High-Fidelity Illumination Normalization for Face Recognition Based on Auto-Encoder

CHUNLU LI, FEIPENG DA, AND CHENXING WANG

Abstract Nonuniform illumination is one of the main issues that hinder the accuracy of face recognition because it makes the intra-person variation more complicated. To minimize the intra-person differences caused by varying illuminations, this paper presents a normalization method based on Convolutional Auto-encoder (CAE). The CAE is employed to map the face images under various illumination conditions to a normalized one, generating preliminary results with blurry and insufficient facial details, which are tricky for recognition. To recover these details, a restoration method based on re-blurring strategy and frequency analysis is proposed to preserve the facial features lying in high-frequency components based on discrete cosine transform (DCT). Therefore, in our method, these components are extracted and re-introduced into the outputs of CAE to enhance the fidelity of outputs. Thus, the facial details are preserved to the largest degree and the following works such as recognition tasks are benefited. Experiments conducted on the AR, extended Yale B, and CAS-PEAL database demonstrate the effectiveness of our method.

Index Terms Illumination normalization, face recognition, convolutional auto encoder.

I. INTRODUCTION

Face recognition has the potential to be widely applied in access control, identity authentication, watch-list surveillance etc. However, the uncontrolled illumination condition poses an obstacle to its robustness [1], because the intra-person variations caused by varying illumination conditions can be more complicated than inter-person variations. For example, the shadow cast over faces varies drastically according to the direction and intensity of lighting, which degrades the accuracy of face recognition. Therefore, it is crucial to conduct illumination normalization before recognition.

Over the last decades, many algorithms have been developed for illumination normalization. From an earlier time, holistic normalization methods have drawn great attention. Histogram equalization (HE) [2] and Histogram match (HM) manage to deal with less complicated illumination problems, by altering the pixel values and adjusting the intensity histogram of a gray-scale image. The gamma intensity correction (GIC) [3] and the logarithm transform (LT) focus on the overall brightness and achieve similar results. However, such holistic methods can only cope with simple illumination variations.

Later, numerous methods were proposed to model the illumination variation. Georghiades et al. [4] reconstruct the shape and albedo of a face using a small number of training samples taken under different lighting directions. This is based on the observation that for a certain identity in fixed pose the face images under all possible illumination conditions form a convex cone in the image space. In [5], illumination variation is modeled by low-dimensional linear spaces and the linear subspaces spanned by the corresponding images is a good approximation to the illumination cone. The main drawback of this kind of methods is that their accuracy relies heavily on precise face alignment to obtain samples under the same pose and expression. Besides, the cost of preparing and collecting face images covering various lighting conditions is quite high and their accuracy relies heavily on precise face alignment to obtain samples under the same pose and expression.

Extracting features invariant to illumination has also been studied extensively. For example, Gabor wavelets [6] is designed to simulate the receptive fields of striate neurons and thus can obtain illumination-robust features. Local
binary patterns (LBP) [7] and Gradientfaces [8] are also proposed as descriptors to extract the illumination invariant features. Many other methods attempt to conduct illumination normalization in frequency domain [9, 10] because the low-frequency component (LF) is highly related to the illumination variations while the high-frequency component (HF) contains the intrinsic features of an image, according to the Lambertian reflectance model [11]. Although reasonable results are reported, the features extracted with these methods cannot properly tackle the illumination normalization under extreme lighting conditions and may ignore useful cues, therefore these features cannot meet the rigorous demand of face recognition.

Recently, deep neural networks (DNNs) have been implemented to conduct illumination normalization. In [12], the local pattern extraction layer and the illumination elimination layer are designed and integrated into a Convolutional Neural Network (CNN) to obtain illumination invariant feature maps. Wu et al. [13] devise a multi-task DNN in order to complete the tasks of normalization and reconstruction. Generative Adversarial Network (GAN) with four types of loss function is utilized in [14] to generate images under several fixed illumination conditions. Han et al. [15] propose to enhance the output quality of a primary GAN by incorporating another GAN, which relights the normalized results. By training the two GANs interactively, the output is premium in quality. However, these methods [14]–[16] usually rely on carefully-designed network architectures or sophisticated loss functions, thus increasing computational cost. Besides, labels indicating lighting conditions are required in their training process.

The Auto-encoder (AE) and its upgraded version, Convolutional Auto-encoder (CAE), have been widely used to tackle a wide range of face-related tasks, including pose alignment [17], [18], 3D face reconstruction [19], de-occlusion [20] etc. Considering the success of Convolutional Neural Networks (CNNs) in various computer vision tasks, the CAE incorporates the feature extraction power of convolutional operation into AE, which improves its ability in understanding 2D image structures [21], [22]. Park et al. [23] achieve good performance in low-light image enhancement by utilizing two networks, including an AE for illumination estimation and a CAE for image restoration. Another stacked sparse demosing autoencoder is employed in [24] to enhance the low-light images. Overall, the CAE is capable of transforming its input images to a constricted domain and is promising in solving the illumination problems in face recognition tasks [25].

In this paper, a CAE combined with a detail restoration method is proposed for illumination normalization. The CAE is adopted to normalize the illumination because of its capability of transforming images into a normalized domain, but its outputs often lack vital details, especially when there are great changes in light and shade. While the traditional methods based on the Distinct Cosine Translation (DCT) can preserve useful facial details invariant to illumination but rely on sophisticated parameter settings. Consequently, we propose to combine the generation power of CAE and the detail-preserving capability of traditional methods to obtain better performance. This paper illustrates how the HF of original images and the LF of CAE outputs can be adaptively extracted and combined to optimize the performance by implementing an iteratively re-blurring strategy [26], [27]. Experiments on the AR database, the CAS-PEAL dataset [28], and the Extended Yale B database validate the power of this combination. The main contributions of this paper are as follows:

- The CAE is firstly introduced to generate preliminary normalization results because of its power of reconstruction. Besides, no other auxiliary network or loss function is required. Therefore it can be easily optimized and the computational cost is reduced.
- The CAE output cannot restore facial details well when the illumination condition is complex because it only focuses on holistic reconstruction. Therefore we propose a detail restoration method to enhance the quality of its outputs, inspired by the existing normalization methods based on the DCT. The HF of the original image is extracted and combined with the LF of the generated image to achieve higher fidelity. This strategy takes the advantages of both the CAE and traditional methods, and can be extended to alleviate the quality degradation problems in other similar fields.
- A re-blurry strategy is introduced to decide the boundary between the HF and LF automatically in DCT domain where the output of CAE is taken as an reference. This strategy avoids sophisticated parameter settings and thus ensures effective and efficient detail restoration.

The rest of the paper is organized as follows: Section 2 provides a summary of the state-of-the-art methods related to our work. Section 3 describes the proposed work. Section 4 presents the results from the experimental evaluation and Section 5 concludes this paper.

II. RELATED WORK

The AE couples an encoder with a decoder to learn some certain mapping between its inputs and targets. The encoder and decoder of CAE are composed of convolutional layers and deconvolutional layers respectively. For example, Hinton et al. [25] conduct view-point transformation with AE. Tewari et al. [19] propose to learn the mapping between 2D images and their corresponding 3D Model is learnt by an expert-designed CAE.

The CAE is implemented in many other face re-rendering tasks. Hinton et al. [25] conduct view-point transformation by translating images into codes describing their pose information with the encoder before computing the transformation parameters with the decoder. In [29], the very generic CAE is employed. The neurons in the last layer of its encoder are separated into several groups, each of which learns to represent a certain type of transformation (e.g. face rotation, lighting direction etc.). Then the decoder re-renders the input.
images to different viewpoints, lighting conditions etc. For each group of neurons, a mini batch of images corresponding to changes in only a single scene variable is used for training. Wu et al. [13] convert images into recon code representing poses and illumination conditions and then reconstruct images in the frontal view and the neutral lighting condition. With the aid of 3D face models, enormous training samples are generated to optimize their network. Their works have achieved reasonable results. However, one fatal drawback is that the outputs of the CAE are often low in quality and cannot meet the rigorous demand of face recognition. Specifically, the CAE often fail to recover facial details well when the illumination condition is complex. For example, if there are both low-contrast and high-contrast regions, which constantly occur in the images under non-uniformed lighting conditions, the CAE fail to recover high-quality details.

As for the underperformance of CAE in reconstructing high-quality images, the reasons are two-folds [30]:

- **The Compression Nature of the Encoder and the Pooling Layer**: The encoder, acting as a feature extractor, converts the images into a lower-dimensional subspace, from which the subtle details may be lost. Particularly, if the network contains pooling layers, which may further discard useful details, the quality will be degraded significantly.

- **The Optimization Objection Inconsistent With Human Visual Perception**: The CAE is trained by minimizing the pixel-wise differences between the original images and the generated ones, and such holistic error is far from human visual perception. It is also tricky that to minimize such error, the local facial details may be ignored because they only account for a small proportion of images.

In terms of the first factor, Mao et al. [31] propose to pass image details from the encoder directly to the decoder. However, for transformation tasks, such as pose alignment and illumination normalization, the feature maps extracted by the encoder and that of the decoder are in different conditions (i.e., poses, lighting directions), and as a result, such algorithm cannot be implemented directly. To address the second factor, meanwhile, the GAN is designed to compensate for this drawback by introducing an extra discriminator to ensure the fidelity of images. However, the hyper-parameters are difficult to select and sophisticated loss functions and structures are needed, or else visually absurd outputs may be generated, putting burden to the training process [30]. In [14], image quality is enhanced by introducing two types of loss functions based on two image quality assessment index, i.e. the Peak Signal-to-noise ratio (PSNR) and Structural Similarity Index (SSIM). By maximizing these indexes, images of higher resolution are obtained. Han et al. [15] propose to enhance the quality of their results by attaching another GAN, GAN2, to a primary one, GAN1. The auxiliary GAN2 is trained to transform its input to a fixed lighting condition, while it is necessary to provide it with the lighting label.

By training the two GANs interactively, the whole network is optimized and can produce high-quality results. These methods have achieved reasonable results, while introducing additional computation and auxiliary architecture.

As illustrated above, in frequency domain, the LF corresponds to illumination variations and the HF contains facial details. Therefore, many existing methods based on frequency analysis preserve the HF while discarding or normalizing the LF and their capability of maintaining image quality is considerable. For example, the images are transformed into logarithm domain to disentangle the illumination component and the reflectance and then DCT is performed independently on the $8 \times 8$ non-overlapping blocks. The LF is discarded to achieve good normalization performance. In [9], images are preprocessed using HE, Contrast Limiting Adaptive Histogram Equalization (CLAHE) and logarithm transformation and in DCT domain, the LF are multiplied by an exponential function over the relative location of the coefficients. These methods all preserve crucial details and yield competitive results. However, many of these methods separate the LF and HF manually and there is no efficient way to automatically locate the exact boundary between the HF and LF in frequency domain. In time domain, the logarithmic total variation (LTV) [10] is proposed to separate large- and small-scale feature corresponding to the LF and HF respectively, but the calculation is relatively complex. Therefore, it is natural to combine the LF of the outputs of CAE and the HF of the original images to enhance image quality.

In this paper, we modify the re-blurry strategy to extract the HF and LF. By comparing the blurred original images with the outputs of CAE, the boundary between the HF and LF is located, making it convenient to conduct combination of the HF and LF of the original images and CAE outputs respectively. The re-blurry strategy performs well in image quality assessment [32], [33] and de-blurring [34], [35]. Taking the corresponding re-blurred restored results as references to the original images, parameters in these algorithms are fine-tuned and optimized results are achieved.

### III. PROPOSED METHODOLOGY

The CAE is an efficient network and has been applied to many face-related tasks. In this case, it translates images under arbitrary lighting conditions into a uniform one. The 3D aided sample augmentation is utilized to boost the performance. However, the outputs of the CAE are relatively poor in quality. Hence, the time-frequency analysis is exploited to restore vital details. In this section, 3.1 introduces the theory and the defect of the CAE and 3.2 illustrates the proposed method. In 3.3, the implementation details are depicted and 3.4 provides the discussion on our method.

#### A. CONVOLUTIONAL AUTO ENCODER

Fig. 1 provides the basic structure of the CAE. This network contains an encoder and a decoder. The former extracts robust features from the original image and then the decoder reconstructs the corresponding images from these features.
Traditionally, the training error is measured by the pixel-wise differences between its input and output so as to ensure the completeness of the extracted features and the accuracy of the reconstruction process.

In our case, however, the aim is to minimize the distance between the output image and the corresponding target image under the uniform illumination condition. As a result, we take the images under arbitrary illumination conditions as inputs. The corresponding images under the frontal lighting condition and the same pose and expression, meanwhile, are regarded as expected results as in [29]. The reconstruction loss is defined as the pixel-wise difference between them, which can be represented as:

$$\text{Loss} = \frac{1}{2} \left\| I_{\text{norm}} - I_{\text{CAE}} \right\|_1$$

where $I_{\text{norm}}$ indicates the image with frontal uniform lighting and $I_{\text{CAE}}$ represents the output of the CAE. The training pairs can be obtained with the aid of 3D models and the Cook-Torrance re-rendering model [36].

Although the images can be normalized to some extent, it is difficult to restore high-quality facial details if the lighting condition is complex. For example, when the lighting is too strong or too weak, the facial regions are prone to be either overexposed or dark. Both circumstances would lead to low contrast in these regions. The CNNs, however, treat every receptive field as the same, no matter how the contrast is, so they cannot effectively extract the features of these regions. Though theoretically increasing the quantity of kernels would bring some benefits, the expansion of the network would inevitably lead to over-fitting. Consequently, the regions with low contrast cannot be restored effectively.

According to the previous researches on image analysis based on frequency domain, the facial details all lies in the HF while the illumination variation mainly affects the LF. Hence, it is reasonable to deduce that the LF, which is affected by the nonuniform illumination, of the original image can be normalized properly by the CAE, although some details lying in the HF are lost. In comparison, for the traditional illumination normalization methods, it is difficult to eliminate the effect of illumination in an effective, efficient and adaptive manner, whereas it is less complicated to extract the HF. Based on this consideration, we propose to compensate for the detail loss of the CAE by re-introducing the HF of the original images as in the traditional methods. By taking the advantages of both the modern networks and the traditional image processing methods, the results of our method are optimized and are ideal for recognition.

**B. FREQUENCY ANALYSIS FOR DETAIL ENHANCEMENT**

To restore the missing details, we propose to integrate the HF from the original image to the output of CAE. To extract the illumination-invariant component lying in HF, previous works either directly separate HF and LF in frequency domain or divide the images into large- and small-scale features in time domain. The former mostly set the boundary between HF and LF to constant values, ignoring the diversity of the frequency distribution of local details. The latter, meanwhile, rely on statistical indicators which cannot represent facial features properly.

In this article, we propose to estimate the boundary between them by analyzing the differences between the DCT coefficients of the original image and that of its filtered version which is most similar in image quality to the output of the CAE. Compared with the mentioned methods [10], our method can adaptively locate the boundary without complex parameter settings.

In our method, the 2D DCT is applied to transfer the $M \times N$ image into coefficient matrix in frequency domain. As shown in Fig. 2, the DC component, which lies in the first column first row of the matrix, represents the proportional average of the image. The remaining part denotes the AC coefficients, whose frequency increases following the direction of the arrows in Fig. 2.

As mentioned above, the CAE is capable of normalizing the LF of its input, where the illumination variation mainly lies, while it also deteriorates the details, which corresponds to the HF of the input image. As a result, the CAE output contains normalized LF and middle-frequency components.
To achieve ideal performance, it is vital to retrieve the lost HF. Therefore, we propose an iterative re-blurring strategy combined with frequency analysis. As illustrated in Fig. 3, we first utilize the re-blurring strategy to discard the HF of the input image gradually, in order to find out the very filtered sample that is most similar to the CAE output in quality. In the frequency domain, this sample can be regarded as the combination of the LF and middle-frequency component of the original image, and it lacks the HF to the same degree as the CAE output. Hence, this sample and the input image are both transferred into the DCT domain and the former is taken as reference for the latter to locate the HF, thus the expected HF of the original image can be extracted and integrated to the CAE output to recover the lost details.

1) ITERATIVE RE-BLURRING STRATEGY
In order to find out the most similar sample $I_{sim}$ to the output of the CAE, $I_{CAE}$, in quality, the original image, $I_{org}$, is first processed with the same Gaussian low-pass filter for 1 to $M$ times, respectively. As the filtering times increases, the ambiguity of the sample rises as well. Then each of the filtered samples is compared with $I_{CAE}$ in quality. Among these samples, there exists one sample $I_{sim}$, which is most similar to $I_{CAE}$ in quality because discarding either too much or too little information will lead to changes in quality. The parameter $M$ is set to a constant. An alternative is to stop filtering once the quality similarity starts to decrease, because it indicates that the filtered sample contains less details than $I_{CAE}$ and the quality dissimilarity would increase if the filtering process continues. The selected $I_{sim}$ lacks details to the same extent as $I_{CAE}$, therefore is taken as the reference to locate the HF.

It is also worth noting that although the ideal low-pass filter can also be employed to conduct the blurring operation, the Gaussian low-pass filter is preferred in practice because it performs better in de-noising.

2) ASSESSMENT OF IMAGE QUALITY
The image quality resemblance is measured with the SSIM index based on local region [37], thus the illumination variation, which mainly affects large-scale features, would barely have effect on its accuracy. Consequently, when assessing the image quality resemblance, the illumination variation can be ignored. The Structural Similarity Index Measure (SSIM) [37] is employed to evaluate the similarity between $I_b$ and $I_g$. This metric is a combination of three factors concerning image distortion, namely the loss of correlation $l$, luminance distortion $c$ and contrast distortion $s$.

$$SSIM(I_b, I_g) = l(I_b, I_g)c(I_b, I_g)s(I_b, I_g)$$

3) LOCATING THE BOUNDARY BETWEEN THE HF AND LF
To locate the boundary between LF and HF, we compare the DCT coefficient matrices of the original image and the manually blurred sample which is most similar to it in image quality. The coefficient matrices are converted into vectors by mapping $(u, v) \mapsto w$ following the direction of the arrows in Fig. 2. The difference between them is measured by the variation between the coefficients, expressed as:

$$D(w) = \frac{|c_{filtered}(w) - c_{org}(w)|}{|c_{org}(w)|}$$
where $c_{\text{filtered}}(w)$ indicates the coefficient of the filtered signal at frequency $w$, and $c_{\text{org}}(w)$ is that of the original signal. The boundary, therefore, is defined as the mutation point (i.e. frequency at which the coefficient variation rises most rapidly) of the coefficient variation, represented as:

$$w_b = \alpha \arg \min_w (D(w))$$

The component that is located in frequency band $[w_b, \infty]$ is regarded as HF, and the rest is defined as LF, i.e. which is located in $[0, w_b]$. Clearly, when $\alpha$ is set to constant, $w_b$ decreases as iteration rises. It is necessary to have the constant $\alpha$ because the boundary between HF and LF is vague, and $\alpha$ controls the actual boundary in our algorithm. The higher $\alpha$ is, the less the reserved HF will be, and the lower the image quality will be, and vice versa. However, when $\alpha$ is too high, the variation caused by illumination may be re-introduced to the result image. In order to achieve the best recognition performance, $\alpha$ should be carefully selected to balance the trade-off between illumination normalization and the quality of image. Usually, our results can be optimized when $\alpha = 0.1$.

4) INTEGRATION OF THE DETAIL

After the HF of the original image is located, the components in frequency band $[w_b, \infty]$ of it is integrated with the CAE output to optimize the result. With the HF of the original image, the details are preserved and the quality of our result is enhanced. By integrating the LF of the output of the CAE, the adverse effect caused by illumination variation is eliminated. As a result, the restored result is both high in quality and invariant to illumination.

5) 3D-AIDED DATA AUGMENTATION

The illumination variation is inextricably related to 3D geometry features of faces, so it is beneficial to generate aligned samples for network training with 3D face models and Cook-Torrance reflectance model [36]. According to this model, the intensity of pixel $I(x, y)$ can be decomposed into the surface reflection coefficient $R(x, y)$ and the illumination coefficient $L(x, y)$. The Cook-Torrance model also enables us to obtain $L'(x, y)$ under arbitrary illumination conditions. Then the corresponding generated images $I'(x, y)$ are fed into the network and the original images are regarded as ‘ground truth’ to optimize the whole network. Some of the generated training samples are provided in Fig. 4.

$$I'(x, y) = R(x, y) \cdot L'(x, y)$$

Here we reaffirm the necessity of the combination of these two process. The traditional illumination normalization methods based on DCT achieve fair results by discarding the LF directly or by replacing them with specific values. However, the contrast ratio of the face images is affected greatly by the non-uniform illumination conditions, and as a result, the distribution of DCT coefficient in different facial regions can be considerably distinct. Hence, separating HF and LF with a unified value cannot cope with complex variation in illumination properly.

By contrast, the CAE is capable of extracting facial feature invariant to illumination and then reconstructs normalized image from it. Although it is blurred to some extent, it contains normalized LF and useful middle-frequency information of the original images. The detail restoration process only makes up for the HF which is undoubtedly robust to illumination changes, so it will not degrade the normalization effect. In addition, the boundary localization process enables us to disentangle HF and LF adaptively according to the input image. Therefore, theoretically, our method yields better results than simply extracting the HF from the original images.

IV. EXPERIMENT

In this section, we first introduce the architecture and the training of our network. The impact of parameter $\alpha$ is illustrated and analyzed by conducting experiments on the AR dataset [38]. Experimental results on the AR dataset also shows the effectiveness of our method on RGB data. Experiments on the CAS-PEAL database [28] and the extended Yale B database demonstrate the superiority of our method in handling complex lighting conditions and the comparison with other state-of-the-art methods is provided.

A. NETWORK SETTINGS

1) NETWORK ARCHITECTURE

The encoder and the decoder both contain three fully convolutional layers and the kernels are all $3 \times 3$.

The memory occupied by the parameters of our network is extremely small compared with many popular networks in illumination normalization. As shown in Tab. 1, our network only needs 6M to restore its parameters. Since our network is smaller, the training process is more time-saving and our algorithm can handle illumination variations with relatively low computational cost. Note that only part of the architecture of the AJGAN [15] is provided because the whole network is not presented in their article.
2) NETWORK TRAINING

Our network is trained on the generated images mentioned in section 3 and fine-tuned in a self-supervised manner with the images under frontal lighting conditions in the AR, Extended Yale B, and the CAS-PEAL datasets. Note that the faces in the generated database are detected by the MTCNN [41]. For the Extended Yale B dataset, we crop the images and discard the hair and facial contour in order to align the generated images with the test images in this dataset.

B. EXPERIMENTS ON THE AR DATASET

In this section, we demonstrate the effectiveness of our proposed algorithm on RGB images and illustrate the impact of the parameter \( \alpha \) by conducting experiments on the AR database. The AR dataset contains 4000 pictures of 126 identities with different expressions, lighting conditions, and occlusions and this dataset is widely referred to in the pattern recognition community. For each identity, 14 images are used to verify our algorithm, since ones with occlusions are excluded because they are irrelevant to our study.

In order to illustrate the benefit of the detail restoration process and discuss the effect of the parameter \( \alpha \), SSIM and PSNR values are used to measure the quality of normalized images. These two indexes are positively associated with image quality. As shown in Tab. 2, the two indexes rise as \( \alpha \) decreases, indicating growing image quality. The reason is that the greater \( \alpha \) is, the less HF from \( I_o \), therefore the two values are smaller. This is consistent with our analysis that the HF is related to facial details, and transferring the HF from the original image brings benefits to detail enhancement.

Fig. 5 shows the enhanced results under different \( \alpha \) s. It is evident that when \( \alpha \) equals 0, the ultimate results are the same as the original images. Also, when alpha approaches infinity, the results are the output of the network. Evidently, the quality of images rises as \( \alpha \) decreases, while the effect of illumination normalization strengthens as \( \alpha \) increases. It is obvious that there is a trade-off between image quality and the normalization effectiveness.

Fig. 6 provides some randomly selected results of our methods. The first row of each block provides the original images while the second row and the third row compares the results of the CAE and the restored images. Apparently, the restored images are both premium in quality and normalized in illumination. As shown in this figure, the eye region and mouth region are enhanced after implementing our proposed detail enhancement method. It is worthy to note that this algorithm will inevitably alter the colors of skin due to the disparity between the training samples and the practical input.

Also, the Receiver Operating Characteristic (ROC) curve is employed to assess the recognition performance on the normalized results. For each normalized raw image, a nearest neighbor classifier is implemented to search for the most similar image in the rest of the whole dataset. The cosine similarity is used to measure the distance between one sample to another. Fig. 7 shows the ROC curves obtained when the parameter \( \alpha \) takes different values.

It is obvious that when \( \alpha \) is 0.05, our method under-performs because of lower image quality. As \( \alpha \) increases, the AUC of these curves rises and reaches a high of 0.3494 when \( \alpha \) equals 0.15, which is 17% more than that of the original, which indicates that the normalization process is conductive to recognition. Evidently, when \( \alpha \) is higher, HF from \( I_o \) is less, so the results of our algorithm is of lower quality and the recognition rate is degraded. When \( \alpha \) is lower, meanwhile, the effect of illumination normalization is deteriorated. In other word, the variation caused by lighting conditions is re-introduced to the images, so the recognition rate is adversely affected. It seems that there is a trade-off between illumination normalization and image quality. In the following experiments on the Extended Yale B and CAS-PEAL databases, \( \alpha \) is set to 0.1 to reach the best performance.

### TABLE 1. Memory of parameters.

| Networks                  | Memory of parameters (M) |
|---------------------------|--------------------------|
| Part of AIRGAN [15]       | 189                      |
| Pix2Pix [39]              | 118                      |
| Cycle GAN [40]            | \( \approx 81 \)         |
| Ours                      | 6                        |

### TABLE 2. Comparison of results on SSIM and PSNR with and without DCT fusion.

| Method                        | Average SSIM | Average PSNR |
|-------------------------------|--------------|--------------|
| Without detail restoration    | 0.5532       | 12.4595      |
| \( \alpha \to \infty \)       |              |              |
| \( \alpha = 0.5 \)            | 0.5806       | 12.5447      |
| \( \alpha = 0.3 \)            | 0.6454       | 12.7491      |
| \( \alpha = 0.2 \)            | 0.7234       | 12.9942      |
| \( \alpha = 0.15 \)           | 0.7790       | 13.2246      |
| \( \alpha = 0.1 \)            | 0.8236       | 13.7319      |
| \( \alpha = 0.05 \)           | 0.8882       | 16.1246      |
| Original image\( (\alpha = 0) \) | 1            | -            |
FIGURE 6. The original images, outputs of the CAE and the overall results of our proposed method, the original images lie in the first row of each block, the outputs of the CAE are in the second row and the final results are in the third row.

FIGURE 7. The ROCs of our proposed method. Each curve corresponds to a different $\alpha$.

C. EXPERIMENTS ON THE EXTENDED YALE B

In this section we compare our method with the state-of-the-art ones on the Extended Yale B dataset. This dataset contains frontal images of 38 identities under 64 illumination conditions.

This database is divided into five subsets according to the lighting incident angles (seen in Tab. 3). For face recognition, subset 1 (lighting angles less than $12^\circ$) is taken as training set and the rest ones as testing sets.

It is noteworthy that the proposed detail enhancement method is conducted on patches for the Extended Yale B dataset because the lighting condition is complex. For an input image with size of $100 \times 100$, an $50 \times 50$ patch is taken every 10 pixels and enhanced with our method. The average of the processed patches are taken as the final result. When the lighting condition is less complex, this process can be conducted on the whole image.

To optimize the recognition result, $\alpha$ is set to 0.1. The CLAHE is implemented to the training samples and the ones in the dataset so that the distribution is consistent. As for the training set, the images are converted to gray-scale and cut to be aligned with that in the Extended Yale B. Fig. 8 shows the optimized results which are randomly selected from five subsets.

To verify the effectiveness of our algorithm in face recognition, pixel-wise comparison is implemented. All of the images are resized to $100 \times 100$ and the nearest neighbor classifier with the Euclidean distance measure is employed as in [42]. We compare our results with that of some previous methods provided in [42], including Histogram Equalization (HE) [2], Multi Scale Retinex (MSR) [43], Self Quotient Image (SQI) [44], Gradient-Face (GF) [8], Weber-Face (WF) [45], Generalized Weber-Face (GWF) [46], Sparse discriminative region selection (SPR) [47], Discriminative Common hybrid Dictionary Learning (DCDL) [48], AJGAN [15], Correction of Large-scale Components (CLC) [49] and API [42] etc. Tab. 4 demonstrates that our method achieves competitive results among these algorithms.

According to Tab. 4, our recognition rates achieve comparable results to the results of existing state-of-the-art methods. The top 3 recognition rates are in bold type. DE refers to the detail enhancement method proposed in our paper.
algorithms. The possible reason is that images in this subset contain complicated local changes. For example, the shadow in the vicinity of one’s nose and eyes often causes sudden changes, leading to disturbance in HF. When the DCT fusion is conducted, it is re-introduced to the results, therefore the gap between the normalized images and the original ones is widened and the recognition rate is reduced. Another reason for this is that there is some disparity between the synthetic training samples and the realistic images. Cook-Torrance reflectance model [36] is not capable enough of simulating complicated illumination conditions, therefore the generalization power of the network is limited. To sum up, although there are some flaws of our proposed algorithm, we achieve comparable results to the state-of-the-art algorithms, demonstrating the effectiveness of our algorithm.

D. EXPERIMENTS ON THE CAS-PEAL DATASET

The lighting subset of the CAS-PEAL-R1 dataset contains 2450 images of 233 subjects under more than 9 different conditions [28]. For each identity, various numbers of images taken under more than 9 illumination conditions are provided.

Fig. 9 shows the normalized results of our method, which are randomly selected and are organized according to the azimuth of their lighting sources. The parameter $\alpha$ is set to 0.1.

It is obvious that our results are premium in quality and can handle illumination changes well. While it is inevitable that some of the images contain high-frequency noise, because the noise is included in the original images. This indicates that during the combination of the HF from original images and the LF of CAE outputs, high-frequency noise is re-introduced to the result, as the aim of our method is to recover as much details as possible.

To prove the effectiveness of our method for face recognition, features extracted from the normalized images by the pre-trained model of VGGFACE2 are used for classification. For every identity, images under frontal lighting source are taken as gallery set and the rest are used as probe set. Altogether, 1973 faces of 188 identities are detected with MTCNN and resized to 160 × 160. Some learning-based methods [15], [39], [40] are compared with our algorithm and the recognition results are shown in Tab. 5.
only provide a rough boundary between HF and LF. Therefore the parameter $\alpha$ is introduced to decide the precise boundary. To enhance the robustness of the algorithm, strategy which can adaptively compute $\alpha$ should be devised. Additionally, our method cannot handle extreme local variation because it treats every region as the same. It is estimated that integrating 3D facial data would mitigate this problem. Besides, the detail restoration method will inevitably re-introduce the noise of original images since it combine all the information in HF into the results.

Overall, as a pre-processing method, the proposed one achieves good results and is beneficial for recognition task.

ACKNOWLEDGMENT
The authors would like to thank their sincere gratitude for them.

REFERENCES

1. H. Han, S. Shan, X. Chen, and W. Gao, “A comparative study on illumination preprocessing in face recognition,” Pattern Recognit., vol. 46, no. 6, pp. 1691–1699, Jun. 2013.

2. S. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” Comput. Vis., Graph., Image Process., vol. 39, no. 3, pp. 355–368, 1987.

3. S. Han, W. Gao, B. Cao, and D. Zhao, “Illumination normalization for robust face recognition against varying lighting conditions,” in Proc. IEEE Int. SOTI Conf., Oct. 2003, pp. 157–164.

4. S.-M. Georgiades, P. N. Belhumeur, and D. J. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 643–660, Jun. 2001.

5. K.-C. Lee, J. Ho, and D. J. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 5, pp. 684–698, May 2005.

6. N. Gudur and V. Asari, “Gabor wavelet based modular PCA approach for expression and illumination invariant face recognition,” in Proc. 35th IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR), Oct. 2006, p. 13.

7. T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

8. T. Zhang, Y. Nan Tang, B. Fang, Z. Shang, and X. Liu, “Face recognition under varying illumination using gradientfaces,” IEEE Trans. Image Process., vol. 18, no. 11, 2598–2606, Nov. 2009.

9. V. P. Vishwakarma, “Illumination normalization using fuzzy filter in DCT domain for face recognition,” Int. J. Mach. Learn. Cybern., vol. 6, no. 1, pp. 17–34, Feb. 2015.

10. X. Xie, W.-S. Zheng, J. Lai, P. C. Yuen, and C. Y. Suen, “Normalisation of face illumination based on large- and small-scale features,” IEEE Trans. Image Process., vol. 20, no. 7, pp. 1807–1821, Jul. 2011.

11. D. W. Jacobs and R. Basri, “Lambertian reflectance and linear subspaces,” U.S. Patent 6,853,745, Feb. 8, 2005.

12. Y.-H. Kim, H. Kim, S.-W. Kim, H.-Y. Kim, and S.-J. Ko, “Illumination normalisation using convolutional neural network with application to face recognition,” Electron. Lett., vol. 53, no. 6, pp. 399–401, Mar. 2017.

13. Z. Wu and W. Deng, “One-shot deep neural network for pose and illumination normalisation face recognition,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2016, pp. 1–6.

14. W. Ma, X. Xie, C. Yin, and J. Lai, “Face image illumination processing based on generative adversarial nets,” in Proc. 24th Int. Conf. Pattern Recognit. (ICPR), Aug. 2018, pp. 2558–2563.

15. X. Han, H. Yang, G. Xing, and Y. Liu, “Asymmetric joint GANs for normalizing face illumination from a single image,” IEEE Trans. Multimedia, early access, Oct. 7, 2019, doi: 10.1109/TMM.2019.2945197.

16. Y. Zhang, C. Hu, and X. Lu, “IL-GAN: Illumination-invariant representation learning for single sample face recognition,” J. Vis. Commun. Image Represent., vol. 59, pp. 501–513, Feb. 2019.
M. Kan, S. Han, H. Chang, and X. Chen, “Stacked progressive auto-encoders (SPAE) for face recognition across poses,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1883–1890.

J. Zhang and H. Hu, “Exemplar-based cascaded stacked auto-encoder networks for robust face alignment,” Comput. Vis. Image Understanding., vol. 171, pp. 95–103, Jun. 2018.

A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, “MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, Oct. 2017, pp. 1274–1283.

F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, “Robust LSTM-autoencoders for face de-occlusion in the wild,” IEEE Trans. Image Process., vol. 27, no. 2, pp. 778–790, Feb. 2018.

J. Xu, M. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” in Proc. Int. Conf. Artif. Neural Netw., Espoo, Finland. Cham, Switzerland: Springer, 2011, pp. 52–59.

M. Suganuma, M. Ozay, and T. Okatani, “Exploiting the potential of standard convolutional autoencoders for image restoration by evolutionary search,” 2018, arXiv:1803.00370. [Online]. Available: http://arxiv.org/abs/1803.00370

S. Park, S. Yu, M. Kim, K. Park, and J. Paik, “Dual autoencoder network for retinaex-based low-light image enhancement,” IEEE Access, vol. 6, pp. 22084–22093, 2018.

K. G. Lore, A. Akintayo, and S. Sarkar, “LLNet: A deep autoencoder approach to natural low-light image enhancement,” Pattern Recognit., vol. 61, pp. 650–662, Jan. 2017.

G. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming auto-encoders,” in Proc. Int. Conf. Artif. Neural Netw., Espoo, Finland. Cham, Switzerland: Springer, 2011, pp. 44–51.

T. Zhang, G. Su, C. Qing, X. Xu, B. Cai, and X. Xing, “Hierarchical lifelong learning by sharing representations and integrating hypothesis,” IEEE Trans. Syst., Man, Cybern., Syst., early access, Feb. 27, 2019, doi: 10.1109/TSMC.2018.2884996.

T. Zhang, C. L. P. Chen, L. Chen, X. Xu, and B. Hu, “Design of highly nonlinear substitution boxes based on I-Ching operators,” IEEE Trans. Cybern., vol. 48, no. 12, pp. 3349–3358, Dec. 2018.

W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, “The CAS-PEAL large scale Chinese face database and evaluation protocols,” Joint Res. Develop. Lab., CAS, Columbus, OH, USA, Tech. Rep. JDL_TR_04_FR_001, 2004.

T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, “Deep convolutional inverse graphics network,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 2539–2547.

S. H. Khan, M. Hayat, and N. Barnes, “Adversarial training of variational auto-encoders for high fidelity image generation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 1312–1320.

X.-J. Mao, C. Shen, and Y.-B. Yang, “Image restoration using convolutional auto-encoders with symmetric skip connections,” 2016, arXiv:1606.08921. [Online]. Available: http://arxiv.org/abs/1606.08921.

D. B. L. Bong and B. E. Khoo, “Blind image blur assessment by using valid reblur range and histogram shape difference,” Signal Process., Image Commun., vol. 29, no. 6, pp. 699–710, Jul. 2014.

T. Wang, C. Hu, S. Wu, J. Cui, L. Zhang, Y. Yang, and L. Zhou, “NRFSIM: A no-reference image blur metric based on FSIM and re-blur approach,” in Proc. IEEE Int. Conf. Inf. Autom. (ICIA), Jul. 2017, pp. 698–703.

Y. Bihari, N. Eftekhar, and M. Irani, “Non-uniform blind deblurring by reblurring,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 3286–3295.

A. Khan, “Efficient methodologies for single-image blind deconvolution and deblurring,” Ph.D. dissertation, Univ. Manchester, Manchester, U.K., 2014.

R. L. Cook and K. E. Torrance, “A reflectance model for computer graphics,” ACM Trans. Graph., vol. 1, no. 1, pp. 7–24, 1982.

A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 27th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

A. Martínez and R. Benavente, “The AR face database,” CVG, Tech. Rep. 24, 1998, vol. 24.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1125–1134.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2223–2232.

X. Yin and X. Liu, “Multi-task convolutional neural network for pose-invariant face recognition,” IEEE Trans. Image Process., vol. 27, no. 2, pp. 964–975, Feb. 2018.

J. Chen, Z. Zeng, R. Zhang, W. Wang, Y. Zheng, and K. Tian, “Adaptive illumination normalization via adaptive illumination preprocessing and modified Weber-face,” Int. J. Speech Technol., vol. 49, no. 3, pp. 872–882, Mar. 2019.

Z. Rahman, D. J. Jobson, and G. A. Woodell, “Multi-scale retinex for color image enhancement,” in Proc. 3rd IEEE Int. Conf. Image Process., vol. 3, Sep. 1996, pp. 1003–1006.

H. Wang, S. Z. Li, Y. Wang, and J. Zhang, “Self-quotient image for face recognition,” in Proc. Int. Conf. Image Process. (ICIP), vol. 2, 2004, pp. 1397–1400.

B. Wang, W. Li, W. Yang, and Q. Liao, “Illumination normalization based on Weber’s law with application to face recognition,” IEEE Signal Process. Lett., vol. 18, no. 8, pp. 462–465, Aug. 2011.

Y. Wu, Y. Jiang, Y. Zhou, W. Li, Z. Lu, and Q. Liao, “Generalized Weber-face for illumination-robust face recognition,” Neurocomputing, vol. 136, pp. 262–267, Jul. 2014.

M. Kim, “Sparse discriminative region selection algorithm for face recognition,” Int. J. Speech Technol., vol. 42, no. 4, pp. 817–828, Jun. 2015.

T. Luo, Y. Yang, D. Yi, and J. Ye, “Robust discriminative feature learning with calibrated data reconstruction and sparse low-rank model,” Appl. Intell., pp. 1–14, 2017.

X. Tu, J. Gao, M. Xie, J. Qi, and Z. Ma, “Illumination normalization based on correction of large-scale components for face recognition,” Neurocomputing, vol. 266, pp. 465–476, Nov. 2017.

T. Matsukawa, T. Okabe, and Y. Sato, “Illumination normalization of face images with cast shadows,” in Proc. 21st Int. Conf. Pattern Recognit. (ICPR), Nov. 2012, pp. 1848–1851.

CHUNLI LI is currently pursuing the Ph.D. degree with the School of Automation, Southeast University, China. She is supervised by Prof. Da and has been studying face recognition and 3D face reconstruction.

FEIPENG DA received the Ph.D. degree in 1998. He is currently a Professor with the School of Automation, Southeast University. He has published an academic monograph, and authored or coauthored over 150 high-quality articles, many of which are retrieved by SCI, EI, and ISTP more than 100 times. He has 40 authorized invention patents, one authorized patent for utility models, four software copyrights, and three international invention patents (PCT applied). He also serves as a Reviewer for the journals from different areas, such as Optics Express, Optics Letters, Optical and Lasers in Engineering, the IEEE TRANSACTIONS on NEURAL NETWORKS, the IEEE TRANSACTIONS on CIRCUITS and SYSTEMS—I, the IEEE TRANSACTIONS on CIRCUITS and SYSTEMS—II, Physics Letters A, Neural Networks, and Pattern Recognition.

CHENXING WANG received the Ph.D. degree in 2013. She was appointed as a Research Fellow at the Multi-platform Game Innovation Center, Nanyang Technological University, Singapore, in 2014. She is currently an Associate Professor with the School of Automation, Southeast University, China. She also serves as the Reviewer of many leading journals, such as Optics Express, Optics Letters, Applied Optic, Optics and Lasers in Engineering, IEEE ACCESS, and IEEE SIGNAL PROCESSING LETTERS. She is a member of the Society of Photo-Optical Instrumentation Engineers (SPIE).