Pattern Recognition 88 (2019) 1–12

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Hard negative generation for identity-disentangled facial expression recognition

Xiaofeng Liu[a,b,c,*], B.V.K. Vijaya Kumar[c,d], Ping Jia[a,b], Jane You[a,b,e]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China
[c] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA
[d] Carnegie Mellon University Africa, Kigali, Rwanda
[e] Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 22 April 2018
Revised 18 October 2018
Accepted 2 November 2018
Available online 3 November 2018

Keywords:
Hard negative generation
Adaptive metric learning
Face normalization
Facial expression recognition

ABSTRACT

Various factors such as identity-specific attributes, pose, illumination and expression affect the appearance of face images. Disentangling the identity-specific factors is potentially beneficial for facial expression recognition (FER). Existing image-based FER systems either use hand-crafted or learned features to represent a single face image. In this paper, we propose a novel FER framework, named *identity-disentangled facial expression recognition machine* (IDFERM), in which we untangle the identity from a query sample by exploiting its difference from its references (*e.g.*, its mined or generated frontal and neutral normalized faces). We demonstrate a possible 'recognition via generation' scheme which consists of a novel hard negative generation (HNG) network and a generalized radial metric learning (RML) network. For FER, generated normalized faces are used as hard negative samples for metric learning. The difficulty of threshold validation and anchor selection are alleviated in RML and its distance comparisons are fewer than those of traditional deep metric learning methods. The expression representations of RML achieve superior performance on the CK+, MMI and Oulu-CASIA datasets, given a single query image for testing.

1. Introduction

Facial expression is the most natural and expressive nonverbal channel for humans to communicate their emotions [1]. Therefore, facial expression recognition (FER) has been an important and active topic for a wide range of applications including soft biometrics, digital entertainment, health care, robot systems and human-computer interaction (HCI). Ekman and Friesen postulated the universality of neutral (Ne) and six prototypical human facial expressions, namely, anger (An), disgust (Di), fear (Fe), happiness (Ha), sadness (Sa) and surprise (Su) [2].

The performances of the FER systems usually depend heavily on facial expression representations, which are affected by pose and illumination variations as well as facial morphology variations (i.e., identity-specific factors). As some facial expressions involve subtle facial muscle movements, the extracted expression-related information from different classes (in this paper, class refers to expression) can be overwhelmed by high-contrast identity-specific geometric or appearance features degrading FER performance. As illustrated schematically in Fig. 1, we want the intra-class distances of happy face images from different people to be smaller than the inter-class distances between face images of different expressions from the same subject. However, the nuisance identity factors often dominate the representation of the image in pixel space causing two images of the same subject with different expressions to be closer to each other than the same-expression images from two different subjects. This is because the extracted facial representation often contains identity-specific information that is irrelevant and may be counter-productive for the FER task. These identity-specific factors may degrade the FER performance on new identities unseen in the training data.

Aided by the advances in deep learning for computer vision [3], much progress has been made on extracting a set of features to represent a single facial expression image [4-5]. The hand-crafted features are constructed by exploiting domain knowledge of the specific relationships within pixels so that the features are invariant to some simple transformations (*e.g.*, translation and scaling). More recently, feature-learning approaches are being investigated because of their ability to produce features that are tolerant to complex transformations. In the case of FER, identity-associated

* Corresponding author.
E-mail addresses: liuxiaofeng@cmu.edu (X. Liu), kumar@ece.cmu.edu (B.V.K. Vijaya Kumar), jiap@ciomp.ac.cn (P. Jia), csyjia@comp.polyu.edu.hk (J. You).

https://doi.org/10.1016/j.patcog.2018.11.001
0031-3203/

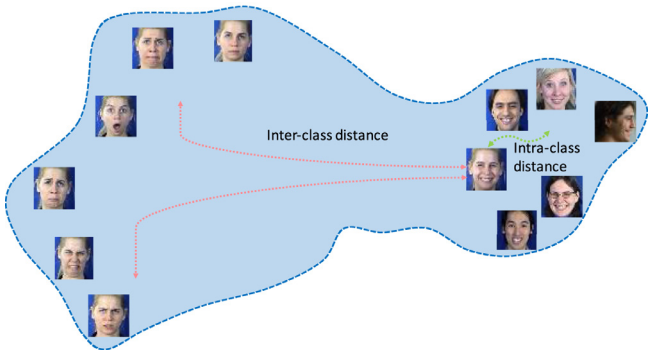

Fig. 1. Illustration of the desired representations in the Face Expression Recognition (FER) feature space. The "class" here refers to the facial expression.

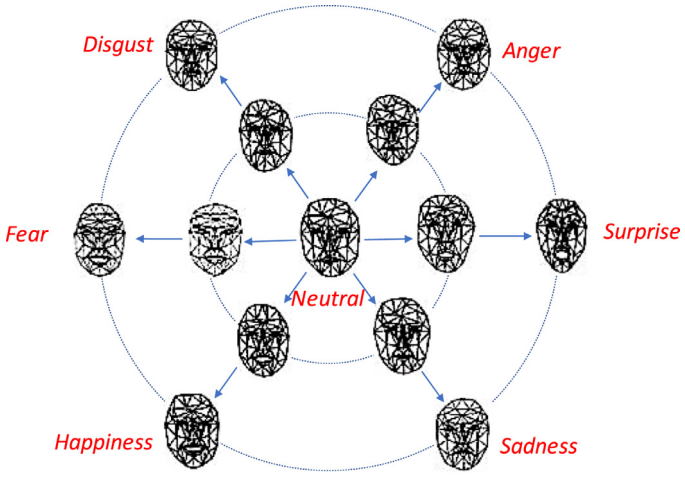

Fig. 2. A schematic depicting the 6 basic facial expressions and their relationships to the neutral face. Representations in the outer ring correspond to higher-intensity facial expressions compared to those in the inner ring.

factors may fall into this category. Yet, unfortunately, due to the tight coupling of the various nuisance factors, when we try to reduce the sensitivity to identity with these state-of-the-art strategies, we are unlikely to satisfactorily disentangle all variations in facial morphology.

Our preliminary work [6,7] proposed to improve the facial expression recognition performance by disentangling the identity factors in a face image through the metric learning method. To alleviate the slow convergence caused by trivial training samples in metric learning, [6] compared the query image to a negative set containing other facial expression images from the same subject as shown in Fig. 3. However, in practice, the structure of real-world FER datasets results in a significant constraint: the dataset may not contain images of every facial expression for every subject. Actually, we may not need to compare a query facial expression with any other class of facial expressions. According to some psychology and anatomy research [8], the muscle activities of different facial expressions initiate from the neutral face as illustrated schematically in Fig. 2. This is also the fundamental principle underlying the action units (AUs) and Facial Action Coding System (FACS) proposed by Ekman [2]. Since the expressive classes are naturally more discriminative from each other than from the neutral face, in the training stage, we may want to emphasize the neutral-expressive distance more than requiring large distance between those expressive classes. The neutral face images can be the ideal hard samples which can improve the learning efficiency of metric learning. However, expressive-neutral face image pairs of every person may not be always available in real world applications and in some FER image datasets.

The above insights suggest that we should generate the frontal and neutral normalized face image of the query image and disentangle the possibly counter-productive identity information and we proposed *identity-disentangled facial expression recognition machine* (IDFERM) towards this goal. IDFERM consists of two main parts, namely, the hard negative generation (HNG) network and the radial metric learning (RML) network. Specifically, given a query facial expression image, its normalized reference will be synthesized using an HNG network trained using expression-normalize face image pairs of the same subject. Then, the query-reference pairs are fed into the radial metric learning (RML) network, which uses an inception style convolutional (Conv) layers group and a unified two-branch fully connected (FC) layers framework to extract the contrast of query-reference pair by simultaneously optimizing the softmax loss and RML loss. By pushing the representation of facial expression images away from their generative references and pulling them close to their cluster centers of each expression, the RML can disentangle those nuisance factors to balance the intra- and inter- class variations.

Unlike other image-based FER systems that produce a representation from a single input image, RML utilizes the expression-reference pair to disentangle the identity-specific factors. In contrast to video-based FER methods [9,10] or real facial expression image pair-based methods [6,7,11], our method employs synthetically generated references to address the real-world limitation that sometimes the dataset does not contain all possible facial expression examples for some subjects.

The preliminary versions of the concepts in this paper were published in the 2017 Biometrics workshop of *IEEE Conference on Computer Vision and Pattern Recognition* [6] and 2018 IEEE International Conference on *Identity, Security and Behavior Analysis* [7]. In this paper, we extend those basic concepts in the following ways:

1) We investigate the prior relationship of different expression classes and propose a novel recognition via generation scheme as a possible substitution of conventional hard sample mining.
2) We design an end-to-end IDFERM to extract identity-disentangled representations for FER without requiring real expression-neutral pair inputs in the FER datasets.
3) The number of needed distance comparisons for adapted RML is orders of magnitude smaller than the number needed for conventional metric learning approach.
4) We conduct all experiments using the new architecture and test on more datasets without neutral expression samples.

In summary, this paper makes the following contributions.

- We propose a generalized metric learning loss function with adaptively learned reference threshold which alleviates the difficulty of threshold validation and anchor selection.
- With the identity-aware HNG for hard negative generation, the adapted RML learns distance metrics with fewer input iterations and distance calculations, without sacrificing the performance for identity-invariant FER.
- We jointly optimize the softmax loss and metric learning loss in a unified two-branch FC layer metric learning CNN framework based on their characteristics and tasks.
- The proposed HNG network is a novel approach to generate photorealistic and identity-preserved normalized face image by combining prior knowledge from data distribution and domain knowledge of faces.
- Using numerical experiments, we demonstrate that the proposed method achieves promising results on CK+, MMI and Oulu-CASIA data sets.

The rest of this paper is organized as follows. Section 2 briefly reviews related work in the literature. Section 3 introduces in detail the proposed normalized face generation with perceptron generative adversarial networks. Section 4 shows its application to FER with RML network. Section 5 reports the experimental results and ablation study of the auxiliary parts in HNG network. Finally, Section 6 provides our conclusions.

2. Related work

Despite receiving considerable research attention, FER remains very challenging [12]. Research developments in deep learning, especially the success of convolutional neural networks (CNN), have made high-accuracy image classification possible in recent years. Deep learning-based FER methods have emerged starting with Bengio et al. [13] who described the use of carefully designed CNN to learn expression features from raw pixels. Despite its popularity, current softmax loss-based approach does not explicitly reward intra-class compactness and inter-class separation, and identity-related factors remain major obstacles for FER. Machine recognition usually is based on similarity metrics, but those metrics may be more sensitive to identity than expressions. To decouple these two types of similarity and exploit the appearance information, substantial efforts have been dedicated to extracting features by learning [14]. Given that the expressions are formed by relaxing or contracting some facial muscles that result in temporally deformed facial features, identity-disentangled representations for FER normally separate a face with expression into a main component neutral face that encodes identity cues and an action component that encodes motion cues (such as movements of eye brows, cheeks, lips, eyelids and nose) which are related to the AUs and FACS [2].

FER is certainly not unique among computer vision applications that have to cope with nuisance factors causing variability in the data. Deep metric learning approaches have been shown to be successful for person and vehicle identification tasks [15-17], which also exhibit large intra-class variations. The initial work in this domain [18] involves training a Siamese network. Pairwise examples are fed into two symmetric sub-networks and the network is updated using contrastive loss function, *i.e.*, their extracted representations should be close to each other if the inputs have the same class label, otherwise the distance between these representations should be large. One improvement is the triplet loss [19], in which, the inputs are triplets, each consisting of a query, a positive example and a negative example. An anchor is chosen from the query or positive examples, then the method requires the difference of the distance from the anchor point to the positive or query example and from the anchor point to the negative example to be larger than a fixed margin τ. Recently, some variants of this offering faster and more stable convergence have been developed. The $(N+1)$-tuplet loss [20] incorporated multiple negative examples while the coupled cluster loss (CCL) [15] incorporated multiple positive examples in a tuplet. The center of positive examples c^+ is set as the anchor in CCL. By comparing each example with c^+ instead of each other, the number of distance evaluations needed are reduced significantly.

For the situation shown in Fig. 3, the triplet loss, $(N+1)$-tuplet loss and CCL are all 0, since the distances between the anchor and positive examples are indeed smaller than the distance between the anchor and negative examples for a margin τ. This means the loss function will learn to neglect such non-trivial samples. We will need many more input iterations with properly selected anchors to correct this situation. The fixed threshold in the contrastive loss was also proven to be sub-optimal [21]. The difficulty of threshold validation and anchor selection have long been significant challenges until our initial work [6], which included an adaptive $(N+M)$-tuplet clusters loss function.

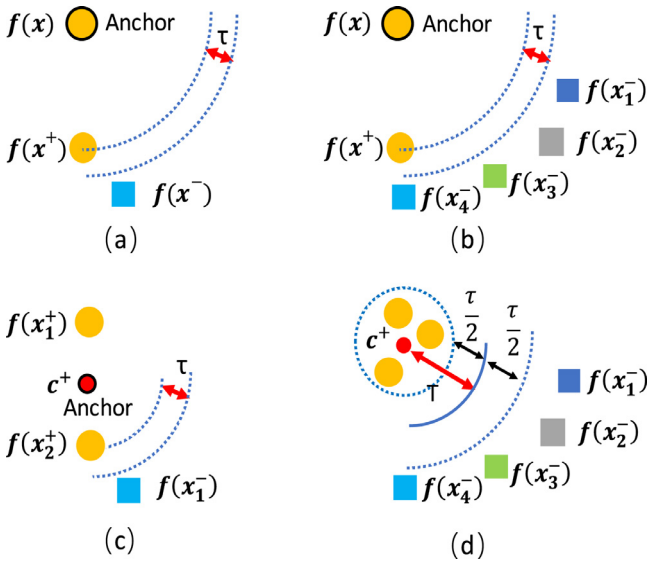

Fig. 3. Failed case of (a) triplet loss, (b) $(N+1)$-tuplet loss, and (c) Coupled clusters loss. The preliminary $(N+M)$-tuplet clusters loss (d) uses two thresholds to avoid the anchor selection issue and threshold-parameters *i.e.*, T and τ, do not need manual tuning [6,7]. We use x^+ (yellow points) and x^- (squares) to denote the positive and negative examples of a query example x, meaning that $x^+$is the same class of x, while $x^-$is not.$f(\cdot)$ is an embedding kernel.

Also, the traditional online or offline mini-batch sample selection is a large additional computational burden and can result in poor local optima [22]. Generating all possible pairs or triplets would result in quadratic and cubic complexity, respectively and most of these pairs or triplets are not very useful for the training [6]. Our initial work [6] utilized identity-aware hard-negative mining and online positive mining for FER, but it still suffers from the dataset-sensitive and computationally-expensive example mining to provide nontrivial tuplets.

Several approaches have been proposed for generative models. Conventional methods such as Principal Components Analysis (PCA), Independent Components Analysis (ICA), Gaussian Mixture Model (GMM), *etc.*, have difficulty in modeling complex patterns of irregular distributions [23]. 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 [24]. However, the limited ability of feature representations restricts further development. Since deep hierarchical architectures of the recent generative models are capable of capturing complex structure of data, generated images from these deep hierarchical structures are more realistic. The 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 $\tilde{x}$ [25]. For the normalized face generation task, pose and expression are regarded as the noise to be mitigated. The main limitation of this approach is that the squared pixel-wise reconstruction error would cause the generated samples to look blurry as they generate the mean image of the distribution. Generative Adversarial Network (GAN) [26] simultaneously trains 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 from generated images. With the GAN, an expected image can be generated from a randomly sampled vector z from a certain distribution.

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, while the boundary

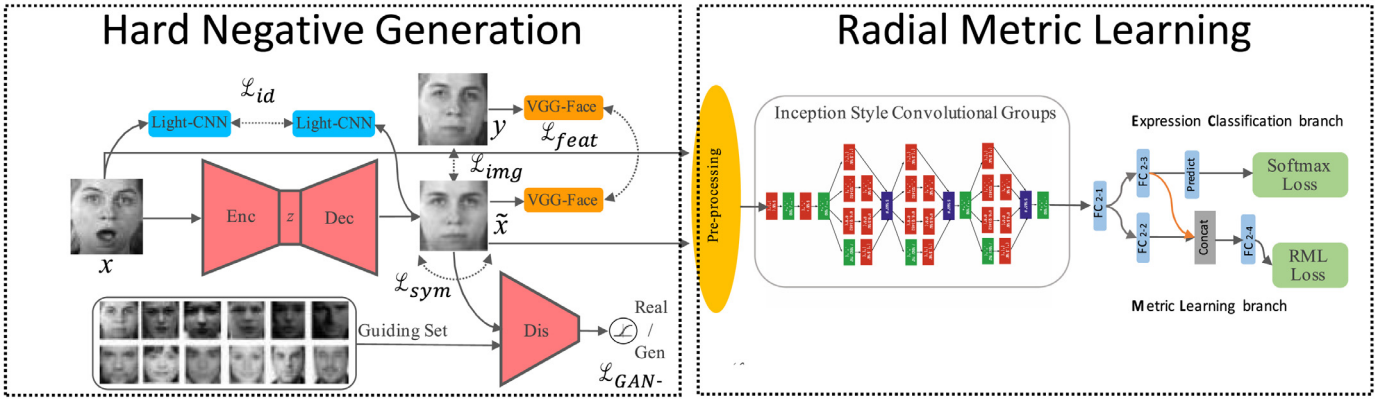

Fig. 4. Framework of our IDFERM, in which, the left and right side are the HNG and RML network, respectively. We feed the input image x to an encoder-decoder structure and generate its transformed output $\tilde{x}$. The *Dis* is trained to determine if its input is from the guiding set (real image) or the encoder-decoder structure (generated image), thus encouraging the encoder-decoder structure to generate images more similar to the images in our guiding set. The guiding set contains all of the target images y, which is a compilation of numerous frontal and neutral real face images. The Light-CNN is used to extract the identity feature for identity similarity measurements, and the VGG-Face is adopted to embed image for feature level perceptual similarity measurements. The mined real expressive face images and their corresponding generted normalized face images are feeded to the RML, an adaptive deep metric learning framework, to disentangle the identity factors in a face image for FER by minimazing both the softmax (cross-entropy) loss and RML loss.

between classes in feature space is more important for classification. Limited research has been devoted to this topic. The semi-GAN [27] adds an extra task for a discriminator network to improve semi-supervised recognition task. The face rotator schemes proposed by Tran [28] generates a frontal face as the preprocessing for the face recognition network.

3. Hard negative generation

As we are trying to disentangle identity-related factors from a facial expression image x, a reference neutral face image from the same subject is required to obtain the difference between the neutral face and x for FER. However, such a neutral face reference image is not always available in real world application scenarios. Instead of mining several negative samples, we directly use the generated normalized face image as negative sample. The goal of the hard negative generation (HNG) network is to produce a photo-realistic and identity-preserved normalized face image $\tilde{x}$ from the probe image x. The network architecture and loss functions are illustrated in Fig. 4.

Our HNG network is composed of five major components: 1) an Encoder network *Enc*, (2) a Decoder network *Dec*, (3) a Discriminator network *Dis*, (4) the *Light CNN* network and (5) the *VGG-facenet*. The function of *Enc* and *Dec* network is the same as that in denoising auto-encoder [25]. In DAE, the output is not required to be exactly the same as the input. For example, the denoising auto-encoder takes in an image that has been corrupted by some form of "noise", and is forced to output a denoised version of that image by requiring the output image to be similar to the original "clean" image. In our application, a face image with expression (input image x) can be regarded as a copy of neutral face image (target image y) that has been corrupted by expressions. The denoising (disentangling expression from a face image) is achieved by requiring our output $\tilde{x}$ to be close to the target image y, as the DAE requires its generated image to be close to a clean target image instead of the original noisy image. 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 [26]. The *Dec* network tries to generate the real distribution by the loss of *Dis* which learns to distinguish between generated image $\tilde{x}$ and real image in the guiding set g. The guiding set contains all of the target images y, which is a compilation of numerous frontal and neutral real face images. The input-target pairs $\{x_i, y_i\}$ from multiple identities are required to learn the parameters θ of the differentiable encoder θ_{Enc} and decoder θ_{Dec}, where x is a face image with expression and y is the frontal neutral face image of that person. In our experimental setting, five different loss functions are used to combine the advantages of high quality GAN and stable auto-encoder which encodes the data into a latent space z. In this section, we show how the multiple objective functions are employed for different parts to generate the facial reference images for FER.

3.1. Feature space perceptual loss

The squared error loss between the CNN feature representations is adopted to represent the feature-level perceptual loss. We make use of the independently-trained and fixed VGG-FaceNet [29] to model this semantic feature-level loss. 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 performance. Denoted by φ_l, the feature map of the l^{th} convolution layer of VGG-FaceNet is 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 following squared-error loss between the two feature maps.

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

where W_l and H_l denote the width, height of the l^{th} feature map and $\varphi_{l,\,n,\,m}$ is the value of the l^{th} feature map at point (n, m). In our experiment, l=5 based on empirical testing.

3.2. Symmetry loss

Symmetry is an inherent property of normal human faces. The symmetry loss of a face image takes the form:

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

where W and H are the width and the height of the images and (n, m) denotes the pixel of the generated image. The $|\bullet|$ denotes the absolute value. For simplicity, when training our model with the symmetry loss, all the inputs are aligned and detected if the occluded parts are on the right side of image. If not, images are flipped so that the occluded parts are on the right side. Real-world

images may not exhibit the strict symmetry of pixel values. Considering the consistency of the pixel difference inside a local area, and the gradients at a point along all directions are largely preserved under different illuminations, minimizing a symmetry loss in the Laplacian space should emphasize human faces.

3.3. Adversarial loss

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 reduce the blur effect and produce visually pleasing results.

The *Dis* aims to discriminate the predicted frontal neutral face image $\tilde{x}_i$ from real ones g_i in the guiding 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 following binary cross entropy loss:

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

With respect to *Dec*, the parameters are trained by minimizing the following loss:

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

3.4. Identity-Preserving loss

Synthesizing the frontal neutral face image while preserving the identity is a critical part of IDFERM. We introduce a direct supervision to reward the perceptual similarity between input and generated images using the face verification network. In our approach, we use the pre-trained Light CNN, a compact network that has only 4 convolution layers with Max-Feature-Map operations and 4 max-pooling layers [30]. In this work, the identity-preserving loss is defined based on the activations of the last two layers of the Light CNN:

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

where W_l, H_l denotes the width and height of the l^{th} layer, $\phi_{l,\,n,\,m}$ is the value of the feature map (n, m) point and $|\bullet|$ denotes the absolute value.

3.5. Pixel-wise loss

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

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

in the image space with a relatively small weight is one method to stabilize the training and accelerate the optimization. $\tilde{x}_{n,m}$ and $x_{n,\,m}$ are the pixel level gray values of the $(n, m)^{\text{th}}$ pixel.

Using judicious selection of aforementioned loss functions, we train the *Enc, Dec* and *Dis* simultaneously. The error signal from adversarial loss and symmetry loss are not back-propagated to *Enc*. Several tradeoff parameters constrained between 0 and 1 are used to balance the aforementioned loss functions. The weights λ_1 and λ_2 shown in Algorithm 1 are the tradeoff parameters for $\mathcal{L}_{feat}$, $\mathcal{L}_{id}$ and $\mathcal{L}_{pixel}$ for the *Enc* and *Dec*. The parameter 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.

We show some of the input-output pairs of our HNG network in Fig. 5. As the common quantitative metrics (e.g., log-likelihood of a set of validation samples) are often not very informative for perceptual generative models [31], we provide a qualitative comparison of visual quality and a quantitative evaluation of identity-preservation in Section 5.

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.

Algorithm 1 Training the HNG network.

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

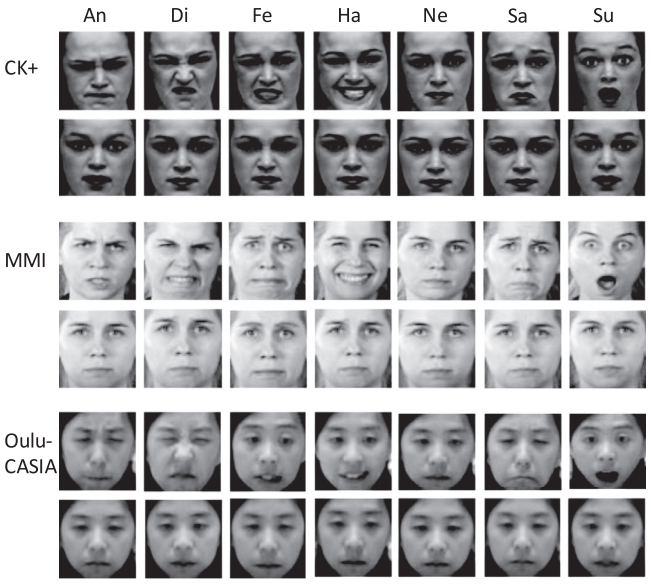

Fig. 5. Input-output pairs of the proposed reference generation (HNG) network from the CK+, MMI and Oulu-CASIA dataset. Top row in each pair: a subject in a database with different expressions (from left to right: angry, disgust, fear, happy, neutral, sad and surprise). Bottom row in each pair: generated normalized face images from the input of the expressive face images in the row above.

4. Radial metric learning

The proposed RML only requires the comparison of the representation of the query sample f_i with the representation of its generated reference $\tilde{f}_i$ and its cluster center C_{y_i}. We introduce a distance T from the query sample x to control the relative boundary $(T - \frac{\tau}{2})$ and $(T + \frac{\tau}{2})$ for the intra-class center and generated references, respectively. The RML loss function is formulated as follows.

$$\mathcal{L}\left(\{x_i\}_{i=1}^{K},\ \{\tilde{x}_i\}_{i=1}^{K}; f\right) = \frac{1}{K} \sum_{i=1}^{K} \left\{ max\left(0, \mathrm{D}(f_i, C_{y_i}) - T + \frac{\tau}{2}\right)\right.$$

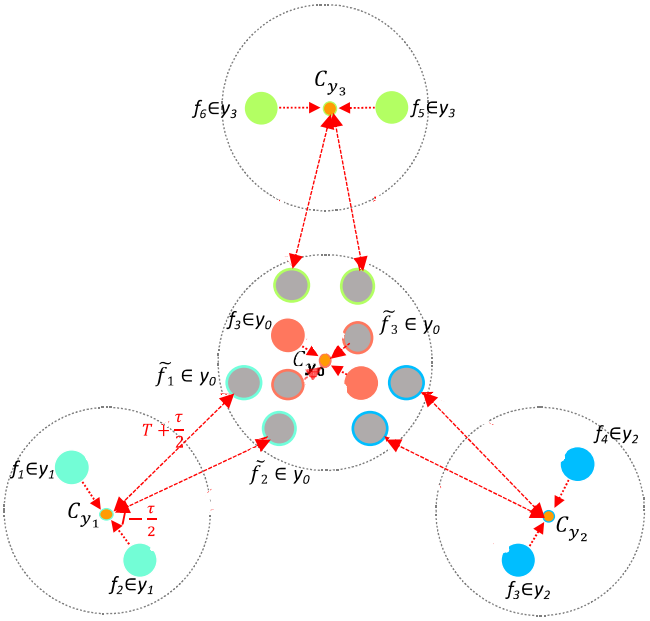

Fig. 6. The proposed radial metric learning (RML) framework. A small circle without border is the representation of a sample (*i.e.*, facial expression image) and the different classes are represented by different colors. The small gray circles with colored border are their corresponding generated references (i.e., normalized face images). The orange points with colored border are the cluster centers of each classes. The big dashed circles are the boundaries of each classes in the feature space, which are expected to have small radius and far away from each other.

$$+ max\Big(0, \frac{\tau}{2} + T - \mathrm{D}\big(\tilde{f}_i, C_{y_i}\big)\Big)\Big\} \tag{8}$$

Only if the distances from all online mined examples f_i to its updated C_{y_i} are smaller than $(T - \frac{\tau}{2})$ and the distances from all the generated references $\tilde{f}_i$ to its updated C_{y_i} are larger than $(T + \frac{\tau}{2})$, the loss $\mathcal{L}(\{x_i\}_{i=1}^K, \{\tilde{x}_i\}_{i=1}^K; f)$ can get a zero value. A simplified geometric interpretation of this is shown in Fig. 6.

By assigning different values for T and τ, we define a flexible learning task with adjustable difficulty for the network. We do not use the special case that requires inter-class variation to be zero (i.e., $T = \tau/2$) as the center loss [32] for the FER training set usually contains some unreliable labels [33]. However, these two hyper-parameters need manual tuning and validation. In here, we formulate the reference distance T to be a function $\mathrm{S}(\cdot, \cdot)$ which should be trained automatically, instead of a constant. Inspired by the Mahalanobis distance matrix M in Mahalanobis distance D (Eq. 9), which is a positive semi-definite (PSD) matrix and can be calculated via the linear fully connected layer as in [34], we try to automatically train both S and D. Since the difference of the reference distance and the distance function need to be calculated in two terms in Eq. 8, a possible solution is to calculate (S−D) function via the linear FC layer.

$$\mathrm{D}(f_1, f_2) = \|f_1 - f_2\|_{\mathrm{M}}^2 = (f_1 - f_2)^{\mathrm{t}}\mathrm{M}(f_1 - f_2) \tag{9}$$

Since the metric M itself is quadratic, we assume that S has a simple quadratic form:

$$\mathrm{S}(f_1, f_2) = \frac{1}{2}f_1{}^{\mathrm{t}}\tilde{\mathrm{A}}f_1 + \frac{1}{2}f_2{}^{\mathrm{t}}\tilde{\mathrm{A}}f_2 + f_1{}^{\mathrm{t}}\tilde{\mathrm{B}}f_2 + c^{\mathrm{t}}(f_1 - f_2) + b \tag{10}$$

where $\tilde{\mathrm{A}}$ and $\tilde{\mathrm{B}}$ are both the $d \times d$ real symmetric matrices (not necessarily positive semi-definite), c is a D-dimensional vector, and b is the bias term.

Then, a new quadratic expression $\mathrm{H}(f_1, f_2) = \mathrm{S}(f_1, f_2) - \mathrm{D}(f_1, f_2)$ is defined to combine the reference distance function S and the Mahalanobis distance metric function D. Substituting $\mathrm{S}(f_1, f_2)$ and $\mathrm{D}(f_1, f_2)$ into $\mathrm{H}(f_1, f_2)$, we get:

$$\mathrm{H}(f_1, f_2) = \frac{1}{2}f_1{}^{\mathrm{t}}\big(\tilde{\mathrm{A}} - 2\mathrm{M}\big)f_1 + \frac{1}{2}f_2{}^{\mathrm{t}}\big(\tilde{\mathrm{A}} - 2\mathrm{M}\big)f_2 + f_1{}^{\mathrm{t}}\big(\tilde{\mathrm{B}} + 2\mathrm{M}\big)f_2 + c^{\mathrm{t}}(f_1 - f_2) + b \tag{11}$$

$$\mathrm{H}(f_1, f_2) = \frac{1}{2}f_1{}^{\mathrm{t}}\mathrm{A}f_1 + \frac{1}{2}f_2{}^{\mathrm{t}}\mathrm{A}f_2 + f_1{}^{\mathrm{t}}\mathrm{B}f_2 + \mathrm{c}^{\mathrm{t}}(f_1 - f_2) + b \tag{12}$$

where $\mathrm{A} = (\tilde{\mathrm{A}} - 2\mathrm{M})$ and $\mathrm{B} = (\tilde{\mathrm{B}} + 2\mathrm{M})$. Suppose A is PSD and B is negative semi-definite (NSD), A and B can be factorized as $\mathrm{L}_{\mathrm{A}}^{\mathrm{T}}\mathrm{L}_{\mathrm{A}}$ and $\mathrm{L}_{\mathrm{B}}^{\mathrm{T}}\mathrm{L}_{\mathrm{B}}$. Then $\mathrm{H}(f_1, f_2)$ can be rewritten as follows:

$$\mathrm{H}(f_1, f_2) = \frac{1}{2}f_1{}^{\mathrm{t}}\mathrm{L}_{\mathrm{A}}^{\mathrm{t}}\mathrm{L}_{\mathrm{A}}f_1 + \frac{1}{2}f_2{}^{\mathrm{t}}\mathrm{L}_{\mathrm{A}}^{\mathrm{t}}\mathrm{L}_{\mathrm{A}}f_2 + f_1{}^{\mathrm{t}}\mathrm{L}_{\mathrm{B}}^{\mathrm{t}}\mathrm{L}_{\mathrm{B}}f_2 + c^{\mathrm{t}}(f_1 - f_2) + b \tag{13}$$

$$\mathrm{H}(f_1, f_2) = \frac{1}{2}(\mathrm{L}_{\mathrm{A}}f_1)^{\mathrm{t}}(\mathrm{L}_{\mathrm{A}}f_1) + \frac{1}{2}(\mathrm{L}_{\mathrm{A}}f_2)^{\mathrm{t}}(\mathrm{L}_{\mathrm{A}}f_2) + (\mathrm{L}_{\mathrm{B}}f_1)^{\mathrm{t}}(\mathrm{L}_{\mathrm{B}}f_2) + c^{\mathrm{t}}(f_1 - f_2) + b \tag{14}$$

Motivated by the above, we propose a general, computationally feasible loss function. Following the notations in the preliminaries and denoting $(\mathrm{L}_{\mathrm{A}}, \mathrm{L}_{\mathrm{B}}, c)^{\mathrm{t}}$ as W which can be learned via the linear fully connected layer, we have:

$$\mathcal{L}\big(\mathrm{W}, \{x_i\}_{i=1}^K, \{\tilde{x}_i\}_{i=1}^K; f\big) = \frac{1}{K}\sum_{i=1}^K \Big\{max\Big(0, \frac{\tau}{2} - \mathrm{H}\big(f_i, C_{y_i}\big)\Big) + max\Big(0, \mathrm{H}\big(\tilde{f}_i, C_{y_i}\big) + \frac{\tau}{2}\Big)\Big\} \tag{15}$$

Moreover, we simplify $\frac{\tau}{2}$ to be the constant 1, since changing it to any other positive value results only in the matrices being multiplied by a corresponding factor. Our hinge-loss like function is given as follows.

$$\mathcal{L}\big(\mathrm{W}, \{x_i\}_{i=1}^K, \{\tilde{x}_i\}_{i=1}^K; f\big) = \frac{1}{K}\sum_{i=1}^K \{max(0, 1 - \mathrm{H}(f_i, C_{y_i})) + max\big(0, \mathrm{H}\big(\tilde{f}_i, C_{y_i}\big) + 1\big)\} \tag{16}$$

By doing this, the adaptive threshold can be seamlessly factorized into a linear-fully connected layer for end-to-end learning [34]. The RML loss can also be easily used as a drop-in replacement for the triplet loss and its variants, as well as used in tandem with other performance-boosting approaches and modules, including modified network architectures, pooling functions, data augmentations or activation functions.

For a training batch consisting of K query samples, the number of input passes required to evaluate the necessary embedding feature vectors in our application is K, and the total number of distance comparisons can be $2K$. Normally, K is much larger than 2. In contrast, triplet loss and $(N+1)$-tuplet loss require $O(K^3)$ comparisons, the contrast loss and CCL require $O(K^2)$ comparisons, and the $(N+M)$-tuplet cluster loss requires $2(N+M)^*K$ comparisons after a strict example mining scheme using the special structure of some FER datasets (i.e., each subject has all 6 expressions). Here N and M are the number of mined positive samples and the number of mined negative samples, respectively. Even for a dataset of a moderate size, it is computationally impractical to load all possible meaningful triplets into the processor memory for model training. With predefined anchors (i.e., C_{y_i} and $\tilde{f}_i$), we also alleviate the difficulty of anchor selection [6].

.....The inception convolutional FER network and two-branch FC layer joint metric learning architecture proposed in our preliminary paper [6] are used in our framework in Fig. 4. The convolutional groups are made up of a 1×1, 3×3 and 5×5 Conv layers in parallel.

Combining the metric learning loss and softmax loss is an intuitive idea to possibly achieve better performance [35]. However,

Algorithm 2 Disentangled feature learning algorithm.

Input
Randomly chose K query examples$\{x_i\}_{i=1}^{K}$
and their generated references $\{\widetilde{x_i}\}_{i=1}^{K}$
Output: The parameters of the FER network θ_{FER}
1. while not converge do
2. map examples to feature plane with CNN to get: $\{f_i\}_{i=1}^{K}$ *and* $\{\tilde{f}_i\}_{i=1}^{K}$
3. calculate the cluster centers C_i for each class
4. $\mathcal{L}_{\mathrm{RML}} \leftarrow \frac{1}{K}\sum_{i=1}^{K}\{\max(0, -\mathrm{H}(f_i, C_{y_i}) + 1) + \max(0, \mathrm{H}(\tilde{f}_i, C_{y_i}) + 1)\}$
5. $\mathcal{L}_{\mathrm{softmax}} \leftarrow -\log(e^{f_{y_i} - \max(f_j)} / \sum_{j} e^{f_j - \max(f_j)})$
6. Compute the joint loss $\mathcal{L}_{\mathrm{softmax}} + \alpha\ \mathcal{L}_{\mathrm{RML}}$
7. Compute the backpropagation error
8. Update the parameters
End while

combining them directly on the last FC layer is sub-optimal. The basic idea of building two-branch FC layers after the deep convolution groups is to combine two losses at different levels of tasks. We learn the detailed features shared between the same expression class with the expression classification (EC) branch, while exploiting semantic representations via the metric learning (ML) branch to handle the significant appearance changes from different subjects. The connecting layer embeds the information learned from the expression label-based detail task to the identity label-based semantic task, and balances the weights in the two task streams. This type of combination can effectively alleviate the interference of identity-specific attributes. The inputs of connecting layer are the output vectors of the former FC layers- $FC_{2\text{-}2}$ and $FC_{2\text{-}3}$, which have the same dimension denoted as D_{input}. The output of the connecting layer, denoted as FC_4 with dimension D_{output}, is the feature vector fed into the second layer of the ML branch. The connecting layer concatenates two input feature vectors into a larger vector and maps it into a D_{output} dimension space:

$$FC_{2-4} = \mathbf{P}^{\mathrm{t}}[\mathrm{FC}_{2-2}; \mathrm{FC}_{2-3}] = \mathbf{P}_1^{\mathrm{t}}\mathrm{FC}_{2-2} + \mathbf{P}_2^{\mathrm{t}}FC_{2-3} \tag{17}$$

where $\mathbf{P}$ is a $(2D_{\mathrm{input}} \times 2D_{\mathrm{output}})$ matrix, $\mathbf{P_1}$ and $\mathbf{P_2}$ are $D_{\mathrm{input}} \times D_{\mathrm{output}}$ matrices.

Regarding the sampling strategy, every training image is used as a query example in an epoch. In practice, the softmax loss will only be calculated for the query examples. The relative importance of the two loss functions is managed by a weight α. During the testing stage, this framework takes one query image and its generated reference image as input, and determines the classification result through the EC branch with the softmax loss function. Our disentangled feature learning scheme is described in Algorithm 2.

5. Numerical experiments

In this section, we compare the IDFERM with state-of-art methods on three benchmark datasets, i.e., CK+, MMI and Oulu-CASIA datasets. Details of our training data are provided in Section 5.1, followed by our preprocessing methods in Section 5.2, and the implementation details in Section 5.3. In Section 5.4, we report a series of ablation experiments to analyze the function of our auxiliary networks. Numerical experiment results are shown in Section 5.5.

5.1. Training data

Besides the FER datasets, a variety of large datasets of facial images for face recognition are publicly available online. We give some samples from the VGG-Face dataset Fig. 7. We use the VGG-Face dataset [29] to extend our data for neutral face generation. It contains approximately 2.6 million face images, but very few of

Fig. 7. Samples from the VGG-Face dataset. Each row contains the face images of the same person.

Fig. 8. Samples from the CMU Multi-PIE dataset. Each row contains images of the same person.

these fit our requirements of neutral expression, front-facing, having no occlusion, and of 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 or tilt or pan angles beyond 5°. These frontal and neutral face images are used as our target and guiding set samples. Their corresponding non-compliant images from the same subject are used as the inputs. All the samples are aligned and cropped to 64×64 Gy images. After filtering, we have about 12 K target images ($< 0.5\%$ of the original set) and 50 K input-target pairs. These data are used for pretraining to initialize the network parameters and then fine-tuned using the CMU Multi-PIE [36].

The CMU Multi-PIE itself is a facial expression dataset, but the facial expression labels it uses are slightly different from modern expression classification system. There are 4 expressions are useable (114,305 neutral images, 19,817 surprise images, 22,696 disgust images and 47,388 happy images), while the squint and scream are not regarded as expression now. Some samples are shown in Fig. 8. It contains more than 750,000 images of 337 people taken from fifteen directions, and in nineteen illumination conditions. There are four recording sessions in which subjects were instructed to display neutral, happy, disgust and surprise facial expressions. It is more close to our FER dataset in testing stage than the filtered VGG face dataset. We selected only the five groups of the nearly frontal view faces ($-45°$ to $+45°$). The neutral images from the 0° view are used as our target image and the guiding set.

5.2. Preprocessing

We follow the [6] 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. An augmentation procedure is employed to increase the number of training images and alleviate the chance of over-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 have zero mean and unit variance. In the testing phase, a single center crop with the size of 60×60 is used as input data.

5.3. Implementation details

We use 64×64 Gy 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 guiding set using the filtered VGG-FaceNet frontal neutral view and the 0° view neutral images from the Multi-PIE. Following the experimental protocol in [6], we pre-train our inception style convolutional groups, two branch FC layers on with 204,156 frontal view (−45°to 45°) face images selected from the CMU Multi-PIE dataset for 300 epochs, optimizing the joint loss using stochastic gradient descent with a momentum coefficient of 0.9. The initial network learning rate, batch size, and weight decay parameter are set to 0.1, 128, 0.0001, respectively based on optimizing the parameter choices using the validation set. If the training loss increased by more than 25% or the validation accuracy does not improve for ten epochs, the learning rate is halved and the previous network with the best loss is reloaded. We select the highest accuracy training epoch as our pre-trained model. In the fine-tuning stage, the mini-batch set size is fixed to two times the number of expression classes of the dataset. Random search is employed to select 2 images from each expression class to form the mini-batch set. The tuplet-size is set to 12. In all our experiments, we set $\eta = 0.1$, $\lambda_1 = 3 \times 10^{-2}$, $\lambda_2 = 10^{-2}$, $\lambda_3 = 0.3$ determined by manual tuning. The weight of joint learning $\alpha = 1$. In the testing phase, only the convolutional groups and expression classification branch with softmax are used to recognize a single facial expression image.

The details of the encoder of HNG can be found in [7]. 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 $z \in \mathbb{R}^{256}$ into a synthetic image $\tilde{x} \in \mathbb{R}^{64 \times 64}$, which is of the same size as x. 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 Leaky ReLU nonlinearities [37] are used in some Conv layers, where $\mathrm{LReLU}(x) = max(x, 0) + \sigma min(x, 0)$. In our experiments, we set $\sigma = 0.1$. Optimizing this minimax 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. All the CNN architectures are implemented with the widely used deep learning tool "Tensorflow [38]."

5.4. Ablation study

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 that 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 Fig. 9, 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. Fig. 10 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. [18] 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. We investigated the effect of the weight of identity preserving loss and show the identity inconsistent percentage in Table 1.

The VGG-FaceNet is 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, without this part, we could never avoid the collapse of the adversarial training to generate the human face structure.

We also analyzed the effect of different hyper-parameter values in RML. The parameter α is used to balance the softmax loss and metric learning loss. We can see from the Fig. 11 (reproduced below) that the highest accuracy is achieved when $\alpha \in [0.95,1]$. As can be seen in Fig. 12 below, the networks were not sensitive to $\sigma \in [0.075, 0.125]$.

5.5. Experimental results of FER

To evaluate the effectiveness of the proposed method, extensive experiments have been conducted on four well-known publicly available facial expressions datasets: CK+, MMI and Oulu-CASIA. Resulting IDFERM confusion matrices are shown in Fig. 13.

CK+Dataset [39]: We conducted both seven-class and eight-class expression recognition experiments (*i.e.*, without or with neutral expression). In the setting without neutral sample, we directly compare to the most nontrivial hard negative samples (i.e., generated normalized face), which not only relaxes the requirements on the dataset (*i.e.*, needing images of all different expressions of the same person) to extract the identity-disentangled expression representation but also reduces the number of comparisons in training stage. The training time of metric learning part is largely reduced as shown in Table 2. We note that [7] and IDFERM do need an additional training stage (around 6 hours) for normalized face generation, but the trained HNG network works for all of the FER datasets in our experiments without fine-tuning and can be regarded as a ready-made tool for several down-stream tasks.

In the testing stage, IDFERM recognizes a query facial expression image in about 50 ms, which is satisfactory for many applications. Video-based methods normally need a relatively longer sampling time (>0.25 s) to collect the whole expression change session.

From Table 3, we can see that the identity-disentangled representation with adaptive metric learning methods achieve higher accuracy than previous works. Therefore, comparing the query sample with its generated normalized face rather than all its other expressive query faces as in [6] is more efficient, which is consistent with the relationship of those expressions as analyzed in Fig. 2. However, just adding the generated neutral face images to original dataset enlarged the number of comparisons [7]. As an efficient hard negative mining scheme, the HNG offers the most nontrivial hard negative samples and the RML can efficiently utilizes them and outperforms the other methods.

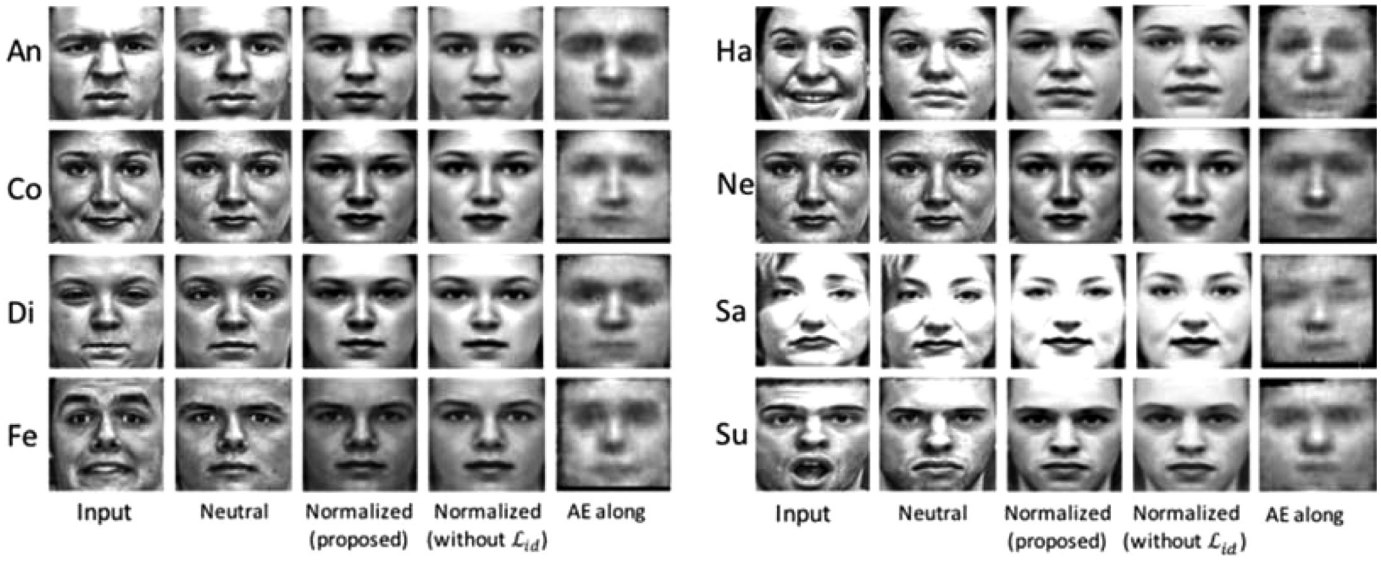

Fig. 9. Examples of input, target, generated normalized face by RG network, generated normalized face by RG network without identity-preserving loss and reconstructed neutral face using only the auto-encoder (AE) structure. Real images are from CK+ dataset which has an additional calss called contempt (Co) class.

Table 1
Percentage of identity errors as a function of the λ_1, the weight of identity-preserving loss term.

λ_1	0	0.001	0.005	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.05
Percentage (%)	2.13	1.77	1.06	0.32	0.13	0.11	0.1	0.1	0.1	0.1	0.1

Table 2
Comparation of the metric learning training time on the datasets with a Titan X GPU.

Metric Learning Training Time	CK+(7-class)	CK+(8-class)	MMI	Oulu-CASIA VIS
Triplet Loss	5 h 24mins	–	3 h 14mins	–
2B(N+M) Softmax [6]	1 h 47mins	–	54 mins	–
Data Augmentation [7]	3 h 11mins	–	2 h 8mins	–
IDFERM	42 mins	1 h 4 mins	30 mins	56 mins

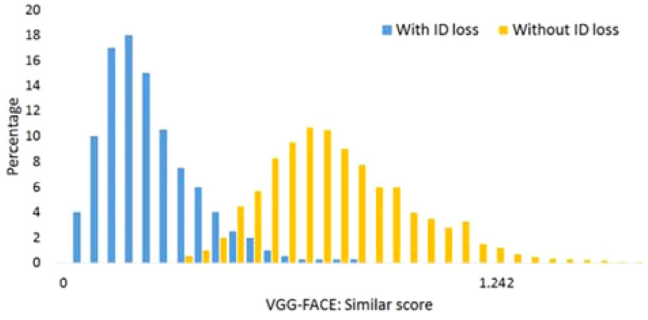

Fig. 10. Histogram of VGG-Face net L2 error between the input face and the normalized pairs on the FER data collection. Blue: with the identity preserving loss which calculated by the Light CNN. Orange: without the identity preserving loss. The 1.242 threshold was used by Schroff et al. [18] to cluster 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.

Table 3
Performance compares of the rank-1 recognition accuracy on the CK+ dataset in terms of 7 expressions and the MMI dataset (without neutral expression).

Methods	CK+(seven-class)	MMI
MSR [13]	91.4%	N/A
BNBN [42]	96.7%	N/A
IBCNN [4]	95.1%	N/A
STM [43]	96.3%	N/A
CER(video) [44]	92.34%	70.12%
CDMML(video) [45]	96.6%	N/A
STMExplet(video) [9]	94.19%	75.12%
DTAGN(video) [10]	97.25%	70.2%
IACNN [11]	95.37%	71.55%
2B(N+M) Softmax [6]	97.1%	78.53%
Data Augmentation [7]	97.49%	80.26%
IDFERM	98.35%	81.13%

Table 4
Performance comparison of the rank-1 recognition accuracy on the CK+ dataset in terms of the 8 expressions (with neutral expression).

Methods	CK+(eight-class)
AUDB [46]	93.70%
CNN+AD [47]	96.4%
FN2EN [5]	96.8%
IDFERM	97.76%

The improved accuracy compared to the other methods is appealing in the image-based 7-class CK+, MMI and Oulu-CASIA setting which do not have real-neutral samples as training data. It also generalized well in dataset with neutral expressions as shown in Table 4. With the added generated samples, the accuracy of neutral class is improved to 99% as shown in Fig. 13(b).

Benefitting from the generated data, the proposed IDFERM outperforms our earlier approach [6] by 2.6% on MMI dataset

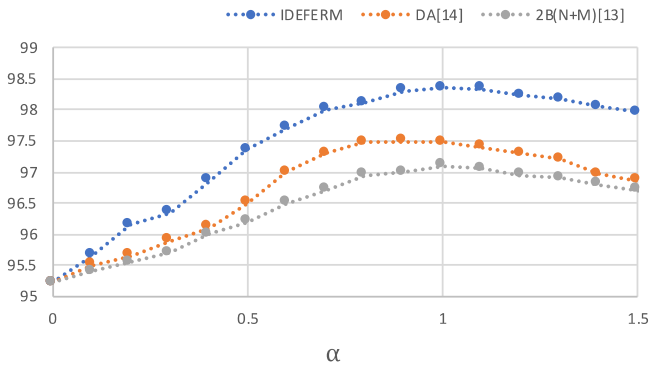

Fig. 11. Facial recognition accuracy on CK+(7-class) dataset as a function of α, the parameter used to balance the softmax loss and metric learning loss.

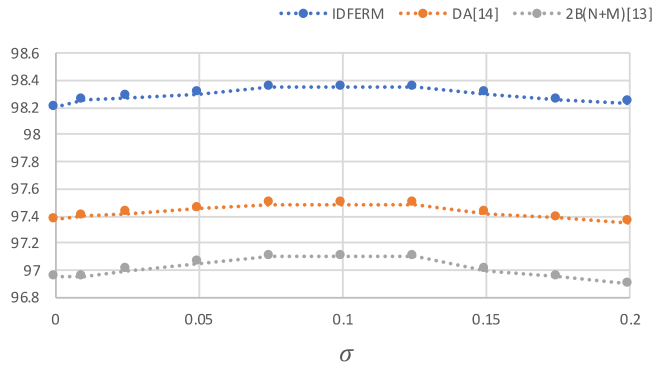

Fig. 12. Facial recognition accuracy on CK+(7-class) dataset as a function of parameter σ in LReLU.

and 1.25% on CK+ dataset. Using the same generated images as in [7], we achieve 0.87% and 0.86% improvements on MMI and CK+ datasets respectively, and the number of comparisons per training batch is reduced from $2(N+M)K$ (the N and M in MMI are 5 and 5 respectively [7]) to 2 K. As a consequence, the training time of the earlier approach [7] for MMI dataset is 2hours 8mins, while the IDFERM needs only 30mins on a single Titan X GPU as shown in Table 2. Considering the improved performance and reduced training time, the IDFERM is significantly more efficient than [7] to utilize the generated data.

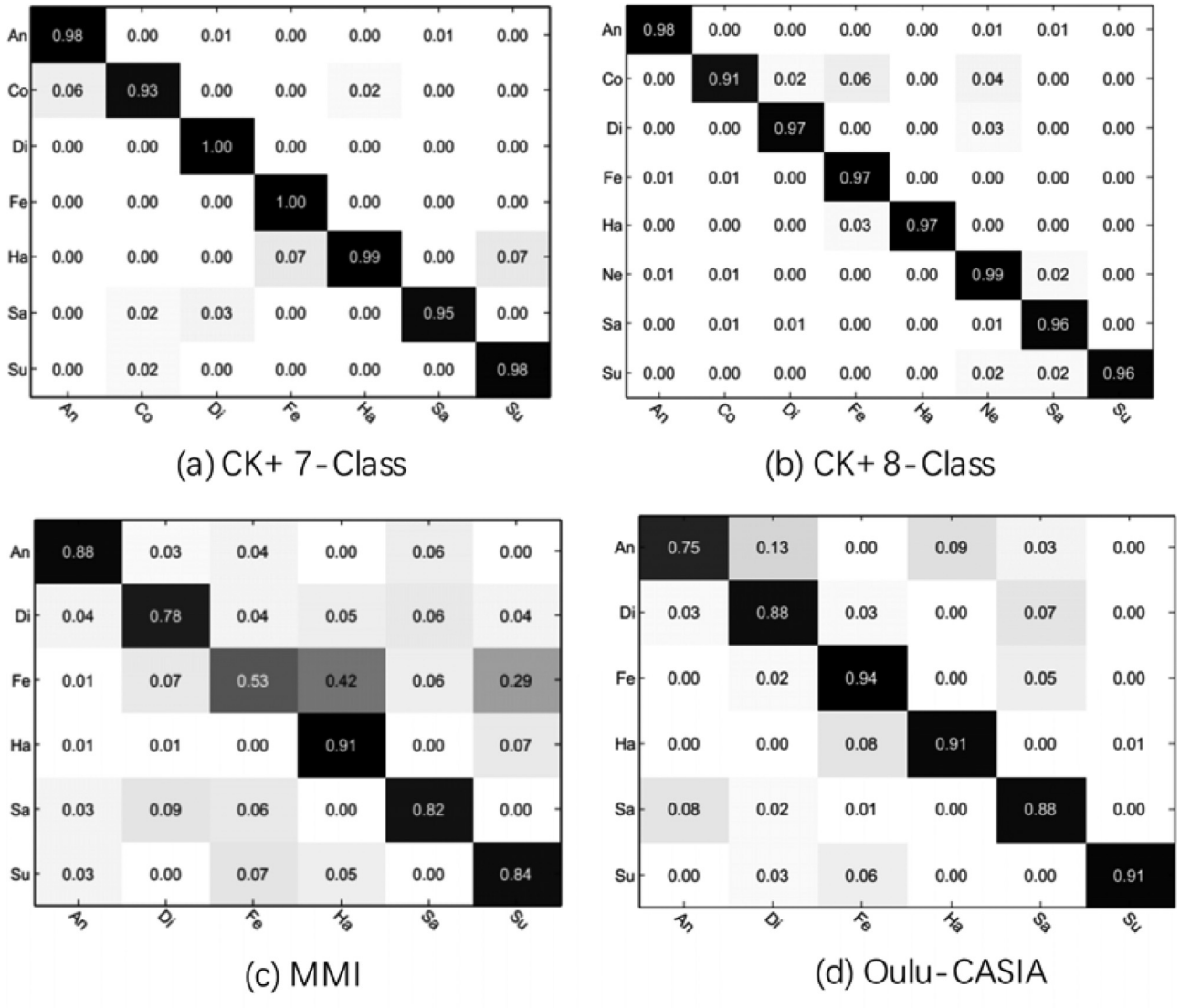

Fig. 13. Confusion matrix of the proposed IDFERM evaluated in the (a) CK+ seven-class, (b) CK+ eight-class, (c) MMI and (d) Oulu-CASIA database. The predicted labels and the ground truth labels are shown in the ordinate and abscissa, respectively.

Table 5
Comparation of the performance with/without pre-training using Multi-PIE on CK + dataset.

	2B(N + M) [6]		IDFERM	
	Without Pre-training	With Pre-training	Without Pre-training	With Pre-training
CK + (7-class)	97.03%	97.10%	98.32%	98.35%
MMI	78.46%	78.53%	81.11%	81.13%

Table 6
Performance comparison of the rank-1 recognition accuracy on the Oulu-CASIA VIS dataset in terms of the 6 expressions (without neutral expression).

Methods	Oulu-CASIA VIS
STM-ExpLet(video) [9]	74.59%
DTAGN(video) [10]	81.46%
PPDN [48]	84.59%
FN2EN [5]	87.1%
IDFERM	88.25%

Limited training data has long been a challenge for facial expression recognition. For example, [11] utilized the FER-2013 dataset for pre-training and then fine-tuned their facial expression recognition network in CK + /MMI datasets. The FER-2013 dataset is even larger than the Multi-PIE dataset. We chose the Multi-PIE pretraining for fair comparison with previous works [6]. We added a comparison of recognition accuracy with/without the pre-training, and show the results in Table 5. We can see that the pre-training can improve the performance consistently.

MMI Dataset [40]: This dataset consists of 213 sequences, 208 sequences from this data set containing frontal-view faces of 31 subjects were used in our experiment as in [6]. Since the actual location of the peak frame is not provided, we collect three frames in the middle of each image sequence and associate them with the labels, which results in 624 images in our experiments as in [6,11]. We divided the MMI dataset into 10 subsets for person-independent ten-fold cross validation. The sequence-level predictions are obtained by choosing the class with the highest average score of the three images. Consequently, 10-fold cross validation was conducted. This dataset could be suitable to measure the recognition performance in realistic situations when compared to other datasets.

With the identity-disentangled FER representation, the proposed methods achieve substantial improvements over the previous best performance in MMI dataset as shown in Table 3 and Fig. 13(c). The HNG can further boost the accuracy by incorporating the prior information of normalized face and the relationships of expressions within an applicable framework. Note that the image sequences in the MMI dataset contain a full temporal pattern of expressions, *i.e.*, from neutral to apex, and then relaxed, and are especially favored by these methods exploiting temporal information.

Oulu-CASIA VIS Dataset [41]: This dataset consists of 480 image sequences of 80 individuals. This dataset is captured under the visible (VIS) normal illumination conditions and is a subset of Oulu-CASIA NIR-VIS dataset. Each individual poses six basic expressions as in MMI dataset. Only the last three frames are used for individual-independent 10-fold cross validation, and the total number of images is 1440 as in [5].

In Oulu-CASIA dataset, the IDFERM performs well in recognizing fear and happy expressions, while angry is the hardest expression, which is mostly confused with disgust as shown in Fig. 13(d). The performance results are shown in Table 6 and are similar to those on the CK + and MMI datasets.

6. Conclusions

We proposed and investigated a novel recognition via generation scheme termed IDFERM to disentangle the identity factors from other factors that are responsible for facial expression. The anchor-selection and threshold-tuning problems present in previous approaches have been addressed in our proposed adaptive deep metric learning paradigm. The identity-preserving neutral face image generation is efficient for hard negative mining which requires fewer similarity comparisons. However, our gray-scale image processing can lead to information loss as the image quality is not emphasized in our framework as in conventional image generation methods. Also, the adversarial game at image-level is usually time-consuming. In future work, we intend to apply some commonly used visual quality assessment methods for the generated images on top of our model for better texture. Recent feature -level GAN's backbone can be utilized to extend our framework for faster, more stable convergence training, and more complex data structure (e.g., color images). We also expect that the application of recognition via generation idea can facilitate several other closely related tasks, e.g., face recognition, person re-ID, and pose-invariant classification.

References

[1] J.F. Cohn, P. Ekman, "Measuring facial action, in: The New Handbook of Methods in Nonverbal Behavior Research, 2005, pp. 9–64.
[2] P. Ekman, L. Erika, What the Face reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.
[3] J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, IEEE Trans. Image Process. 27 (5) (2018) 2420–2432.
[4] S. Han, Z. Meng, A. Khan, Y. Tong, Incremental boosting convolutional neural network for facial action unit recognition, in: Advances in Neural Information Processing Systems, 2016, pp. 109–117.
[5] H. Ding, S. Zhou, R. Chellappa, Facenet2expnet: regularizing a deep face recognition net for expression recognition, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 118–126.
[6] X. Liu, B.V.K. Vijaya Kumar, J. You, P. Jia, Adaptive deep metric learning for identity-aware facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 522–531.
[7] X. Liu, B.V.K. Vijaya Kumar, Y. Ge, C. Yang, J. You, P. Jia, Normalized face generation via perceptral generative adversarial networks, in: Identity, Security, and Behavior Analysis (ISBA), 2018 IEEE 4th International Conference on, IEEE, 2018, pp. 1–8.
[8] J. Haxby, E. Hoffman, M. Gobbini, "The distributed human neural system for face perception, Trends Cogn. Sci. 4 (6) (2000) 223–233.
[9] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1749–1756.
[10] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991.
[11] Z. Meng, P. Liu, J. Cai, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 558–565.
[12] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: a survey of registration, representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 1113–1133.
[13] S. Rifai, Y. Bengio, A. Courville, Disentangling factors of variation for facial expression recognition, in: Computer Vision–ECCV 2012, Berlin, Heidelberg, Springer, 2012, pp. 808–822.
[14] P. Liu, J. Zhou, I. Tsang, Z. Meng, S. Han, Y. Tong, Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis, in: European Conference on Computer Vision, Cham, Springer, 2014, pp. 151–166.

[15] H. Liu, Y. Tian, Y. Yang, L. Pang, T. Huang, Deep relative distance learning: tell the difference between similar vehicles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2167–2175.
[16] W. Deng, J. Hu, Z. Wu, J. Guo, From one to many: pose-aware metric learning for single-sample face recognition, Pattern Recognit. 77 (2018) 426–437.
[17] J. Wang, Z. Wang, C. Liang, C. Gao, N. Sang, Equidistance constrained metric learning for person re-identification, Pattern Recognit. 74 (2018) 38–51.
[18] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1, IEEE, 2005, pp. 539–546.
[19] S. Ding, L. lin, G Wang, H. Chao, Deep feature learning with relative distance comparison for Person *Re* identification, Pattern Recognit. 48 (2015) 2993–3003.
[20] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Advances in Neural Information Processing Systems, 2016, pp. 1857–1865.
[21] Y. Dong, B Du, L. Zhang, L. Zhang, D. Tao, LAM3L: Locally adaptive maximum margin metric learning for visual data classification, Neurocomputing 235 (2017) 1–9.
[22] S. Thrun, others, Robotic mapping: A survey, Exploring Artif. Intell. New Millennium 1 (1-35) (2002) 1.
[23] ChristopherM. Bishop, Pattern recognition and machine learning (information science and statistics) springer-verlag new york, Inc. Secaucus, 2006.
[24] Sebastian Thrun, Robotic mapping: A survey, *Explor. Artif. Intell. New Millenn.* 1 (2002) 1–35.
[25] P. Vincent, H. Larochelle, Y. Bengio, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 1096–1103.
[26] I. Goodfellow, J. Pouget, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[27] I. Goodfellow, Ian. "NIPS 2016 tutorial: Generative adversarial networks." In arXiv: 1701.00160, 2016.
[28] L. Tran, X. Yin, X. Liu, Disentangled representation learning gan for pose-invariant face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3, 2017, p. 7.
[29] P. Omkar, A. Vedaldi, A. Zisserman, Deep Face Recognition, BMVC 1 (3) (2015) 6.
[30] X. Wu, R. He, Z. Sun, T. Tan, A Light CNN for Deep Face Representation with Noisy Labels, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2884–2896.
[31] L. Theis, A. van den Oord, and M. Bethge. "A note on the evaluation of generative models." arXiv preprint arXiv:1511.01844 (2015).
[32] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, Cham, Springer, 2016, pp. 499–515.
[33] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 279–283.
[34] H. Shi, Y. Yang, X. Zhu, L. Zhen, W. Zheng, S.Z. Li., Embedding deep metric for person re-identification: a study against large variations, in: European Conference on Computer Vision, Cham, Springer, 2016, pp. 732–748.
[35] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.
[36] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vision Comput. 28 (5) (2010) 807–813.
[37] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.
[38] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: a system for large-scale machine learning, OSDI 16 (2016) 265–283.
[39] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 94–101.
[40] M. Pantic, M Valstar, R Rademaker, L Maat, Web-based database for facial expression analysis, in: 2005 IEEE international conference on multimedia and Expo, IEEE, 2005, p. 5.
[41] G. Zhao, X. Huang, M. Taini, S.Z. Li, M. Pietikainen, Facial expression recognition from near-infrared videos, Image Vision Comput. 29 (9) (2011) 607–619.
[42] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805–1812.
[43] W. Chu, F. Torre, J.F. Cohn, Selective transfer machine for personalized facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 39 (3) (2017) 529–545.
[44] S. Lee, W. Baddar, Y. Ro., Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos, Pattern Recognit. 54 (2016) 52–67.
[45] H. Yan, Collaborative discriminative multi-metric learning for facial expression recognition in video, Pattern Recognit. 75 (2018) 33–40.
[46] M. Liu, S. Li, S. Shan, X. Chen, Au-inspired deep net- works for facial expression feature learning, Neurocomputing 159 (2015) 126–136.
[47] P. Khorrami, T. Paine, T. Huang, Do deep neural networks learn facial action units when doing expression recognition? in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 19–27.
[48] X. Zhao, X. Liang, L. Liu, T. Li, N. Vasconcelos, S. Yan, Peakpiloted deep network for facial expression recognition, in: European conference on computer vision, Cham, Springer, 2016, pp. 425–442.

Xiaofeng Liu received the B.Eng. degree in automation and B.A. degree in communication from the University of Science and Technology of China, Hefei, China, in 2014. He was the research assistant in MSRA, CMU and NTU. He is currently pursuing his Ph.D. at the University of Chinese Academy of Sciences, Beijing and a Research Associate in the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh. He was a recipient of the Best Paper award of the IEEE International Conference on Identity, Security and Behavior Analysis 2018. His research interests include image processing, computer vision, and pattern recognition.

B.V.K. Vijaya Kumar is the U.A. & Helen Whitaker Professor of Electrical and Computer Engineering at Carnegie Mellon University, Pittsburgh. He is also the Director of Carnegie Mellon University Africa in Rwanda. His research interests include computer vision, pattern recognition algorithms, and applications, coding, and signal processing for data storage systems. His publications include a book entitled Correlation Pattern Recognition, 22 book chapters, more than 400 conference papers, and more than 200 journal papers. He is also a co-inventor of 12 patents. He served as a topical editor of *Applied Optics* and as an associate editor of the IEEE *Transactions on Information Forensics and Security*. He has served on many conference program committees and was co-chair of several conference program committees. In 2003, he received the Eta Kappa Nu Award for Excellence in Teaching in the ECE Department, CMU and the Carnegie Institute of Technology's Dowd Fellowship for educational contributions and he was a co-recipient of the 2008 Outstanding Faculty Research Award in CMU's College of Engineering. He is a fellow of IEEE, SPIE, OSA, AAAS, IAPR and NAI.

Ping Jia received his B.Eng and MSc degree in computer science from the University of Science and Technology of China, Hefei and his Ph.D. from the Graduate University of the Chinese Academy of Sciences. He is the President of Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His current research interests include image processing, computer vision, and optical engineering.

Jane You received the Ph.D. degree from La Trobe University, Melbourne, VIC, Australia, in 1992. She is currently a Professor with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, and the Chair of Department Research Committee. She has researched extensively in the fields of image processing, medical imaging, computer-aided diagnosis, and pattern recognition. She has been a Principal Investigator for one ITF project, three GRF projects, and many other joint grants since she joined PolyU in 1998. Prof. You was a recipient of three awards including Hong Kong Government Industrial Awards, the Special Prize and Gold Medal with Jury's Commendation at the 39th International Exhibition of Inventions of Geneva in 2011 for her current work on retinal imaging, and the Second Place in an International Competition [SPIE Medical Imaging'2009 Retinopathy Online Challenge in (ROC'2009)]. Her research output on retinal imaging has been successfully led to technology transfer with clinical applications. She is an Associate Editor of Pattern Recognition and other journals.