (BN) and non-linearity are removed from the last layer. We fixed the latent vector dimension to be 256 and found this configuration to be sufficient for generating images for FER. A series of fractional-stride convolutions (FConv) transforms the 256-dim vector $\mathrm{z} \in \mathbb{R}^{256}$ into a synthetic image $\tilde{x} \in \mathbb{R}^{64\times64}$, which is of the same size as x. Details of the decoder architecture are shown in Table 2. To further incorporate the prior knowledge of the frontal neutral face's distribution into the training process, we introduce a discriminator *Dis* to distinguish the generated face image from the real images in the guiding set. The decoder architecture of the *Dis* can be seen in Table 3.

There are two auxiliary networks in our architecture, their functions and structures are introduced in Section 4.1 and 4.4 and their impacts are discussed in subsection 5.3.

Table 1. Detailed Encoder architecture

Layer	Kernel/stride	Filters	BN	Activation
Conv1	3×3/2	32	No	ReLU
Conv2	3×3/2	64	Yes	ReLU
Conv3	3×3/2	128	Yes	ReLU
Conv4	3×3/2	256	Yes	ReLU
FC1	-	512	Yes	ReLU
FC2	-	256	No	None

Table 2. Detailed decoder architecture

Layer	Kernel/stride	Filters	BN	Activation
FConv1	3×3/2	512	Yes	ReLU
FConv2	3×3/2	256	Yes	ReLU
FConv3	3×3/2	128	Yes	ReLU
FConv4	3×3/2	64	Yes	ReLU
FConv5	3×3/2	1	No	Tanh

Table 3. Detailed discriminator architecture

Layer	Kernel/stride	Filters	BN	Activation
Conv1	3×3/2	64	No	LReLU
Conv2	3×3/2	128	Yes	LReLU
Conv3	3×3/2	256	Yes	LReLU
Conv4	3×3/2	512	Yes	LReLU
Conv5	3×3/1	1	No	Sigmoid

The Leaky ReLU nonlinearities are used in some Conv layers, where LReLU(x) = max(x,0) + α min(x,0). In our experiments, we set $\alpha = 0.3$. Optimizing this min-max objective function will continuously push the output of the generator to match the target distribution of the guiding set thus making the synthesized facial images to be more photorealistic.

4. Synthesis Loss Functions

In the training phase, the input-target pairs $\{x_i, y_i\}$ from multiple identities are required to learn the parameters θ of the differentiable encoder $\boldsymbol{\theta}_{Enc}$ and decoder $\boldsymbol{\theta}_{Dec}$, where x is a face image with expression and y is the frontal neutral face image for that person. Typical choices for the loss function are the squared Euclidean loss $\|x-y\|_2^2$ or $L1$ loss $\|x-y\|_1$. However, these may lead to overly blurry results. In this section, we show how the multiple objective functions are employed for different network parts to generate the photorealistic normalized face images. A group of five individual loss functions is used in our network to combine

the advantages of high quality GAN and stable auto-encoder which encodes the data into a latent space z. Each of the loss function corresponds to some specific network architectures, an overview of which is shown in Figure 2, and we will give the detailed descriptions in the following subsections.

4.1. Feature Space Perceptual Loss

Typically, the pixel-wise similarity measure is used in the image space to facilitate the image content consistency. However, it often yields overly smooth results. This is especially the case for the image prediction task which has inherent uncertainty in reconstructing an image from its feature representation. The precise location of all details may not be preserved in the features after the dimension reduction process by the encoder. Using the squared Euclidean distance in the image space corresponds to averaging over all possible locations, making reconstructions look blurry. The exact locations of all fine details are not essential for measuring the perceptual similarity of images. But the distribution of these details plays a key role in FER. Our main insight is that invariance to irrelevant transformations and sensitivity to local image statistics can be achieved by measuring distances in an alternate, suitable feature space. In fact, convolutional networks provide a feature representation with such desirable properties. They are invariant to small smooth deformations, but sensitive to perceptually important image properties, for example sharp edges and textures.

Instead of only requiring two images to be the same or similar at the pixel-level, we also require their similarity in the feature space. The squared error loss between the CNN feature representations is adopted to measure the feature-level perceptual loss. Based on earlier research [19], the measuring system may be a fixed or trained differentiable network and could be a part of the generator or discriminator. We make use of the independently-trained and fixed VGG-Face network to model this semantic feature-level loss. It is pre-trained on a very large scale face dataset and has been shown to yield excellent performance on face recognition tasks. Although it comprises 14 Conv layers and 3 FC layers, we omitted the deeper layers after the 5th Conv layer because their limited spatial resolution cannot support good image reconstruction. Denoted by φ_l, the feature map of the l^{th} convolution layer of VGG-Face are used to extract the feature representations using the standard forward-propagation process. The semantic perceptual loss between two images $\tilde{x}$ and y on the l^{th} convolutional layer is defined as the squared-error loss between the two feature maps.

$$\mathcal{L}_{feat}(\tilde{x}, y) = \frac{1}{W_l \times H_l} \sum_{n=1}^{W_l} \sum_{m=1}^{H_l} \left\| \varphi_{l,n,m}(\tilde{x}) - \varphi_{l,n,m}(y) \right\|_2^2 \quad (2)$$

where W_l and H_l denote the width, height of the l^{th} feature map, $\varphi_{l,n,m}$ is the value of the l^{th} feature map at point (n, m). In our experiment setting, l=5. $\mathcal{L}_{feat}$ alone does not provide a good loss for training. It is known that optimizing just for similarity in the feature space typically leads to high frequency artifacts [19]. This is because for each natural image there are many non-natural images mapped to the same feature vector. Therefore, a natural image prior is necessary to constrain the generated images to the manifold of natural images.

4.2. Symmetry Loss

Symmetry is an inherent feature of human faces. Exploiting this domain knowledge as a prior and imposing a symmetry constraint on the generated images may effectively alleviate the self-occlusion problem and thus improve performance for large pose cases. The symmetry loss of a face image takes the form

$$\mathcal{L}_{sym}(\tilde{x}) = \frac{1}{H \times W/2} \sum_{n=1}^{W/2} \sum_{m=1}^{H} \left| \tilde{x}_{n,m} - \tilde{x}_{W-(n-1),m} \right| \quad (3)$$

where W and H are the width and the height of the images and (n, m) denotes the pixel of the generated image. For simplicity, when training our model with the symmetry loss, all the inputs are aligned and detected with the occluded parts on the right side of image. If not, images are flipped so that the occluded parts are on the right side. Thus, we may transfer the appearance of the visible part to the occluded part easily by teaching the network to utilize the information on the left side of image to recover the lost information on the right side of image. The input face images are not required to be vertical since our deep neural networks can tackle this kind of simple transformations. However, the illumination changes and intrinsic textures are more complicated in the original pixel space. Real-world images may not exhibit the strict symmetry of the gray value. Considering the consistency of the pixel difference inside a local area, and the gradients of a point along all directions are largely preserved under different illuminations, defining a symmetry loss on the Laplacian space is a good way to capture the human face symmetry.

4.3. Adversarial Loss

Instead of manually designing a prior, as in [22], we learn it with an approach similar to GAN [10]. We introduce a discriminator *Dis* which serves as a supervisor to push the synthesized image to reside in the manifold of frontal neutral face images. It can prevent the blurry effect and produce visually pleasing results.

The *Dis* aims to discriminate the predicted frontal neutral face $\tilde{x}_i$ from real ones g_i in the guided set, and is trained concurrently with the transform network (*Enc* and *Dec*). The transform network tries to "trick" the *Dis* to classify the

generated images as real. Formally, the discriminator is trained to minimize the binary cross entropy:

$$\mathcal{L}_{GAN-Dis}(g_i, \tilde{x}_j) = -log(Dis(g_i)) - log\left(1 - Dis(\tilde{x}_j)\right) \quad (4)$$

with respect to *Dec*, the parameters of the generator are trained by minimizing:

$$\mathcal{L}_{GAN-Dec}(\tilde{x}_j) = -log\left(Dis(\tilde{x}_j)\right) \quad (5)$$

4.4. Identity-Preserving Loss

Synthesizing the frontal neutral face image while preserving the identity is a critical part of developing the NFGAN. By forcing the generated image and the target image (from the same person as the input image) to be similar in both the pixel and feature domains makes the transform network learn to maintain some identity information. However, there is no direct supervision to reward the perceptual similarity between input and generated images. As a high-level semantic concept, the identity is better preserved by penalizing the dissimilarity of their face verification network embedding. In our approach, we use the Light CNN, a compact network that has only 4 convolution layers with Max-Feature-Map operations and 4 max-pooling layers [24]. Despite the small size of the Light CNN, computing this loss adds considerable cost. See Section 5 for the discussion of its impact. In this work, the identity-preserving loss is defined based on the activations of the last two layers of the Light CNN:

$$\mathcal{L}_{id} = \sum_{l=1}^{2} \frac{1}{W_l \times H_l} \sum_{n=1}^{W_l} \sum_{m=1}^{H_l} \left| \phi_{l,n,m}(\tilde{x}) - \phi_{l,n,m}(x) \right| \quad (6)$$

where W_l, H_l denotes the width and height of the last l^{th} layer, $\phi_{l,n,m}$ is the value of the feature map (*n, m*) point. The identity-preserving loss enforces the prediction image to have a small distance with the input on the compact identity feature space. Since the Light CNN is pre-trained to classify tens of thousands of identities, it can capture the most prominent feature or face structure for identity discrimination. Therefore, it is possible to leverage this loss to enforce an identity-preserving frontal neutral face synthesis.

$\mathcal{L}_{id}$ has better performance when used with the adversarial loss. Using $\mathcal{L}_{id}$ alone makes the results prone to annoying artifacts, because the search for a local minimum of $\mathcal{L}_{id}$ may go through a path that resides outside the manifold of natural face images. Using adversary loss and $\mathcal{L}_{id}$ together can ensure that the search resides in that manifold and produces a photorealistic image.

4.5. Pixel-wise Loss

Adversarial training is known to be unstable and sensitive to hyper parameters. Adding the following pixel-wise L1 loss:

$$\mathcal{L}_{pixel} = \frac{1}{W \times H} \sum_{n=1}^{W} \sum_{m=1}^{H} \left| \tilde{x}_{n,m} - y_{n,m} \right| \quad (7)$$

in the image space with a relatively small weight is an important method to stabilize the training and accelerate the optimization, even though it will lead to blurry synthesis images [19]. $\tilde{x}_{n,m}$and $x_{n,m}$ are the pixel level gray value of the image's (*n, m*) point.

4.6. Limiting error signals to relevant networks

Using the aforementioned loss functions, we train the *Enc*, *Dec* and *Dis* simultaneously. This is possible because we do not update all network parameters to minimize a combination of all the loss function. In particular, the *Dis* should only try to minimize $\mathcal{L}_{Dis}$ as this could avoid collapsing the discriminator to 0. The error signal from adversarial loss and symmetry loss will not back-propagate to *Enc*.

Algorithm 1 Training the NFGAN

$\boldsymbol{\theta}_{Enc}, \boldsymbol{\theta}_{Dec}, \boldsymbol{\theta}_{Dis}$ ← initialize network parameters
Repeat
  $\boldsymbol{X}$ ← random mini-batch from dataset
  $\boldsymbol{Z} \leftarrow Enc(\boldsymbol{X})$
  $\tilde{\boldsymbol{X}} \leftarrow Dec(\boldsymbol{Z})$
  $\mathcal{L}_{feat} \leftarrow \frac{1}{2W_l \times H_l} \sum_{n=1}^{W_l} \sum_{m=1}^{H_l} \left\| \varphi_{l,n,m}(\tilde{x}) - \varphi_{l,n,m}(y) \right\|_2^2$
  $\mathcal{L}_{sym} \leftarrow \frac{1}{W/2 \times H} \sum_{n=1}^{W/2} \sum_{m=1}^{H} \left| \tilde{x}_{n,m} - \tilde{x}_{W-(n-1),m} \right|$
  $\mathcal{L}_{GAN-Dis} \leftarrow -log(Dis(g_i)) - log\left(1 - Dis(\tilde{x}_j)\right)$
  $\mathcal{L}_{GAN-Dec} \leftarrow -log(Dis(x_i))$
  $\mathcal{L}_{id} \leftarrow \sum_{l=1}^{2} \frac{1}{W_l \times H_l} \sum_{n=1}^{W_l} \sum_{m=1}^{H_l} \left| \phi_{l,n,m}(\tilde{x}) - \phi_{l,n,m}(x) \right|$
  $\mathcal{L}_{pixel} \leftarrow \frac{1}{W \times H} \sum_{n=1}^{W} \sum_{m=1}^{H} \left| \tilde{x}_{n,m} - y_{n,m} \right|$
  // Update parameters according to gradients
  $\boldsymbol{\theta}_{Enc} \leftarrow -\nabla_{\boldsymbol{\theta}_{Enc}} (\mathcal{L}_{feat} + \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{pixel})$
  $\boldsymbol{\theta}_{Dec} \leftarrow -\nabla_{\boldsymbol{\theta}_{Dec}} (\eta \mathcal{L}_{GAN-Dec} + (\mathcal{L}_{feat} + \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{pixel} + \lambda_3 \mathcal{L}_{sym}))$
  $\boldsymbol{\theta}_{Dis} \leftarrow -\nabla_{\boldsymbol{\theta}_{Dis}} (\mathcal{L}_{GAN-Dis})$
Until deadline

4.7. Weighting loss function

Several tradeoff parameters constrained between 0 and 1 are used to balance the aforementioned loss functions. The λ_1 and λ_2 as shown in Algorithm 1 are the tradeoff parameters of the $\mathcal{L}_{feat}$, $\mathcal{L}_{id}$ and $\mathcal{L}_{pixel}$ for the *Enc* and *Dec*. The λ_3 is used to weight the $\mathcal{L}_{sym}$ in *Dec*. As *Dec* also receives the error signal from the *Dis*, a parameter η is used to weight the ability of fooling the discriminator. This can also be interpreted as weighting prior knowledge from data distribution and domain knowledge of faces. The whole training procedure is described in Algorithm 1.

5. Experiments

In this section, we introduce the normalized face generation with NFGAN and report our experiment results on supervised recognition tasks (i.e., FER).

5.1. Datasets and preprocessing

A variety of large datasets of facial photographs for face recognition are publicly available. We adopt the database for VGG-Face network [24] training to further extend our data for neutral face generation. It contains approximately 2.6 Million face images, but very few of these fit our requirements of neutral expression, front-facing, no occlusion, and sufficient resolution for face region. We use the Google Cloud Vision API to remove those images that look blurry, with high emotion score or eyeglasses, tilt or pan angles beyond 5◦. The remaining images and few of their corresponding non-compliant images from the same subject are filtered to get about 12K target images (< 0.5% of the original set) and 50K input-target pairs. This dataset is used to pre-train the NFGAN network. The CMU Multi-PIE [25] contains more than 750,000 images from 337 people under fifteen viewpoints, and nineteen illumination conditions. There are four recording sessions in which subjects were instructed to display the neutral, happy, disgust and surprise facial expressions. We selected only the five groups of the nearly frontal view faces (-45° to +45°), giving us a total of 204156 images for pre-training. The 0° faces with neutral expression are used as the targets, while the other images from the same subjects are used as the inputs.

Besides the training datasets, we also evaluate the FER performance on CK+ and MMI dataset. The extended Cohn-Kanade database (CK+) [26] contains 593 videos from 123 subjects, while only 327 sequences from 118 subjects contain facial expression labels that range across 7 different expressions (i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise). The label is only provided for the last frame (peak frame) of each sequence. We split the CK+ database to 8 subsets in a strict subject-independent manner, and an 8-fold cross-validation is employed. The MMI Database [27] has 208 sequences captured in a front view and three frames in the middle of each image sequence are usually collected for static FER. We divided them into ten different subject-independent subsets.

For a raw image in the database, face registration is a crucial step for good performance. The bidirectional warping of Active Appearance Model (AAM) [28] and a Supervised Descent Method (SDM) called IntraFace model [32] are used to locate the 49 facial landmarks. Then, face alignment is done to reduce in-plane rotation and crop the region of interest based on the coordinates of these landmarks to a size of 64×64. The limited number of images of FER datasets is a bottleneck to deep model implementation. Thus, an augmentation procedure is employed to increase the number of training images and alleviate the chance of over-

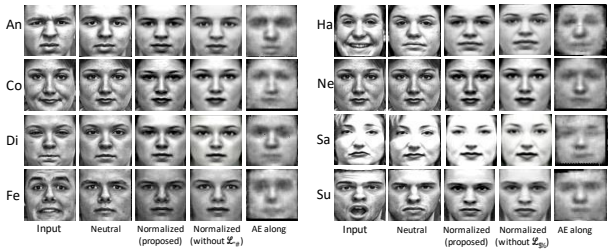

Figure 3. Examples of input, target, generated normalized face by RG network, generated normalized face by RG network without identity-preserving loss and reconstructed neutral face using only the auto-encoder (AE) structure [30].

fitting. We crop five 60×60 size patches from the center and four corners, flip them horizontally and transfer them to grayscale images. All the images are processed with the standard histogram equalization and linear plane fitting to remove unbalanced illumination. Finally, we normalize them to a zero mean and unit variance vector. In the testing phase, a single center crop with the size of 60×60 is used as input data.

5.2. Implementation Details

We use 64×64 gray images as the input-target pairs for the neutral face generation training. The filtered VGG-FaceNet and Multi-PIE images are used to pre-train the neutral face generation network. We construct the guided set using the filtered VGG-Facenet frontal neutral view and the 0° view neutral images from the Multi-PIE.

In all our experiments, we set $\eta = 0.1$, $\lambda_1 = 3 \times 10^{-3}$, $\lambda_2 = 10^{-3}$, $\lambda_3 = 0.3$ determined by manual tuning. As for the FER task, we follow the protocol in [2]. All the CNN architectures are implemented with the widely used deep learning tool "Tensorflow."

5.3. Impact of the auxiliary networks

The Light CNN and the first five layers of the VGG-FaceNet are used to embed the input, target or output images for the similarity measurements in different feature spaces. It is obvious that these two networks incur additional computation cost. We show in this section why they are needed.

The difference of our models trained with and without the $\mathcal{L}_{id}$ is subtle in visual appearance, as can be seen in Figure 3, but its effect on improving the identity likeness of the generated faces can be measured by evaluating the similarity of the input-outputs pairs using VGG-FaceNet. Figure 4 shows the distributions of L2 distances between the embeddings of the facial expression images and their corresponding synthesized results, for models trained with and without this loss. Schroff et al. [24] consider two FaceNet embeddings to encode the same person if their L2 distance is less than 1.242. All of the synthesized images using the identity-preserving loss pass this test using FaceNet, but about 2%

of the images would be identified as a different subject by FaceNet when not using the identity-preserving loss.

The VGG-FaceNet are employed to calculate the feature level perceptual loss, which is expected to make the generated result to keep more perceptually important image attributes, for example sharp edges and textures. This loss was empirically given the largest weight in our experiments. In practice, we failed to avoid the collapse of the adversarial training to generate the human face structure without this part. We observe the characteristics of the filters learned in the proposed networks. As shown in Figure 5, the *Enc* filters learned with feature level perceptual loss, (b), has more Gabor like edges than (a) without the feature level perceptual loss.

5.4. Application on FER

Recently, the 2B[N+M]Softmax is proposed to improve the FER. In this, the face images with the same expression label are expected to be close to each other in the feature space, while face images with different expressions are expected to be farther apart from each other by incorporating a deep metric learning scheme to disentangle the identity-related factors for FER. However, not all subjects reveal every expression in the dataset for similarity comparison. We supplement the normalized face image to the dataset in the training phase and observed the performance increasing and show the results in Table 4 and Figure 6.

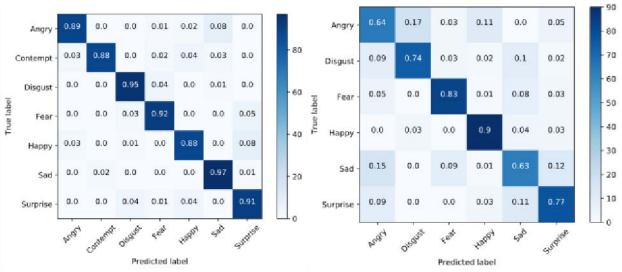

Figure 6. The confusion matrix obtained from the NFGAN augmented CK+ (left) and MMI (right) dataset using the 2*B*[*N*+*M*] Softmax FER method [2].

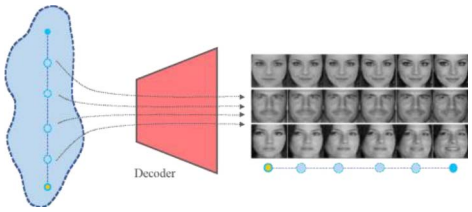

Figure 7. The blue and orange point denote the corresponding z and $\tilde{z}$ mapped from the input faces x and generated target image $\tilde{x}$ by the encoder, and the dotted line indicates the traversing from z to $\tilde{z}$. The intermediate circles along the traversing are supposed to generate a series of plausible morphing faces from x to $\tilde{x}$.

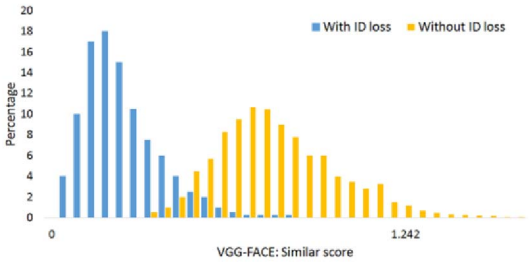

Figure 4. Histogram of VGG-Face net L2 error between the input face and the normalized pairs on the FER data collection. Blue: with the identity preserving loss which calculated by the Light CNN. Orange: without the identity preserving loss. The 1.242 threshold was used by Schroff et al. to cluster identities in the LFW dataset. Without the Light CNN, about 2% of the generated neutral faces would not be considered from the same subject as the query faces.

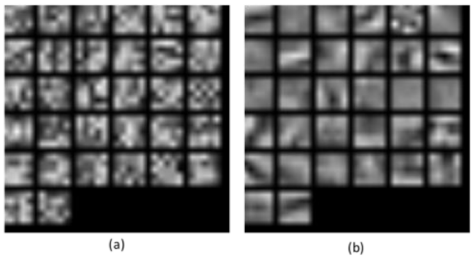

Figure 5: Learned the first Conv. layer's filters (32 filters) at the encoder networks examples on the CK+ database. (a) Conv. filters in *Enc* trained without the feature level perceptual loss, (b) Conv. filters in the proposed *Enc*.

Table 4. Recognition accuracy comparison

Dataset	State-of-the-Art Methods	NFGAN-2B[N+M]
CK+	97.25% [33], 97.1% [2]	97.49%
MMI	75.12% [34], 78.53% [2]	80.26%

5.5. Traversing in the manifold

Traversing the manifold that the autoencoder network learned can usually tell us about signs of memorization (if there are sharp transitions) and about the way in which space is hierarchically collapsed. If walking in this latent space results in semantic changes to the image generations (such as objects being added and removed), we can reason that the model has learned relevant and interesting representations. The results are shown in Figure 7. A vector was created from the query image and its generated normalized image via the encoder network. By adding interpolations along this axis, we were able to reliably transform their expressions. The video-based FER methods could be directly used to model the change and predict the expression labels of the query examples.

6. Conclusions and feature work

In this paper, we have proposed and investigated a novel strategy to disentangle the factors of expression and pose from the factors that are responsible for identity etc. As a novel identity-preserving neutral face generation scheme, it can synthesize the normalized face based on the prior knowledge of guiding set data distribution and domain knowledge of human face. We also show how the smooth expression change sequences can be generated by interpolation on the learned manifold, which may be compatible with video-based methods to explore the information. In future work, we intend to leverage the generative normalized face images for several downstream tasks (e.g., face recognition and attribute estimation).

Acknowledgements

The funding support from the China Scholarship Council, Youth Innovation Promotion Association(2017264), CAS, and Hong Kong Government General Research Fund GRF (No.152202/14E) is greatly appreciated.

References

[1] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In CVPR, 2014.

[2] X. Liu, B.V.K. Vijaya Kumar, J. You, P. Jia. Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition. Biometrics workshop, In CVPR 2017.

[3] X. Liu, L. Kong, Z. Diao and P. Jia. Line-scan system for continuous hand authentication. Optical Engineering, 56(3), 033106-033106, 2017.

[4] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In ICCV, 2013.

[5] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In CVPR, 2015.

[6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In CVPR, 2014.

[7] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In NIPS, 2014.

[8] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, arXiv preprint arXiv:1603.08155, 2016.

[9] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-Resolution Image Inpainting using Multi-Scale Neural Patch Synthesis. In CVPR, 2017.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.

[11] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, A. A. Efros, Learning a discriminative model for the perception of realism in composite images. In ICCV, 2015.

[12] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[13] Thrun, Sebastian. "Robotic mapping: A survey." Exploring artificial intelligence in the new millennium 1 (2002): 1-35.

[14] P. Vincent, H. Larochelle, Y. Bengio. Extracting and composing robust features with denoising autoencoders. In ICML 2008.

[15] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, W. Freeman. Synthesizing Normalized Faces from Facial Identity Features. In CVPR 2017.

[16] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114, 2013.

[17] K. Gregor, I. Danihelka, A. Graves, D. Wierstra, Draw: A recurrent neural network for image generation, arXiv preprint arXiv:1502.04623, 2015.

[18] L. Gatys, A. Ecker, M. Bethge. A Neural Algorithm of Artistic Style. In arXiv:1508.06576, 2015.

[19] Goodfellow, Ian. "NIPS 2016 tutorial: Generative adversarial networks." In arXiv: 1701.00160, 2016.

[20] L. Tran, X. Yin, X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In CVPR, 2017.A. Alpher. Frobnication. Journal of Foo, 12(1):234–778, 2002.

[21] R. Huang, S. Zhang, T, Li, R. He. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In arXiv:1704.04086v1, 2017.

[22] R. Huang, S. Zhang, T, Li, R. He. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In arXiv:1704.04086v1, 2017.

[23] X. Wu, R. He, Z. Sun, T. Tan. A Light CNN for Deep Face Representation with Noisy Labels. In arXiv: 1511.02683, 2017.

[24] P. Omkar, A. Vedaldi, and A. Zisserman. "Deep Face Recognition." In BMVC, 2015.

[25] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker. Multi-pie. Image and Vision Computing, 28(5):807–813, 2010.

[26] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews. The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression. In CVPRW, 2010.

[27] M. Pantic, M Valstar, R Rademaker, and L Maat. Web-based database for facial expression analysis. In ICME, 2005.

[28] M. Pantic, M Valstar, R Rademaker, and L Maat. Web-based database for facial expression analysis. In ICME, 2005.

[29] Y. Ge, X. Liu, Y. Chao and P. Jia. Radial Metric Learning with Generative References for Identity-Disentangled Facial Expression Recognition. Journal of Electronic Imaging, 2017

[30] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In CVPR, pages 532–539, 2013.

[31] X. Liu, B.V.K. Kumar, Y. Ge, L. Kong, J. You. Radial Metric Learning with Generative References for Identity-Disentangled Facial Expression Recognition, in FG 2018.

[32] Y. Kim, B. Yoo, Y. Kwak, C. Choi, J. Kim. Deep generative-contrastive networks for facial expression recognition. In arXiv:1703.07140, 2017.

[33] H. Jung, S. Lee, J. Yim, S. Park and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In ICCV 2015.

[34] A. Mollahosseini, D. Chan and M.H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In WACV, 2016.