Unsupervised Face Normalization with Extreme Pose and Expression in the Wild

Yichen Qian[1,2], Weihong Deng[1]*, Jiani Hu[1]

[1]Beijing University of Posts and Telecommunications
[2]AI Labs, Didi Chuxing, Beijing 100193, China
{mx54039q, whdeng, jnhu}@bupt.edu.cn

Abstract

Face recognition achieves great success thanks to the emergence of deep learning. However, many contemporary face recognition models still have limited invariance to strong intra-personal variations such as large pose changes. Face normalization provides an effective and cheap way to distill face identity and dispell face variances for recognition. We focus on face generation in the wild with unpaired data. To this end, we propose a Face Normalization Model (FNM) to generate a frontal, neutral expression, photorealistic face image for face recognition. FNM is a well-designed Generative Adversarial Network (GAN) with three distinct novelties. First, a face expert network is introduced to construct generator and provide the ability of retaining face identity. Second, with the reconstruction of normal face, a pixel-wise loss is applied to stabilize optimization process. Third, we present a series of face attention discriminators to refine local textures. FNM could recover canonical-view, expression-free image and directly improve the performance of face recognition model. Extensive qualitative and quantitative experiments on both controlled and in-the-wild databases demonstrate the superiority of our face normalization method. Code is available at https://github.com/mx54039q/fnm

1. Introduction

Unconstrained face recognition [25] is an important but extremely challenging problem. Large pose, expression[19] and lighting remain main obstacles for further pushing unconstrained face recognition performance. Some works [3, 22, 29] address the pose problem by learning pose-invariant features, while some others [6, 20, 31, 16, 33] try to synthesize an identity-preserved frontal face. Photorealistic frontal view synthesis from a single face image has an abundance of applications other than face recognition.

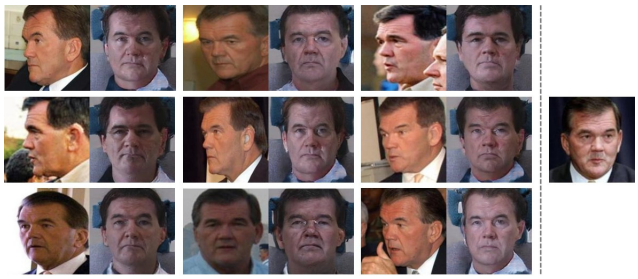

Figure 1. Face normalization results under the same identity in unconstrained environment. Face images are under different views across pose, lighting, expression and background. FNM can keep a high-level consistency in preserving identity. On the right of the dashed line is a near-normal face of the same identity.

However, face rotation (especially face frontalization) in unconstrained environment has two major difficulties: complex face variations besides pose, and unpaired data. Compared to controlled environment, there are more complex face variations, *e.g.*, lighting, head pose, expression, self-occlusion in real-world scenarios. It is difficult to directly warp input face to a normalized view. While obtaining strictly normalized face is undoubtedly expensive and time-consuming, we can not get effective supervision of target normalized face (*i.e.*, front-facing, neutral expression) corresponding to an input face. Face synthesis lays great stress on facial texture, which is difficult without supervision of target normalized face.

We present a method to normalize face into a front-facing, neutral expression view. We propose a Face Normalization Model (FNM) completely based on neural network to solve the two problems above simultaneously. Introduced by Goodfellow *et al.* [10], the Generative Adversarial Network (GAN) maps from a source data distribution to a target data distribution using a min-max two-player game between a generator network and a discriminator network. In this

*Corresponding author

work, adversarial loss encourages our generator to synthesize normal face (target distribution) from non-normal face (source distribution). Inspired by Cole *et al*. [6], we employ a face expert network in our generator to produce identity features. Trained on large-scale face dataset with large variations in pose, age, lighting, ethnicity and profession, the face expert network is robust enough across complex face variations in real-world scenarios. Many previous works commonly import existing knowledge and make learning much more efficient. While previous face synthesis methods usually directly warp image to image, our model eases the task of the generator and maps from identity features to face image. Intuitively, the Face expert network provides strong prior knowledge. Rewarding features similarity of normalized face and input face, it can keep the identity of face during the transformation.

Integrating GAN with expert face network, we can achieve the target of face normalization theoretically. However, the result is dissatisfied with such two simple parts due to the speciality of face synthesis. Normalized faces are similar in outline but different in detail, so the generator's task is arduous without paired data (*i.e*., input face and target view face of the same person). While the image-wise loss of GAN works on the whole face image, we need a pixel-wise loss to guide and stabilize the face synthesis process.

We suggest to employ a more elaborate architecture that contains two modifications. First, we introduce a new term of pixel-wise loss in unpaired data problem. When the generator reconstructs normal face to itself, a pixel-wise loss could be applied to stabilize the optimization process. Second, face attention mechanism is proposed to refine local facial texture. Without any assistant techniques (*e.g*., 3D face model and landmark localization), our face attention model is simple but effective. Unlike previous methods that crop local areas of input face with landmark localization, we construct a series of attention discriminators in the fixed areas of generated normalized face. With the prior knowledge of facial attribute, attention discriminators would automatically enhance the quality of local facial texture.

Our model employs a publicly-available face expert network, VGG-Face2 [4], to produce face identity features and preserve identity while generating the normalized face image. Recovering a face image from a particular feature vector presents an interesting approach in understanding deep networks' predictions. In other words, our model provides a novel method to analyse and visualize the feature space of the face recognition model.

This paper makes the following contributions. 1) A Face Normalization Model (FNM) is proposed to synthesize a canonical view and identity-preserved face from a single face image. Incorporating with face expert network in a novel way, it develops an effective and novel training strategy for unpaired data and extreme face variations in the wild. 2) Introduce a pixel-wise loss by normal face reconstruct, which leads to a healthy optimization process. Compared with the image-wise adversarial loss, the pixel-wise loss encourages image content consistency and greatly stabilizes the training process in unsupervised face normalization. 3) Attention mechanism is applied to reinforce the realism and quality of normalized face. 4) Although FNM does not contain any recognition module, it can improve the performance of traditional face recognition frameworks by "stitching" face normalization to them. As a preprocessing procedure, FNM helps to distill face identity and dispell face variations before face recognition procedure. We conduct qualitative and quantitative experiments on various benchmarks, including both controlled and in-the-wild datasets. The results demonstrate the effectiveness of FNM on boosting face recognition model.

2. Related Works

Generative Adversarial Network (GAN) Since GAN first introduced by Goodfellow *et al*. [10], its surprising performance on generative task has drawn substantial attention from the deep learning and computer vision community. The GAN framework learns a generator network G and a discriminator network D with competing loss. The min-max two-player game provides a simple yet powerful way to estimate target distribution and generate novel image samples [7]. Mirza and Osindero [23] introduce the conditional GAN, to control the generator and discriminator for effective image-to-image generating. Arjovsky *et al*. [2] introduce Wassertain distance and propose Wasserstein GAN (WGAN), which makes progress toward stable training of GANs. These successful improvements of GAN motivate us to develop face normal view synthesis, in the harsh conditions of unconstrained environment, unpaired data and no auxiliary 3D face model.

Face Normalization Synthesizing a frontal, neutral expression face from a single image in unconstrained environment is very challenging because of extreme variations such as large pose. Hassner *et al*. [13] adopt 3D face model to register and produce frontal face. Zhu *et al*. [34] provide a high fidelity pose and expression normalization method based on 3DMM. The results of 3D-based methods are often not realistic enough with artifacts and severe texture losses. These methods suffer big performance drop on large pose.

Benefiting from deep learning, FF-GAN [31] incorporates 3D face model into GAN to solve the problem of large-pose face frontalization in the wild. Considering photorealistic and identity preserving frontal view synthesis, Huang *et al*. [16] propose TP-GAN with global and local aware networks under large pose. Extending TP-GAN, Zhao *et al*. [33] propose PIM with introducing a domain adaptation strategy for pose invariant face recognition.

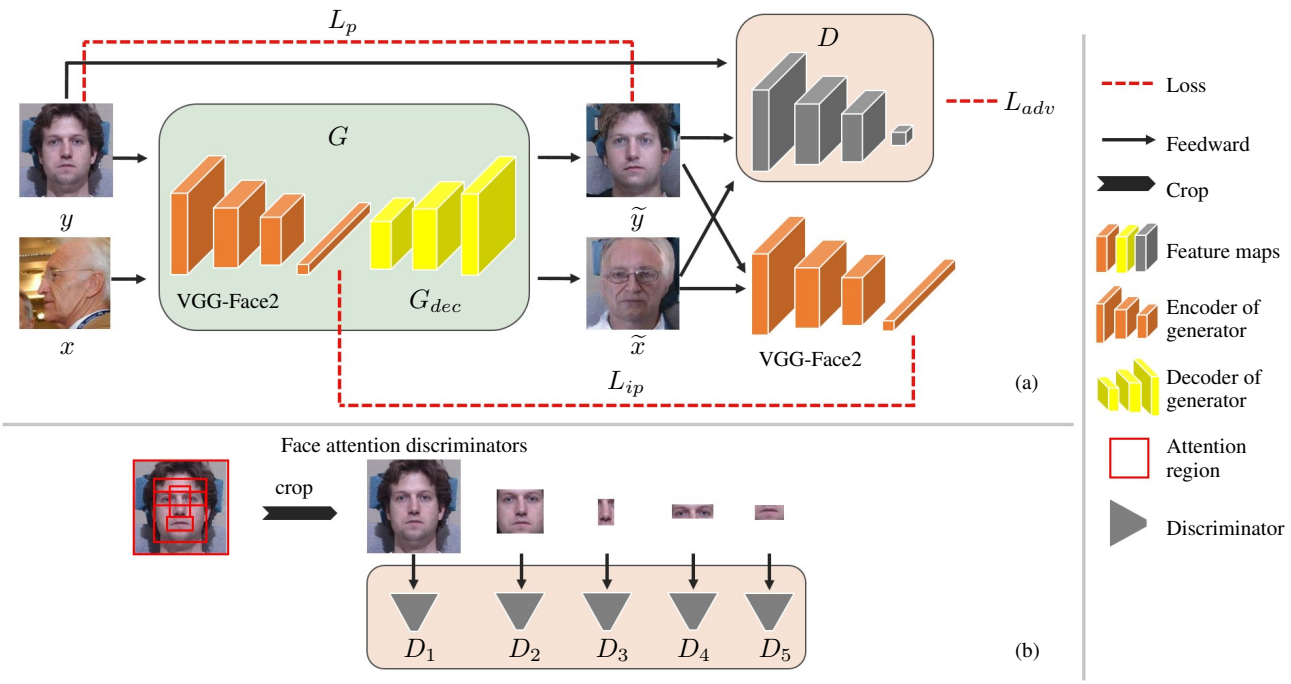

Figure 2. Architecture of FNM. The feed-ward process is shown in (a), where x from non-normal face set is fed into the generator to extract identity features and generate the normalized view face $\widetilde{x}$. In order to introduce pixel-wise loss, the generator also produces a virtual normalized image $\widetilde{y}$ for y from normal face set. Attention discriminators are used to distinguish the normalized samples $\widetilde{x}$ and $\widetilde{y}$ from the real samples y. The details of the proposed attention discriminators are shown in (b). Fixed regions of face image are cropped and fed into corresponding attention discriminators, which are marked by red box. Losses are drawn with red dashed lines, where L_p is a pixel-wise consistency for normal face, L_{adv} is the adversarial loss for encouraging G to synthesize a photorealistic normalized face, L_{ip} is the identity perception loss for preserving the identity information. All of the notations are listed in (c).

The proposed FNM differs from prior works in following aspects: 1) FNM incorporates a face expert network to solve the problems of complex variations and unpaired data in the wild. As trained on paired data in controlled environment, TP-GAN [16] and PIM [33] might degenerate performance in unconstrained environment. 2) While most prior works focus on face frontalization [33, 16, 31], FNM considers face variations besides pose and makes further efforts to distill identity and dispell face variations. 3) Incorporating with 3D face model, FF-GAN [31] employs a complex architecture and suffers from great optimization difficulty. Our FNM is an end-to-end deep learning model. FNM introduces a pixel-wise loss for reconstruction of faces from the normal face set (frontal faces with neutral expression). This pixel-wise loss leads to a more stable optimization.

Disentangled Representation via Generation Face normalization (or face frontalization) may be considered an image-level disentangled representation. The advantage of this category is that it can be easily incorporated into off-the-shelf face recognition framework as a pre-processing procedure. Tran *et al*. [20] propose DR-GAN to rotate face and explicitly disentangle the identity representation by using the pose code. Hu *et al*. [29] propose CAPG-GAN that uses a landmark heatmap to control the face rotation. FF-GAN [31], TP-GAN [16] and PIM [33] share the same objective of pose-invariant recognition via face frontalization. While most prior works learn a representation that is only invariant to pose, our method can learn a representation invariant to other attributes besides pose. From a face image under arbitrary condition, FNM can synthesize a virtual canonical-view face image while preserving face identity.

3. Approach

Face normalization aims to synthesize a canonical-view face from a single face image, while preserving face identity. GAN [10] transforms non-normal face set X to normal face set Y, while the face expert network preserves face identity. In addition, a pixel-wise loss and face attention mechanism are applied for high-quality synthesis. The overall framework of our proposed Face Normalization Model (FNM) is depicted in Fig. 2

3.1. Employing Face Expert Network

Our key target is to synthesize from a face taken in the wild to a front-facing and neutral expression face, retaining identity as much as possible. Face expert network has prominent discriminative capability and is efficient in map-

ping face image to identity feature. Most previous works only use face model to maintain perceptual similarity of generated face and input face. In our model, face recognition network is also used as encoder network to initially distill identity information and dispell non-identity information.

As shown in Fig. 2, we employ a publicly-available face expert network, VGG-Face2 [4] network, both as the encoder of generator G_{enc} and as a source of identity loss. Our assumption is that the face expert network is a strong prior knowledge on face recognition. In response, we keep the face expert network fixed and do not update its parameters during training. For preserving face identity, we penalize the feature distance between normalized face and input face.

Both employed in generator and used for identity preserving, the face expert network $G_{enc} : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^D$ produces reliable identity features of input face image, where H, W, and C denote the image height, width, and channel number respectively, and D is the size of feature vector.

3.2. Generator

We employ the face expert network as the encoder of the generator to map input face to feature space, denoted as G_{enc}. A decoder G_{dec} is constructed to make a further effort on distilling identity by recovering a normalized (*i.e.*, front-facing, neutral expression) face. Specifically, the generator is stated as $G = G_{enc} \circ G_{dec}$, for a learned function G_{enc} (VGG-Face2).

More formally, let face image from non-normal set be denoted by $x \in \mathbb{R}^{H\times W\times C}$ and the responding generated face image be denoted by $\widetilde{x} \in \mathbb{R}^{H\times W\times C}$, then

$$\widetilde{x} := G_{dec}(G_{enc}(x)), \tag{1}$$

When function of G is normalizing arbitrary input face, it is obvious that G would reconstruct normal face to itself. To this end, we also input normal face to G:

$$\widetilde{y} := G_{dec}(G_{enc}(y)), \tag{2}$$

where y denotes face image from normal face set, and $\widetilde{y}$ denotes corresponding normalized face image. The new term of normal-to-normal mapping makes it possible to introduce pixel-wise loss, which guides and stabilizes the optimization of GAN in condition of unpaired data.

We generate normal face image using a fully convolutional network. As applied on identity features from G_{enc}, G_{dec} consists of a set of stacked layer groups (transposed convolution layer [9], ReLU layer and Residual block [14]). Finally, we apply an 1×1 convolution to yield $224\times224\times3$ RGB values.

An ideal generator will warp non-normal face to photorealistic normalized face, and keep consistent on normal face. Meanwhile, preserving the identity information is crucial for face recognition. Employing face expert network in generator is an essential part in our model. Extracting face identity features from a face in unconstrained environment is more difficult than generating a normalized face from identity features. Intuitively, this idea eases the difficulty of the generator by more than a half.

3.3. Face Attention Discriminators

We introduce a series of discriminators to distinguish between real normal face images and generated normal face images. Considering face characteristic, these discriminators have different receptive fields. More specifically, we crop the regions of eyes, nose, mouth and face to construct face attention discriminators, while an addition discriminator receives the entire image. As shown in Fig. 2, we construct five discriminators ($D_k, k = 1, 2, 3, 4, 5$) with five corresponding attention regions respectively.

Different from general image generation task, face synthesis attaches great importance to local facial texture. The idea of image attention is applied in DA-GAN [21] and TP-GAN [16]. Unlike previous works that crop attention regions of input image, our FNM pays attention on fixed regions of output normalized face. Integrating the attention mechanism into face synthesis produces photorealistic face image with great quality.

3.4. Loss Function

The key objective of our FNM is to normalize an arbitrary face, while the synthesized face should keep the identity of the input face and look like a real face. Two losses are proposed to basically meet the requirements, denoted by L_{adv}, L_{ip} as following:

$$L_{adv} = \sum_{k=1}^{5} D_k(\widetilde{x}_k) + \sum_{k=1}^{5} D_k(\widetilde{y}_k) - \sum_{k=1}^{5} D_k(y_k), \tag{3}$$

$$L_{ip} = ||G_{enc}(x) - G_{enc}(\widetilde{x})||_2^2 + ||G_{enc}(y) - G_{enc}(\widetilde{y})||_2^2, \tag{4}$$

where subscript k is number of attention discriminators and corresponding regions, L_{adv} is the **adv**ersarial loss for domain adaptation from source distribution (*i.e.*, non-normal face set) to target distribution (*i.e.*, normal face set) and adding realism to the synthesized images, L_{ip} is the **i**dentity **p**erception loss for preserving the identity information. $||\cdot||_2^2$means the vector 2-norm.

Proposed by Arjovsky *et al*. [2], Wasserstein distance is effective in stabilizing the optimization process of GAN. We apply WGAN-GP [12] loss in our model instead of original cross entropy loss. To be specific, the outputs of the discriminators are directly applied to loss function without sigmoid activating.

Figure 3. Face normalization results on IJB-A [18] under extreme pose, express, lighting, occlusion and front view.

Besides the elaborate architecture presented in Sec. 3.3 and Sec. 3.2, we suggest to employ a new term for the problem of unpaired data. The image-wise adversarial loss would result in distorted face contour and volatile optimization process. We would like the generator G to behave like an identity matrix when applied to faces from the normal set. It is an essential part for both texture information preserving and stable optimization. We introduce a **p**ixel-wise consistency before and after face normalization on normal face, denoted by L_p as following:

$$L_p = \frac{1}{W \times H \times C} \sum_{w,h,c}^{W,H,C} |y_{w,h,c} - \widetilde{y}_{w,h,c}|, \qquad (5)$$

where w, h, c traverse all pixels and channels of y and $\widetilde{y}$.

We have three forms of loss for warping face, preserving identity and promoting performance respectively. The overall objective function for FNM is:

$$\begin{cases} L_D = L_{adv}, \\ L_{G_{dec}} = -L_{adv} + \lambda_1 L_{ip} + \lambda_2 L_p. \end{cases} \qquad (6)$$

We optimize FNM by alternatively optimizing D and G_{dec} for each training iteration.

4. Experimental Results

FNM aims for synthesizing a canonical view and identity-preserved face in extreme unconstrained environment. Face normalization is an image-level disentangled representation. In Sec. 4.2, we show qualitative face normalization results of FNM. In Sec. 4.3, we quantitatively evaluate face recognition performance on boosting face recognition models under both the controlled and in-the-wild settings. In Sec. 4.4, we further conduct experiments with different architectures and loss functions to analyse respective roles.

4.1. Experimental Settings

Databases: IJB-A [18] is one of the most challenging unconstrained face recognition benchmark dataset with uncontrolled pose variations. IJB-A[18] contains both images and video frames from 500 subjects with 5,397 images and 2,042 videos that are split into 20,412 frames, 11.4 images and 4.2 videos per subject, captured from in-the-wild environment to avoid the near frontal bias, along with protocols for evaluation of both *verification* (1:1 comparison) and *identification* (1:N search) tasks. For testing, 10 random splits are provided by each protocol, respectively.

The CMU Multi-PIE database [11] is the largest database for evaluating face synthesis and recognition in the controlled setting. Multi-PIE allows for a graded evaluation with respect to pose, illumination, and expression variations. Thus, it is an important database to validate the performance of our method with respect to prior works on face synthesis. We conduct experiments on Multi-PIE with Setting-1 [30, 16, 29], which contains faces of 250 subjects. The training set is composed of all the images (13 poses and 20 illumination levels) of the first 150 identities, *i.e.*, $150 \times 13 \times 20 = 39,000$ images in total. For testing, one gallery image under frontal view and normal illumination is used for each of the remaining 100 subjects. The numbers of the probe and gallery sets are 24,000 and 100 respectively.

Implementation details: In unconstrained experiment, the non-normal face set contains 297,369 face images from unconstrained dataset CASIA-WebFace [28], while the normal face set contains 5,000 face images (front pose and 20 illuminations of 250 identities) from Multi-PIE [11]. In constrained experiment, we separate training set of the Multi-PIE Setting-1 into non-normal set (12 poses and 20 illuminations of 150 identities) and normal set (front pose

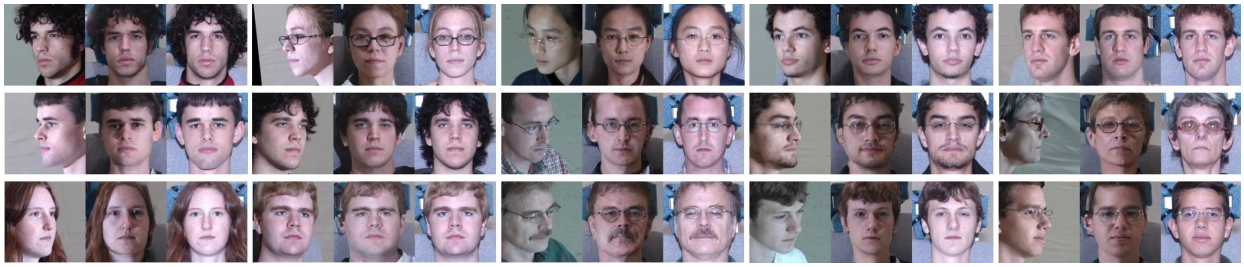

Figure 4. Synthesis results on Multi-PIE. Each pair presents profile (left), normalized face (middile) and ground truth normal face (right).

Table 1. Performance comparison on IJB-A. The results are averaged over 10 testing splits. Symbol "-" implies that the result is not reported for that method. FNM is incorporated into two face recognition framework VGG-Face [24] and Light CNN [26] as a pre-processing procedure.

Method	Verification		Identification	
	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
OpenBR [18]	23.6±0.9	10.4±1.4	24.6±1.1	37.5±0.8
GOTS [18]	40.6±1.4	19.8±0.8	43.3±2.1	59.5±2.0
PAM [22]	73.3±1.8	55.2±3.2	77.1±1.6	88.7±0.9
DCNN [5]	78.7±4.3	-	85.2±1.8	93.7±1.0
DR-GAN [20]	77.4±2.7	53.9±4.3	85.5±1.5	94.7±1.1
FF-GAN [31]	85.2±1.0	66.3±3.3	90.2±0.6	95.4±0.5
VGG-Face [24]	86.8±1.8	68.4±3.3	92.8±0.8	97.9±0.6
FNM+VGG-Face	88.8±1.9	69.0±4.6	94.6±0.5	98.4±0.5
Light CNN [26]	82.7±2.0	67.4±2.2	84.5±1.7	92.6±0.9
FNM+Light CNN	**93.4**±0.9	**83.8**±2.6	**96.0**±0.5	**98.6**±0.3

and 20 illuminations of 150 identities). As our objective is face normalization with unpaired data, we strictly keep the same setting on constrained and unconstrained environments. In particular, we do not use paired data and identity information under both environments, which is available in controlled environment.

We pre-process the images by applying an off-the-shelf face detection algorithm [32] and crop to 250×250 image size across all the databases. The G_{enc} is constructed on public-available pretrained ResNet-50 [14] from VGG-Face2 [4]. We keep the G_{enc} fixed both in training and testing process. Our network is implemented on Tensorflow [1]. We train the discriminators and the generator by iteratively minimizing the discriminator loss function and the generator loss function in sequence with Adam [17]. We empirically set the hyperparameters of the loss functions as follows: $\lambda_1 = 10$, $\lambda_2 = 0.001$. We set the hyperparameters of the optimizer as follows: $\alpha = 10^{-4}$, $\beta_1 = 0$, $\beta_2 = 0.9$, $\epsilon = 10^{-8}$. Please refer to supplementary material for full details on network architectures and training procedures.

4.2. Qualitative Results

As shown in Fig. 3, FNM can generate high-fidelity and identity-preserved normal face on unconstrained dataset IJB-A [18]. There are intricate face variations in unconstrained environment. These results demonstrate robustness of FNM to large pose, lighting, occlusion and expression. Specially, our FNM performs well in large-pose challenge with 90° yaw angle. It's worth noting that the difficulty of face normalization on unconstrained environment lies in not only extreme variations but mixture of these variations. With the robustness of face recognition model, our model generates normalized face from high-level semantic feature instead of image. Surprisingly, we observe FNM's ability in super resolution, while we don't have special setting in training process.

Table 2. Rank-1 recognition rates (%) across poses and illuminations under Multi-PIE Setting-1. FNM is incorporated into two face recognition framework VGG-Face [24] and Light CNN [26] as a pre-processing procedure.

Method	±90°	±75°	±60°	±45°	±30°	±15°
HPN [8]	29.82	47.57	61.24	72.77	78.26	84.23
c-CNN [27]	47.26	60.7	74.4	89.0	94.1	97.0
TP-GAN [16]	64.0	84.1	92.9	98.6	99.9	99.8
PIM [33]	75.0	91.2	97.7	98.3	99.4	99.8
CAPG-GAN [29]	77.1	87.4	93.7	98.3	99.4	99.9
VGG-Face [24]	2.1	5.8	38.0	73.5	85.8	94.9
FNM+VGG-Face	41.1	67.3	83.6	93.6	97.2	99.0
Light CNN [26]	2.6	10.5	32.7	71.2	95.1	99.8
FNM+Light CNN	55.8	81.3	93.7	98.2	99.5	99.9

The proposed FNM provides a further insight into the structure of the identity feature space. Moustache, hair and glasses are preserved in normalizing procedure. Normal face set is from controlled environment with the same background, which likely results in overfitting on the images' backgrounds. From another aspect, it also verifies that face recognition would not be affected by background.

Further synthesis results on the controlled database Multi-PIE are shown in Fig. 4.

4.3. Quantitative Results

We conduct unconstrained face recognition (*i.e.*, verification and identification) on IJB-A database to quantitatively verify the superiority of FNM on "recognition via generation". In addition, we evaluate our model on controlled database Multi-PIE for comparison. FNM is incorporated into two pre-trained face recognition models, VGG-Face [24] and Light CNN [26]. More specifically, we use

Figure 5. The results produced by two variations of FNM. (a) Input face. (b) FNM without L_p. (c) FNM without face attention mechanism. (d) our FNM.

Table 3. Component analysis: rank-1 recognition rates (%) across poses and illuminations under Multi-PIE Setting-1. Light CNN [26] is choose as baseline.

Method	±90°	±75°	±60°	±45°	±30°	±15°
Light CNN [26]	2.6	10.5	32.7	71.2	95.1	99.8
w/o L_p	46.9	62.0	70.4	78.5	81.2	90.3
w/o attention	41.3	66.6	83.4	92.3	96.0	97.6
FNM+Light CNN	55.8	81.3	93.7	98.2	99.5	99.9

FNM as a pre-processing procedure and then apply the same face recognition process. The results of VGG-Face and Light CNN on the original images are used as the baselines of our method.

As shown in Table 1, we evaluate the face recognition performance on our normalized images of IJB-A database with two baselines and other state-of-the-art methods. The results of VGG-Face [24] and Light CNN [26] demonstrate that FNM has a clear advantage to enhance the performance of face recognition model. Our method achieves consistently significant improvement compared to the baseline methods. In particular, FNM with Light CNN achives 10.7% improvement at FAR 0.01 and 16.4% improvement at FAR 0.001 on face verification, 7.5% improvement at Rank-1 and 6.0% improvement at Rank-5 on identification. Face alignment is not necessary for VGG-Face, but necessary for Light CNN. Our FNM provide a different way of face alignment in unconstrained environment. It might be the reason that FNM performs better on Light CNN than on VGG-Face.

As shown in Table 2, we evaluate the face recognition performance on our normalized images of Multi-PIE database. The accuracies of FNM achieve comparable performance with the state-of-the-art methods. The performance gap with the other methods lies in two points: 1) These methods fine-tune the baseline (Light-CNN) on the Multi-PIE database, while our FNM is directly incorporated to face recognition model; 2) These methods train with paired data and identity information while our FNM keeps the same training strategy with uncontrolled environments. Under extreme pose, our FNM achieves incredible improvements (*i.e*. 2.6% to 55.8% under ±90°)

Figure 6. Comparison of face frontalization on LFW[15].

4.4. Ablation Study

To verify the superiority of FNM as well as the contribution of each component, we train two partial variants of FNM in terms of without L_p and without face attention mechanism (*i.e*., only one discriminator for the whole image). Fig. 5 illustrates the perceptual performance of these variants. The results on the first row show dropping of quality even in the case of near-front input face. The results on the second row demonstrate that our elaborate face attention discriminators have a notable performance in perceiving local texture. Detailed recognition performance is reported in Table 3.

5. Conclusion

In this paper, we propose a novel Face Normalization Model (FNM) for unsupervised face normalization in condition of unconstrained environment. FNM uses a face expert network to produce face identity features and preserve identity, which decomposes the task of generator to employ high-level semantic feature instead of image. A pixel-wise loss is introduced by a novel way for stabilizing training optimization and high quality result. Face attention mechanism helps to refine the local texture effectively. The advantage of FNM is that it can be easily incorporated into off-the-shelf face recognition framework as a pre-processing procedure. Extensive quantitative and qualitative results validate the superiority of FNM on visual applications and boosting recognition performance of face recognition.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 61573068, 61871052, 61471048, and 61375031, and by the Beijing Nova Program under Grant No. Z161100004916088, DiDi GAIA Research Collaboration Initiative.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan.

[3] K. Cao, Y. Rong, C. Li, X. Tang, and C. L. Chen. Pose-robust face recognition via deep residual equivariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. 2017.

[5] J. C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *Applications of Computer Vision*, pages 1–9, 2016.

[6] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3395, 2017.

[7] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. pages 1486–1494, 2015.

[8] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144–152, 2017.

[9] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

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

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. 2017.

[13] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. pages 4295–4304, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, 2007.

[16] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, pages 2458–2467, 2017.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015.

[19] S. Li and W. Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[20] T. Luan, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Computer Vision and Pattern Recognition*, pages 1283–1292, 2017.

[21] S. Ma, J. Fu, C. W. Chen, and T. Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks (with supplementary materials). *arXiv preprint arXiv:1802.06454*, 2018.

[22] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4838–4846, 2016.

[23] M. Mirza and S. Osindero. Conditional generative adversarial nets. *Computer Science*, pages 2672–2680, 2014.

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[25] M. Wang and W. Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[26] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics & Security*, PP(99):1–1, 2015.

[27] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T. K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *IEEE International Conference on Computer Vision*, pages 3667–3675, 2016.

[28] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[29] B. Y. R. H. Z. S. Yibo Hu, Xiang Wu. Pose-guided photorealistic face rotation. In *CVPR*, 2018.

[30] J. Yim, H. Jung, B. I. Yoo, and C. Choi. Rotating your face using multi-task deep neural network. In *Computer Vision and Pattern Recognition*, pages 676–684, 2015.

[31] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[33] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, and J. Xing. Towards pose invariant face recognition in the wild. In *CVPR*, 2018.

[34] X. Zhu, Z. Lei, J. Yan, Y. Dong, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Computer Vision and Pattern Recognition*, pages 787–796, 2015.