Multi-factor joint normalisation for face recognition in the wild

Yanfei Liu | Junhua Chen

Abstract
Face recognition has become very challenging in unconstrained conditions due to strong intra-personal variations, such as large pose changes. Face normalisation can help to resolve these problems and effectively improve the face recognition performance in unconstrained conditions by converting non-frontal faces to frontal ones. However, there are other complex facial variations in addition to pose, such as illumination and expression, which will also influence face recognition performance. The authors propose a well-designed generative adversarial network-based multi-factor joint normalisation network (MFJNN) to normalise multiple factors simultaneously. First, a multi-encoder generator and a feature fusion strategy are designed and implemented in the MFJNN to realise the joint normalisation of multiple factors in addition to pose. Second, a convolutional neural network-based (CNN-based) network is applied in the MFJNN, which allows the MFJNN to simultaneously realise image synthesis and facial representation learning. Moreover, an identity perceptive loss is introduced based on the CNN-based network to produce reliable identity-preserving features of the input face images. The experimental results demonstrate that the proposed method can synthesise multi-factor normalisation results with identity preservation and effectively improve the face recognition performance.

1 | INTRODUCTION

Face recognition is a typical biometric identification technology which has extensive application prospects in the fields of public security of state, military security, financial security, human–computer interaction etc. Currently, face recognition has achieved outstanding results due to the development of deep learning technology and the easy access to large-scale labelled datasets. However, in an unconstrained environment with extreme pose, illumination, expression, variations etc., it is still very challenging to achieve robust face recognition. Face normalisation, which aims to synthesise a photorealistic frontal view face image from one with arbitrary pose, can effectively address the problem of face recognition in unconstrained conditions.

Current face normalisation methods can be divided into three categories: 2D-based methods, 3D-based methods and deep learning-based methods. 2D-based methods realise pose normalisation using 2D image matching or test image encoding using a basis function, and 3D-based methods normally capture 3D face data or evaluate 3D models from 2D images and match them to the 2D test images. Compared with the first two methods, deep learning which has achieved outstanding normalisation performance [1–4] due to its highly nonlinear mapping capability and it is a new way to realise face normalisation. However, the above face normalisation methods have certain limitations. For example, 3D data capturing requires additional computation complexity and resources. Moreover, 3D-based approaches always lead to distortion and inauthenticity in the synthesised image. Although deep learning related methods, especially algorithms based on generative adversarial network (GAN) [5], have achieved impressive results in face normalisation, face normalisation in unconstrained conditions still faces a great challenge because there are many other complex factors besides pose that need to be normalised.

In the real world, illumination, pose, expression and other factors are always intertwined with each other, as shown in Figure 1. Hence, it is difficult to achieve real practicality by handling only one of these factors (such as pose).

Therefore, the authors aim to achieve the joint normalisation of various factors that affect the face recognition performance. So, they propose a multi-factor joint normalisation
network (MFJNN), which is designed based on the GAN, for face normalisation. A multi-encoder generator and a fusion strategy are designed and implemented in the MFJNN to achieve multi-factor normalisation (MFN). Each encoder inputs an image with some abnormal factor in the training stage and is mainly responsible for this factor’s normalisation. For example, the expression encoder takes abnormal expression face images as its input and mainly normalises the expressions. Using a weighted averaging feature fusion strategy, the features extracted from each encoder are fused as one feature representation and fed to the decoder to generate a synthesised virtual canonical face image.

To produce identity-preserving synthesised face image for face recognition, the authors introduce a multi-input based identity perceptive loss in the proposed model. Moreover, we apply the source of the identity perceptive loss (a convolutional neural network-based (CNN-based) network) as a feature learning subnetwork, which allows the proposed MFJNN to simultaneously realise face normalisation and facial representation learning.

The work of authors are summarised as follows.

1) The author propose a MFJNN based on the GAN is proposed to synthesise frontal face images and learn pose-invariant representations for face recognition from face images under arbitrary condition. Different from current GAN-based models that usually use one encoder and one decoder for the generator to synthesise images, the proposed GAN uses multiple encoders to input face images with different variations (e.g. pose and illumination) and fuse the outputs of these encoders to one decoder to achieve multi-factor joint normalisation.

2) Further, a multi-input-based identity perceptive loss is introduced to the proposed model to preserve the identity information of synthesised face images. Moreover, the CNN-based network used for computing the identity perceptive loss function is also used by us for feature learning, and this allows the proposed model to realise face normalisation and facial feature learning in one framework.

2 | RELATED WORK

Face recognition methods can be divided into two categories: hand-crafted methods and deep learning based methods. Early hand-crafted methods directly extract hand-crafted features for face recognition. Typical methods include local binary patterns [6], histograms of oriented gradients [7] etc. In recent years, deep learning based face recognition methods have achieved promising results [8–10]. The most popular deep learning network for face recognition is the Deep convolutional neural network (DCNN). However, on datasets such as IJB-A [11], which has larger pose variation than traditional LFW [12] dataset, the face recognition performance is still not up to the mark. Pose variations can be considered as the most important and challenging problem in unconstrained face recognition.

Face normalisation is an efficient way to overcome pose problem. Early face normalisation algorithms can be divided
into two categories: 3D-based and 2D-based normalisation methods. In the first category, the morphable model fits a 3D model to an input face using the prior knowledge of human faces and image-based reconstruction. Li et al. generated one virtual pose for each probe face image using a series of 3D displacement fields sampled from 3D face database, and matched this synthetic face to the gallery face [13]. Similarly, Asthana et al. used a pose-based active appearance model (AAM) to match a 3D model and 2D image [14]. Compared with 3D-based approaches, 2D techniques do not need 3D prior information for pose normalisation. For example, a dynamic programming stereo matching method was adopted to calculate the similarity of two faces for 2D face recognition in [15]. The tested face image was represented by linear combination of training images and linear regression coefficient was used as the feature for 2D face recognition in [16].

Recently, the deep neural network has been used in face normalisation and has achieved significant performance. In [2] the authors used a DCNN to normalise the illumination and pose variations of face image, and extracted the conventional features on the canonical face image after normalisation for face recognition. Experimental results on CMU Multi-PIE [17] dataset have shown that this method can achieve a high recognition rate compared with the method that directly extracts features. [1,3] recovered face images with noises and occlusions by utilising the restricted boltzmann machine (RBM) and gated markov random field, respectively, which have achieved good performance on face verification and face expression recognition, respectively. [4] applied the stacked denoising auto-encoder (SDA) to recover occluded face image, and then adopted the DCNN for face recognition. Experiments on the occluded face images from the AR [18] dataset demonstrated that the face image restoration method based on the SDA had a comparable recognition rate similar to the method based on the sparse representation.

It is worth mentioning that the GAN has attracted substantial attention in image synthesis due to its ability to synthesise photorealistic images with plausable high frequency details. The state-of-the-art face normalisation methods are almost based on the GAN [19–24]. For instance, the Two Pathway GAN (TP-GAN) [19] proposed a two-pathway GAN architecture to simultaneously perceive global and local information. The face frontalization GAN (FF-GAN) [20] incorporated a 3D face model into the GAN to maintain the visual quality under occlusion during frontal view synthesis. Zhao et al. [22] further extended the TP-GAN and proposed the PIM (pose invariant model) that introduces a domain adaptation strategy for face normalisation. A couple-agent pose-guided GAN [21] was proposed to synthesise both frontal and profile face images. Bao et al. [23] proposed a GAN-based framework for synthesising face images recombining different identities and attributes in open domains. [25] proposed a face normalisation model to generate frontal, neutral expression, photorealistic face images. While most of the prior works learn a representation that is mainly invariant to the pose, our method can learn a representation invariant to other variations like illumination and expression besides pose. From a face image under an arbitrary condition, the MFJNN can synthesise a virtual frontal face image and learn a multi-factor invariant representation while preserving face identity.

3 | APPROACH

Here, the authors propose a MFJNN to synthesise a photorealistic and identity-preserving frontal face and learn robust feature representation for face recognition from face image with multi-factor variations. The overview of the MFJNN is detailed in Section 3.1. In the MFJNN, the multi-encoder generator, which is introduced in Section 3.2, is the main structure used to achieve MFN. Finally, the loss function of our method including identity preserving loss is detailed in Section 3.3.

3.1 | Multi-factor joint normalisation network

The general architecture of the MFJNN is shown in Figure 2. To jointly normalise multiple factors (including pose, illumination, expression, etc.), we design a multi-encoder generator \( G_\theta : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H \times W \times C} \) that has one decoder but multiple encoders for different factors, where \( H \), \( W \), and \( C \) represent the height, the width and the number of channels of the input image of the generator, respectively; and \( \theta \) denotes the network parameters for the generator. In the training stage, each encoder takes face images with some factor changes as its input and maps each input face to the feature space. Then, the extracted features from the encoders are fused by a weighted averaging fusion strategy in the feature fusion block. The decoder takes the fused feature as its input and generates the synthesised normal face image. In the testing stage, the image that needs to be normalised is taken as the input of each encoder of the multi-encoder generator and will be multi-factor normalised by the multi-encoder generator.

To synthesise identity-preserving face images and simultaneously to learn facial representations for face recognition in addition to image synthesis, we apply the Light CNN [26] to the synthesised face images. The Light CNN is used as a source of the identity perspective loss to make the synthesised image identity-preserving and also as a feature learning sub-network to learn the pose-invariant representation from a synthesised face image for face recognition. There are three reasons to choose the Light CNN as our CNN architecture. First, the Light CNN has achieved good face recognition performance in many face recognition via generation works. For example, the Light CNN demonstrates better recognition performance than Visual Geometry Group Face (VGG-Face) [27] and residual network (ResNet) [28] in [22,25], respectively. Second, many works use the Light CNN to generate the identity perception loss and have demonstrated good results [19,21,29]. Finally, the Light CNN has relatively simpler network architecture than other CNNs like ResNet-50 or ResNet-100, which are the most common ResNet network architectures.
The structure of the discriminator is shown in Figure 3. The discriminator denoted as $D_\emptyset$, where $\emptyset$ represents the network parameters for the discriminator, consists of a set of convolutional layers to execute down-sampling operations iteratively. The LReLU [30] is adopted as the activation function after each convolutional layer. The final layer is a fully connected layer that transforms the input feature into a one-dimensional output feature.

### 3.2 Multi-encoder generator

To normalise multiple factors in a unified framework, we design a multi-encoder generator that consists of multiple encoders denoted as $G_{enc} : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{D}$, where $D$ denotes the dimension of the output feature of the encoder, and one decoder is denoted as $G_{dec}$. In the proposed experiment, we use three encoders corresponding to pose, illumination and expression normalisation. Therefore, the proposed MFJNN is considered as a three encoder architecture for simplicity here. The three encoders correspond to the input $I = \{I_p, I_i, I_e\} \in \mathbb{R}^{H \times W \times C}$ and aim to learn features $G_{enc}(I_p)$, $G_{enc}(I_i)$ and $G_{enc}(I_e)$ from face images $I_p$, $I_i$ and $I_e$, respectively, where $I_p$, $I_i$ and $I_e$ denote face images with pose, illumination and expression variations, respectively. With the input $\{I_p, I_i, I_e\}$, the fused representation is the weighted average of the three representations.

**Figure 2** Visual illustration of the MFJNN. The generator contains multiple encoders and one decoder for multi-factor normalisation. The features extracted from the encoders are fused as one feature through a weighted averaging fusion strategy and fed to the decoder. The discriminator (D) distinguishes between the synthesised frontal views and the ground-truth frontal views. Images with some factor variance are taken as the input of one encoder for training. During testing, the test image will be fed to the multiple encoders as the input of all encoders. Losses are drawn with dashed lines, where $L_{adv}$ is the adversarial loss that encourages the generator to synthesise the photorealistic normalised face image, $L_i$ is the identity perceptive loss for preserving identity information, $L_{sym}$ is the symmetry loss to alleviate the self-occlusion problem in large pose cases and $L_p$ is the pixel-wise loss to facilitate image content consistency.

**Figure 3** Structure of the discriminator $D_\emptyset$ and the encoder of the generator $G_{enc}$ in the MFJNN. Each rectangle corresponds to one network module, and the feature map dimension of each layer is marked by dotted lines when the input image has a size of 224×224. Note that the last fully connected layer is only included in the discriminator architecture and the encoder of the generator directly outputs features with a size of 14×14×256.
The multi-encoder generator is a fully convolutional network. The structure of the encoder is as same as the discriminator except the last fully connected layer, as shown in Figure 3. The decoder is composed of a set of modules, each of which comprises a transposed convolution layer [31], a ReLU layer and a residual block [28], as shown in Figure 4. The transposed convolution layer with a $4 \times 4$ filter and a stride of 2 executes an un-sampling operation and doubles the input feature. The residual block is composed of a convolutional layer ($4 \times 4$ filter, stride = 1), batch normalisation (BN), and a ReLU layer, cascaded by another convolutional layer ($4 \times 4$ filter, stride = 1) and BN. Finally, a $1 \times 1$ convolution is applied to generate the output image.

### 3.3 Loss functions

(1) Adversarial Loss: To generate a photorealistic normalised face image, we adopt the adversarial loss term as follows:

$$L_{adv-D} = -\log D_{\phi}(F_{g}^{\ast}) - \log(1 - D_{\phi}(\tilde{I}))$$  
$$L_{adv-G} = -\log D_{\phi}(\tilde{I})$$

where $L_{adv-D}$ and $L_{adv-G}$ represent the adversarial loss for the discriminator and the generator, respectively, and $F_{g}^{\ast}$ is the ground truth frontal face image of the synthesised image.

(2) Multi-input-based identity perceptive loss: Preserving the identity of the frontal face image is the most important part in developing the recognition via generation framework. Our identity perception loss for preserving identity information is based on the activations of the last two layers of the Light CNN. Since our model has multiple inputs, we introduce an identity perceptive loss based on multiple inputs. In a situation with three inputs, we define our identity perceptive loss as

$$L_{id} = \sum_{l=1}^{2} \frac{1}{W_l \times H_l} \sum_{m=1}^{W_l} \sum_{n=1}^{H_l} (F(l_{m,n}^{1}) - (\omega_p F(I_{m,n}^{1})$$

$$+ \omega_i F(I_{m,n}^{2}) + \omega_e F(I_{m,n}^{3}))^2$$

where $\| \cdot \|_2$ means the vector 2-norm, $W_l$ and $H_l$ respectively denote the width and the height of the $l$th layer, and $F(l_{m,n}^{i})$ is the value of point $(m, n)$ in feature map of the $l$th layer.

(3) Pixel-wise loss: For a fine-grained task like face normalisation, the details of the facial features will cause great challenges for the synthesis task. The synthesised images may contain distortions and lack authenticity without the supervision of the pixel-wise loss. Therefore, we use the pixel-wise loss to maintain the consistency of the image content. Since the $L_2$ loss tends to generate a blurry output, we adopt the $L_1$ loss as in [19,25] as our pixel-wise loss instead to better preserve high frequency signals.

$$L_p = \frac{1}{W \times H \times C} \sum_{w,h,c} |\tilde{I}_{w,h,c} - F_{g,w,h,c}^{\ast}|$$

where $F_{g,w,h,c}^{\ast}$ is the ground truth image of the synthesised image.

(4) Symmetric loss: Symmetry is a normal property of a face image that can be used to constrain the synthesised frontal face image as prior knowledge. We use the symmetric loss as in [19] formulated as follows:

\[F_{g,w,h,c}^{\ast} = \begin{pmatrix} \tilde{I}_{w,h,c} & \tilde{I}_{w,h,c} & \tilde{I}_{w,h,c} \\ \tilde{I}_{w,h,c} & \tilde{I}_{w,h,c} & \tilde{I}_{w,h,c} \\ \tilde{I}_{w,h,c} & \tilde{I}_{w,h,c} & \tilde{I}_{w,h,c} \end{pmatrix}
\]

\[L_{sym} = \frac{1}{3} \sum_{w,h,c} |F_{g,w,h,c}^{\ast} - F_{g,w,h,c}^{\ast}|\]
\[ L_{\text{sym}} = \frac{1}{W/2 \times H} \sum_{w,b}^{W/2,H} |I_{w,b} - \tilde{I}_{W-(w-1)b}| \quad (7) \]

Face images in real world are usually not strictly symmetric on the grey level. However, the pixel variance in the local region and gradient of each point are consistent, so it is more reasonable to define the symmetry loss in Laplacian space. Thus, we first transform the face image to Laplacian space and then adopt the symmetric loss on transformed images.

(5) Overall Loss: Consequently, the complete objective loss is defined as a weighted sum of four individual loss functions:

| Methods/Pose | W/o MFN | W/o exp. | W/o illu. | MFJNN |
|--------------|---------|----------|-----------|--------|
| ±90°         | 60.9    | 61.7     | 62.1      | 62.8   |
| ±75°         | 78.9    | 79.6     | 80.1      | 81.7   |
| ±60°         | 85.1    | 86.4     | 86.9      | 87.6   |
| ±45°         | 78.1    | 79.9     | 79.3      | 81.3   |
| ±30°         | 87.7    | 88.4     | 89.1      | 90.2   |
| ±15°         | 90.9    | 91.8     | 92.6      | 93.1   |
| Avg          | 80.2    | 81.3     | 81.7      | 82.8   |

**TABLE 1** Recognition rates on different poses on Multi-PIE. ‘w/o MFN’ represents the proposed MFJNN without multi-factor normalisation (i.e., only one encoder is included in the generator of the MFJNN), while ‘w/o illu.’ and ‘w/o exp.’, respectively represent the MFJNN without the illumination normalisation encoder and expression normalisation encoder.

**FIGURE 5** Comparison of face synthesis results with MFN and w/o MFN. From left to right, each column represents input images and the normalisation results of w/o MFN and with MFN, respectively.
\[
\begin{align*}
L_{D_0} &= L_{\text{adv}D_0}, \\
L_{G_0} &= L_{\text{adv}G} + \lambda_1 L_{id} + \lambda_2 L_p + \lambda_3 L_{\text{sym}}.
\end{align*}
\]  

where \(\lambda_1, \lambda_2\) and \(\lambda_3\) are parameters controlling the relative relation of objective terms. The MFJNN is optimised by alternatively optimising \(G_0\) and \(D_0\) for each training iteration.

4 | EXPERIMENTAL RESULTS

Here, the authors conduct experiments to evaluate the effectiveness of MFN, the synthetic image result and the face recognition performance of the proposed method. Now, we first describe the datasets and implementation details of our experiment. Then, we compare the performances of models with and without MFN to evaluate the effectiveness of MFN. We also present the visualisation of our synthesised normal face images results and compare the face recognition performance of the proposed method with those of the state-of-the-art methods. Finally, an ablation study is conducted to evaluate the effectiveness of each loss.

4.1 | Datasets and implementation details

Datasets Here, we train our model on the Multi-PIE [17] dataset. Multi-PIE consists of 754,204 images from 337 individuals. Each individual includes images under 20 illuminations, 15 poses, and six different expressions. These images can be divided into four sessions which are captured in different periods. In the proposed work, we use the entire Multi-PIE dataset including all pose, illumination and expression variations. For the two cameras above the head, their poses are labelled as \(\pm 45^\circ\). For the constrained evaluation, the first 200 identities are used for training and the remaining 137 identities are used for testing, where one image with a frontal pose and neutral illumination and expression for each identity is selected as the gallery set and the remaining identities are used as the probe set. Since multi-PIE has three kinds of labels (illumination, pose and expression), the MFJNN contains three encoders in our experiment. Correspondingly, the images with the same identity \(m\) but with non-normal illumination \(I_p^m\), a non-frontal pose \(I_p^m\), a non-normal expression \(I_p^m\) and a frontal pose and neutral illumination and expression \(I_p^m\) are selected to constitute a training sample pair \((I_p^m, I_p^m, I_p^m, I_p^m)\) for training. It is worth mentioning that the encoder in the MFJNN is mainly responsible for some factor normalisation but it may also normalise other factors incidentally to an inferior degree because training examples with only one non-normal factor are difficult to acquire.

To compare the results of the proposed method with those of the state-of-the-art methods, we also conduct experiments on Multi-PIE Setting-1 [19,21,25,32], which contains images of 250 identities in session one. The first 150 identities are used for training and the remaining 100 identities are used for

**Figure 6** Face normalisation results on Multi-PIE. Each pair presents the input image (left), the face normalised by the MFJNN (middle) and the face normalised by the TP-GAN [19] (right)
testing. Since the images in Multi-PIE Setting-1 do not have expression variation, we include only two encoders in the MFJNN (one is for pose and the other is for illumination) and the corresponding fusion coefficients are set as $\omega_p = 0.75$ and $\omega_i = 0.25$.

In the unconstrained experiment, we train the MFJNN on the entire Multi-PIE and test it on the LFW [29] and IJB-A [11] datasets. LFW consists of 13,233 images of 5749 individuals collected from the web. In the verification protocols of LFW, the test set contains 10 folds, each with 300 matched pairs and 300 unmatched pairs. IJB-A is a more challenging large pose dataset that has 5396 images and 20,412 video frames for 500 subjects in uncontrolled settings.

The Light CNN used as the feature extraction network and the source of the identity perceptive loss is pre-trained on MS-Celeb-1M [33] and kept fixed both in both the training and testing processes.

Implementation details The images from all datasets are pre-processed by applying the TCDCN [34] as the face detection algorithm and then cropped and resized to 224×224 images. The pixel values are normalised to the range of [−1,1]. Our network is implemented on TensorFlow [35]. We use Adam [36] with a learning rate of 0.0001 to train the generator and the discriminators by iteratively minimising the generator loss function and the discriminator loss function. We empirically set the hyperparameters of the loss function as follows: $\lambda_1 = 10$, $\lambda_2 = 0.001$ and $\lambda_3 = 0.3$. For the face recognition performance evaluation, the cosine-distance metric is used on the extracted feature from the Light CNN.

![Comparison of face synthesis](image)
4.2 | Effectiveness of multi-factor normalisation

To evaluate the effectiveness of MFN, we remove some factor normalisation encoders from the MFJNN and train three other models on the same training set (the first 200 identities of Multi-PIE), that is only including one or two encoders in the generator of the proposed model, and compare these models with the proposed model. We report the rank-1 recognition rates to compare the different models in Table 1. We can observe that the MFJNN can improve the performance compared to w/o MFN, w/o illumination, and w/o expression. Especially, the MFJNN achieves a 2.6% average performance improvement compared to w/o MFN. These results demonstrate that each factor normalisation is useful to improve the face recognition performance.

The face synthesis results of the proposed MFJNN and w/o MFN are shown in Figure 5. Although there is not a visually obvious illumination normalisation effect difference between the synthesised images from the MFJNN and w/o MFN, the recognition results from Table 1 have verified the effectiveness of the illumination normalisation in MFJNN on boosting recognition performance. In addition, it can be seen from Figure 5 that the MFJNN can achieve better expression normalisation results compared to w/o MFN, though w/o MFN can also slightly normalise expression.

4.3 | Face synthesis

This section mainly evaluates the MFJNN from two aspects, which are image quality and the ability to preserve an identity. The face synthesis results on Multi-PIE are illustrated in Figure 6. For comparison, the faces normalised by the TP-GAN [19] are also illustrated. We can see that our model can effectively normalise the pose, expression and illumination.

Figure 8 | Synthesis results on IJB-A. The odd columns are the input faces, and the even columns are the synthesis results of the MFJNN.
in various pose cases and synthesise photorealistic normalised face images while maintaining the identity information.

Figure 7 visualises the results of five images selected from LFW. From Figure 7, we see that the MFJNN synthesises face images with both good image quality and identity maintenance in an unconstrained environment. Although the TP-GAN and method in [23] can also generate photorealistic frontal faces on the LFW dataset, our MFJNN can normalise illumination and expression simultaneously while other methods only normalise pose variance (see results in the last three columns).

Figure 8 shows the synthesis results of the MFJNN on IJB-A. These results further demonstrate that the proposed model can simultaneously normalise pose, illumination and expression variations and synthesise high-quality and identity-preserving face images. Meanwhile, note that our model is trained on constrained Multi-PIE but can achieve better synthesis results on unconstrained datasets, which demonstrates that the proposed model has a good generalisation ability.

### 4.4 Face recognition

Here, the authors use the features extracted from the MFJNN and conduct face recognition on the Multi-PIE, LFW and IJB-A datasets. To compare the results of our method with those of other face recognition methods in a constrained environment, the training and testing data from Multi-PIE Setting-1 are used to perform the experiments.

Table 2 compares the recognition rates on Multi-PIE Setting-1. We observe that our method achieves comparable results with the state-of-the-art methods and better performance than the Light CNN on each different pose. Especially in extreme poses, our model achieves incredible improvements compared with the Light CNN, for example, from 2.6% to 59.2% under ±90°. Note that the recognition performance of our method is higher than that of the Light CNN that extracts features from non-normal face images using the Light CNN, which demonstrates the advantage of our face normalisation method.

For the unconstrained evaluation, we compare the proposed method with the state-of-the-art pose invariant face recognition methods, the DCNN [41], the DR-GAN [42] and the FF-GAN [20], and others on the LFW and IJB-A datasets. The comparison results on the LFW and IJB-A datasets are respectively shown in Tables 3 and 4. Our method outperforms many state-of-the-art methods. In addition, our method that extracts features from normal face images achieves consistently significant improvements compared with the Light CNN that extracts features from non-normal face images using the same feature extraction network, as shown in Table 4. Moreover, we also trained the model without multi-factor normalisation (i.e., w/o MFN) on the same training set (the entire Multi-PIE) and test it on LFW and IJB-A. The results from Tables 3 and 4 indicate that our model including multi-factor normalisation has better recognition performance than that of w/o MFN, which demonstrate the effectiveness of MFN in an unconstrained environment.

#### 4.5 Ablation study

To further evaluate the effectiveness of the proposed MFJNN and the contribution of each loss, we train three partial variants of the MFJNN: without the identity perceptive loss, without the pixel-wise loss and without the symmetric loss, that is, w/o \( L_{id} \), w/o \( L_{p} \) and w/o \( L_{sym} \), respectively.

| Table 2 | Recognition rate (%) on Multi-PIE Setting-1 |
|---------|---------------------------------------------|
| Method  | ±90° | ±75° | ±60° | ±45° | ±30° | ±15° |
| HPN [39] | 29.8 | 47.6 | 61.2 | 72.8 | 78.3 | 84.2 |
| c-CNN [40] | 47.3 | 60.7 | 74.4 | 89.0 | 94.1 | 97.0 |
| TP-GAN [19] | 64.0 | 84.1 | 92.9 | 98.6 | 99.9 | 99.8 |
| PIM [22] | 75.0 | 91.2 | 97.7 | 98.3 | 99.4 | 99.8 |
| CAPG-GAN [21] | 77.1 | 87.4 | 93.7 | 98.3 | 99.4 | 99.9 |
| Light CNN [26] | 2.6 | 10.5 | 32.7 | 71.2 | 95.1 | 99.8 |
| MFJNN | 59.2 | 80.9 | 93.6 | 98.6 | 99.6 | 99.9 |

| Table 3 | Face verification results on LFW |
|---------|---------------------------------|
| Methods | ACC (%) | AUC (%) |
| TP-GAN [19] | 91.17 ± 1.44 | 92.78 ± 0.02 |
| Hassner et al. [37] | 93.62 ± 1.17 | 98.38 ± 0.06 |
| HPEN [43] | 96.25 ± 0.76 | 99.39 ± 0.02 |
| FF-GAN [20] | 96.42 ± 0.89 | 99.45 ± 0.03 |
| AIF-CNN [44] | 96.63 ± 0.99 | 99.29 ± 0.42 |
| W/o MFN | 96.13 ± 0.92 | 98.31 ± 1.01 |
| MFJNN | 96.74 ± 0.79 | 99.39 ± 0.87 |

| Table 4 | Performance comparison on IJB-A dataset |
|---------|----------------------------------------|
| Method | Verification | Identification |
|        | FAR = 0.01 | FAR = 0.001 | Rank-1 | Rank-5 |
| OpenBR [11] | 23.6 ± 0.9 | 10.4 ± 1.4 | 24.6 ± 1.1 | 37.5 ± 0.8 |
| GOTS [11] | 40.6 ± 1.4 | 19.8 ± 0.8 | 43.3 ± 2.1 | 59.5 ± 2.0 |
| PAM [45] | 73.3 ± 1.8 | 55.2 ± 3.2 | 77.1 ± 1.6 | 88.7 ± 0.9 |
| DCNN [41] | 78.7 ± 4.3 | -- | 85.2 ± 1.8 | 93.7 ± 1.0 |
| DR-GAN [42] | 77.4 ± 2.7 | 53.9 ± 4.3 | 85.5 ± 1.5 | 94.7 ± 1.1 |
| FF-GAN [20] | 85.2 ± 1.0 | 66.3 ± 3.3 | 90.2 ± 0.6 | 95.4 ± 0.5 |
| AIF-CNN [44] | 80.4 ± 3.3 | 60.0 ± 8.6 | 92.2 ± 2.3 | 97.4 ± 0.9 |
| Light CNN [20] | 82.7 ± 2.0 | 67.4 ± 2.2 | 84.5 ± 1.7 | 92.6 ± 0.9 |
| W/o MFN | 84.4 ± 1.1 | 69.8 ± 1.7 | 92.3 ± 0.7 | 96.1 ± 0.8 |
| Proposed MFJNN | 86.5 ± 0.9 | 71.7 ± 4.6 | 93.9 ± 0.5 | 97.6 ± 0.6 |
Figure 9 illustrates the face visual effects of the synthesised face images generated by the MFJNN and its different variants. We can observe that the synthesis images of the MFJNN have better visual effects and preserve more identity information of the input face image compared with the results of other variants. Especially, it can be seen that the synthesis images without the constraint of $L_{id}$ tend to deviate from the identity feature of the input image due to lack of a constraint of the consistency of identity information, which indicates the effectiveness of the identity perceptive loss.

| Method | ±90° | ±75° | ±60° | ±45° | ±30° | ±15° |
|--------|------|------|------|------|------|------|
| W/o $L_{id}$ | 48.8 | 74.6 | 82.2 | 81.8 | 90.3 | 92.4 |
| W/o $L_p$ | 55.3 | 76.7 | 90.4 | 96.9 | 94.8 | 96.7 |
| W/o $L_{sym}$ | 57.5 | 77.9 | 92.4 | 98.1 | 99.1 | 99.3 |
| MFJNN | 59.2 | 80.9 | 93.6 | 98.6 | 99.6 | 99.9 |
Table 5 compares the recognition rates of different variants of the MFJNN on Multi-PIE Setting-1. The proposed MFJNN achieves better performance than all of its variants under different poses. These results suggest that each loss in the MFJNN is essential to obtaining a photorealistic and identity-preserving face image and improving the face recognition performance.

5 | CONCLUSION

Here, the authors have proposed a MFJNN for pose-invariant face recognition. The proposed method can realise pose normalisation when other factors vary except pose using multi-factor normalisation, which can improve face recognition performance on normalised results. From the experimental evidence, it can be concluded that the proposed method can effectively normalise poses, illuminations and expressions; the features extracted from normalised results can boost the face recognition performance; and also achieves better face recognition performance than the state-of-the-art methods.

ACKNOWLEDGEMENTS

The work of authors is supported by the National Natural Science Foundation of China (Grant No. 61,502,444), the youth project of science and technology research programme of Chongqing Education Commission of China (Grant No. KJQN20180119) and the Scientific Research Foundation of Chongqing University of Technology (Grant No. 2017ZD58).

REFERENCES

1. Ranzato, M., et al.: On deep generative models with applications to recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), pp. 2857–2864. IEEE, Colorado Springs (2011)
2. Zhu, Z., et al.: Deep learning identity preserving face space. Proc. ICCV, Sydney 1, 113–120 (2013)
3. Tang, Y., Sakhbudinov, R., Hinton, G.: Robust Boltzmann machines for recognition and denoising. In: IEEE conference on computer vision and pattern recognition (CVPR), pp. 2264–2271. IEEE, Providence (2012)
4. Zhang, Y., et al.: Occlusion-robust face recognition using iterative stacked denoising autoencoder. In: Neural information processing, pp. 352–359. Springer (2013)
5. Ian J Goodfellow, D.W.-F., et al.: Generative adversarial nets. In: International conference on neural information processing systems, pp. 2672–2680. Cambridge (2014)
6. Chen, D., et al.: Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In: Computer vision and Pattern recognition, pp. 3025–3032. Portland (2015)
7. Tan, H., Yang, B., Ma, Z.: Face recognition based on the fusion of global and local HOG features of face images. IET Comput. Vis. 8(3), 224–234 (2013)
8. Deng, J., et al.: ArcFace: additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF conference on computer vision and Pattern recognition (CVPR), pp. 4685–4694. Long Beach (2019)
9. Li, Y., et al.: Distance metric Optimization Driven convolutional neural network for Age invariant face recognition. Pattern recognition, pp. 51–62 (2017).
10. Minace, S., et al.: Biometric recognition using deep learning: a Survey (2019). arXiv:1912.00271
11. Klare, B.F., et al.: Pushing the frontiers of unconstrained face detection and recognition: LARPA Janus Benchmark A, pp. 1931–1939 (2015)
12. Huang, G.B., et al.: Labelled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on faces in ‘real-life’ images: detection, alignment, and recognition (2008)
13. Li, S., et al.: Morphable displacement field based image matching for face recognition across pose. In: Eccv, pp. 102–115. Springer Florence (2012)
14. Asthana, A., et al.: Fully automatic pose-invariant face recognition via 3D pose normalisation. In: IEEE international conference on computer vision (ICCV), pp. 937–944. IEEE, Barcelona (2011)
15. Castillo, C.D., Jacobs, D.W.: Wide-baseline stereo for face recognition with large pose variation. In: IEEE conference on computer vision and pattern recognition (CVPR), pp. 537–544. IEEE, Colorado Springs (2011)
16. Li, A., Shan, S., Gao, W.: Coupled bias-variance tradeoff for cross-face recognition. IEEE Trans. Image Process. 21(1), 305–315 (2012)
17. Gross, R., et al.: Multi-PIE. In: 8th IEEE international conference on automatic face & gesture recognition (FG), pp. 1–8. Amsterdam (2008)
18. Martinez, A.M.: The AR face database. CVC Technical Report. 24 (1998)
19. Huang, R., et al.: Beyond face Rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis. 2017 IEEE international conference on computer vision (ICCV), Venice, Italy (2017)
20. Yin, X., et al.: Towards large-pose face frontalization in the wild, pp. 3990–3999. ICCV Venice, Italy (2017)
21. Hu, Y., et al.: Pose-guided photorealistic face Rotation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp. 8398–8406. Salt Lake City (2018)
22. Jian, Z., et al.: Towards pose invariant face recognition in the wild. In: Cvpr, Salt Lake City (2018)
23. Bao, J., et al.: Towards open-set identity preserving face synthesis. CVPR, Salt Lake City (2018)
24. Yujun Shen, P.L., et al.: FaceID-GAN: learning a symmetry three-player GAN for identity-preserving face synthesis. CVPR, Salt Lake City (2018)
25. Qian, Y., Deng, W., Hu, J.: Unsupervised face normalisation with extreme pose and expression in the wild. CVPR, Long Beach (2019)
26. Xiang, W., et al.: A Light CNN for deep face representation with Noisy labels. IEEE Trans. Inf. Forens. Secur. 13(11), 2884–2896 (2018)
27. Parkhi, A.V.O.M., Zisserman, A.: Deep face recognition. In: Presented at the British Machine vision Conference, Swansea (2015)
28. Khan, A., Wabah, N.: Deep residual learning for image recognition. CVPR, Las Vegas (2016)
29. Jie, C., et al.: Learning a high fidelity pose invariant model for high-resolution face frontalization. NIPS’16: Proceedings of the 32nd International Conference on Neural Information Processing Systems (2016)
30. Maas, A.L., Hannum, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICMIL Workshop on Deep Learning for Audio, Speech and Language Processing. vol. 30, pp. 1–6 (2013)
31. Dumoulin, V., Visin, F.: A guide to convolution arithmetic for deep learning, eprint arXiv:1603.07285, pp. 03–23 (2016)
32. Junho, Y., et al.: Rotating your face using multi-task deep neural network. In: 2015 IEEE conference on computer vision and Pattern recognition (CVPR), pp. 676–684. Boston (2015)
33. Guo, Y., et al.: MS-Celeb-1M: a dataset and Benchmark for large-scale face recognition. In: European conference on computer vision, pp. 87–102. Amsterdam (2016)
34. Zhang, Z., et al.: Facial Landmark detection by deep multi-task learning. In: Computer vision ECCV, pp. 94–108. Springer (2014)
35. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning (2016). arXiv:1605.08669
36. Kingma, D., Ba, J.: Adam: a method for Stochastic Optimization. International Conference on Learning Representations (ICLR), Comput. Sci, San Diego (2014)
37. Hassner, T., et al.: Effective face frontalization in unconstrained images. CVPR, Boston (2015)
38. Zhu, X., et al.: High-fidelity pose and expression normalization for face recognition in the wild. In: 2015 IEEE Conference on computer vision and Pattern recognition (CVPR). Boston (2015)
39. Ding, C., Tao, D.: Pose-invariant face recognition with homography-based normalization. Pattern Recogn. 66, 144–152 (2017)
40. Xiong, C., et al.: Conditional convolutional neural network for modality-Aware face recognition. In: 2015 IEEE international conference on computer vision (ICCV), pp. 3667–3675. Santiago (2015)
41. Chen, J., Patel, V.M., Chellappa, R.: Unconstrained face verification using deep CNN features. In: 2016 IEEE winter conference on applications of computer vision (WACV), pp. 1–9. Lake Placid (2016)
42. Luan, T., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: Computer vision and pattern recognition, pp. 1283–1292. Honolulu (2017)
43. Zhao, H., et al.: Loss functions for image restoration with neural networks. IEEE transactions on computational imaging. 3(1), 47–57 (2017)
44. Zhang, Z., et al.: Face frontalization using an appearance-flow-based convolutional neural network. IEEE Trans. Image Process. 28(5), 2187–2199 (2019)
45. Masi, I., et al.: Pose-aware face recognition in the wild. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp. 4838–4846. Las Vegas (2016)

How to cite this article: Liu Y, Chen J. Multi-factor joint normalization for face recognition in the wild. IET Comput. Vis. 2021;1–13. https://doi.org/10.1049/cvi2.12025