Face illumination processing via dense feature maps and multiple receptive fields

Shenggui Ling, Keren Fu, Ye Lin, Di You, and Peng Cheng

1 The National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, Sichuan, China
2 The College of Computer Science, Sichuan University, Chengdu, Sichuan, China
3 The School of Aeronautics and Astronautics, Sichuan University, Chengdu, Sichuan, China
4 Neijiang Vocational & Technical College, Neijiang, Sichuan, China

Correspondence
Peng Cheng, The School of Aeronautics and Astronautics, Sichuan University, Chengdu, Sichuan 610065, China.
Email: chengpeng_scu@163.com

Introduction: The performance of numerous computer vision tasks such as face recognition [1, 2], expression recognition [3] and so on reduces significantly when the images encounter severe illumination. Therefore, illumination processing of diverse images under a variety of illumination conditions is highly desired. In this letter, we mainly focus on the illumination processing of facial images.

In the past several decades, experts around the world have put forward various methods to process illumination. Though these traditional methods have promoted the development of illumination processing, their image quality is unsatisfactory and identification accuracy is also required to be improved further. Inspired by the success of deep convolutional neural networks (DCNNs) [4] in computer vision tasks such as image classification [5, 6], object detection [7] and image synthesis [8], some researchers begin to use deep-learning-based techniques to process face illumination. Ma et al. [9] processes illumination of facial images by using multi-stage feature maps loss that computes the face image under extreme illumination conditions. For these reasons, an elaborately-designed architecture based on convolutional neural network and generative adversarial networks for processing face illumination is presented. A novel dense feature maps loss that computes the loss by using thevarisized feature maps extracted from different convolutional layers of pre-trained feature network is put forward. Moreover, multiple-receptive-fields-based generator that uses multiple encoders during encoding phase is also proposed, and these encoders have the same structure with different kernel size. A variety of experimental results demonstrate that the method is superior to the state-of-the-art methods under various illumination challenges. Code will be available soon at https://github.com/ling20cn/IP-GAN

Method: In this letter, the x means poor-lighted faces and Y denotes standard illumination faces. Given sets xi ∈ X, yj ∈ Y, i identifies identity and j means light type. We hope the synthesized face images G(xi) and the corresponding standard illumination yj have the same identity i, namely, H(G(xi)) = H(yj). H is the feature extractor such as ResNet-50 [19], LightCNN-9 [20] and LightCNN-29v1 [20]. Next, we denote xi, yj with x, y for short.

Framework: As illustrated in Figure 1, our method mainly includes four parts: encoding network, feature fusion network, decoding network and loss network. In order to expand the receptive field of the encoding network, multiple encoders are used. Although we only draw two encoders on the chart, we can use multiple encoders during the encoding phase according to actual performance. The ellipsis means encoder. The sum-up convolution loss formula is as follows:

\[
    f_{mc} = \frac{1}{N} \sum_{w=1}^{W} \sum_{h=1}^{H} (f_{mc}^{11} + f_{mc}^{12} + \ldots + f_{mc}^{Ni})
\]

where \( f_{mc} \in \mathbb{R}^{T \times C \times W \times H} \) is a tensor (feature map) including T images that have C channels, W is width and H is height. \( f_{mc}^{11} \) means the feature maps outputted by the i-th layer of encoder 1. N means the number of encoder.

The architectures of generator and discriminator: As shown in Figure 2, the generator is made up of 3 parts: encoding network, feature fusion network and decoding network. The encoding network contains an encoder with kernal size = 3, an encoder with kernal size = 5. These encoders have the same architecture except for having different kernel size, which is called multiple receptive fields. This unique design expands the receptive field of the encoding network. After encoding, the feature fusion network fuses the encoding results obtained from the multi-encoders. The feature fusion network includes a
convolution layer and four residual blocks. The decoding network contains four up-sampling layers and six convolution layers. Each up-sampling layer is appended a convolution layer. After four up-sampling operations, the feature maps and the original input have the same size in width and height. The last two convolution layers do not enlarge their feature maps size in horizontal and vertical directions.

The discriminator of our method is inspired by the components of Pix2Pix [17]. The ReLU is used as activation after each convolution layer and we replace BatchNorm with InstanceNorm [21]. The input size of the discriminator is designed to be a $128 \times 128$ 3-channels image.

### Objective function:

**Adversarial loss:** The adversarial process is formed by the training procedure of the generator $G$ and the discriminator $D$. $G$ strives to generate lifelike fake image $G(x)$ to fool the discriminator $D$, whereas the latter attempts to discriminate the generated fake image $G(x)$ and the ground truth image $y$. The adversarial loss is as follows:

$$L_{Adv}(G, D, x, y) = E_{y \sim P_y} [\log D(y)] + E_{x \sim P_x} [\log (1 - D(G(x)))] .$$

(2)

in the above equation, $x$ is the input image (poor-lighted face), whereas $y$ is standard illumination face (well-lighted face).

**Dense feature maps loss:** If two images are very similar in structure and texture, their feature maps obtained from each convolution layer of feature extraction network are also similar. Inspired by this, we propose the dense feature maps loss. Because the higher layers of feature extraction network cause colour distortion of the generated image. Therefore, we use the lower and middle convolution layers of VGG-16 [6] to obtain feature maps and then use these maps to compute dense feature maps (DFM) loss. The former loss can be computed by:

$$L_{Dense}(G, x, y) = \sum_{r=1}^{5} ||F_i(y) - F_i(G(x))||^2_2$$

$$= \sum_{r=1}^{5} ||F_i(y) - F_i(y')||^2_2$$

(3)

In the above equation, $F$ means the VGG-16 [6] network that contains 13 convolution layers. $F_i$ denotes the first convolution layer and the subsequent activation layer of VGG-16. Likewise, we can get $F_2, F_3, F_4$ and $F_5$. $y$ means the output result of our generator and $y$ is the ground truth, namely, face image with standard illumination.

### Experiment results:

In this section, we compare the state-of-the-art methods such as Ling et al. [14], CycleGAN [15], EDIT [16], Pix2Pix [17] with ours. Moreover, we conduct a lot of objective evaluations by using SSIM [22], MS-SSIM [23], FSIM [24] and identification accuracy. The higher SSIM, MS-SSIM and FSIM values are, the better image quality is. We use ResNet-50 [19], LightCNN-9 [20] and LightCNN-29v1 [20] to extract face feature.

The Extended YaleB that has 38 subjects under 64 illumination conditions is widely used to evaluate the performance of various illumination processing methods. According to the literature [25] and [26], 1 to 28 (1792 images) are used to train all the deep-learning-based approaches, 29 to 38 (630 images) are used for testing. Next, we call it YaleB for short.

**Comparisons of illumination processing methods:** Figure 3 shows some face images after processing illumination by five methods on the YaleB database. It is noticeable that the illumination of the original varies greatly. From the second column of Figure 3, we can learn that the face images generated by CycleGAN are unnatural and noisy. In the third column, we can see that these face images are not natural and have many highlight areas. Though the images in the fourth column are natural and clear, their identities do not preserve effectively and some faces have been distorted. In the fifth and sixth column of Figure 3, although our method and Ling et al. seem to have similar visual effects, but Ling et al. has some colourful artefacts above the eyebrows after enlarging these images. Moreover, the right eye is darker than the left eye in the third face image of Ling et al.

From the Table 1, we can learn that all the methods improve the image quality after processing the face illumination. The proposed method obtains higher image quality than other state-of-the-art methods.

In the Table 2, we can learn that all the methods improve the identification accuracy after processing the face illumination except for Pix2Pix and EDIT. The main reason is that Pix2Pix and EDIT cannot preserve identities well and some faces are distorted when they meet extreme illumination. From the final row of Table 2, we can see that the identification accuracy reaches to 96.67% after processing illumination by our method then using Light-CNN-29v1 to identify.

**Ablation study:** As shown in the Table 3, we can know that our DFM loss obtains higher identification accuracy than other losses when we use ResNet-50, LightCNN-9 and LightCNN-29v1 to extract feature for face identification. In the Table 4, we illustrate the identification accuracy of our model trained under different receptive field and one encoder with atrous spatial pyramid pooling (ASPP) [29] that is used to expand receptive field. From the fifth row of Table 4, comparing to one encoder, we can see that the identification accuracy is improved by 3.33%, 2.38% and 1.43% when we use two encoders to train our model and then use ResNet-50, LightCNN-9 and LightCNN-29v1 to extract feature for identification.

| Method      | SSIM ↑ | MS-SSIM ↑ | FSIM ↑ |
|-------------|--------|-----------|--------|
| Original    | 0.3569 | 0.3683    | 0.7410 |
| CycleGAN    | 0.5849 | 0.6403    | 0.8318 |
| EDIT        | 0.5656 | 0.6713    | 0.8101 |
| Pix2Pix     | 0.6031 | 0.6925    | 0.8316 |
| Ling et al. | 0.6946 | 0.7773    | 0.8660 |
| Ours        | 0.7218 | 0.8057    | 0.8731 |

| Method      | ResNet-50 ↑ | LightCNN-9 ↑ | LightCNN-29v1 ↑ |
|-------------|-------------|--------------|-----------------|
| Original    | 0.5126      | 0.7302       | 0.8006          |
| CycleGAN    | 0.6920      | 0.7793       | 0.8111          |
| EDIT        | 0.5032      | 0.6302       | 0.7159          |
| Pix2Pix     | 0.1889      | 0.2016       | 0.1048          |
| Ling et al. | 0.8952      | 0.9016       | 0.9429          |
| Ours        | 0.9270      | 0.9143       | 0.9667          |
Table 3. The top-1 identification accuracy of the original faces and the faces synthesized by our model trained with six losses

| Method          | ResNet-50↑ | LightCNN-9↑ | LightCNN-29v1↑ |
|-----------------|------------|-------------|----------------|
| Original        | 0.5126     | 0.7302      | 0.8064         |
| L1 loss         | 0.7141     | 0.7825      | 0.8841         |
| L2 loss         | 0.6492     | 0.8000      | 0.9286         |
| SSIM loss       | 0.6939     | 0.8714      | 0.9556         |
| Perceptual loss | 0.8159     | 0.8921      | 0.9460         |
| LPIPS loss      | 0.8651     | 0.8837      | 0.9524         |
| DFM loss        | 0.9270     | 0.9143      | 0.9667         |

Table 4. Top-1 identification accuracy of the original faces and the faces synthesized by our model trained under various receptive field

| Method          | ResNet-50↑ | LightCNN-9↑ | LightCNN-29v1↑ |
|-----------------|------------|-------------|----------------|
| Original        | 0.5126     | 0.7302      | 0.8064         |
| 1 en+ASPP [29]  | 0.8571     | 0.8413      | 0.9460         |
| 1 encoder       | 0.8937     | 0.8905      | 0.9524         |
| 2 encoders      | 0.9270     | 0.9143      | 0.9667         |
| 3 encoders      | 0.9127     | 0.9111      | 0.9556         |

Conclusion: We present a novel scheme for processing face illumination in this letter. Experimental results demonstrate that our method surpasses the state-of-the-art methods in image quality and identification accuracy. In the future, we will extend it to other image-to-image translation tasks such as photo to cartoon, photo to sketch, sketch to photo and so forth.

Acknowledgement: This work was supported by the Joint Funds of the National Natural Science Foundation of China under Grant U1833128.

© 2021 The Authors. Electronics Letters published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 6 January 2021 Accepted: 30 March 2021
doi: 10.1049/el2.12181

References
1 Lian, Z., Er, M.J.: Illumination normalisation for face recognition in transformed domain. Electron. Lett. 46(15), 1060–1061 (2010)
2 Wang, M., Deng, W.: Deep face recognition: A survey. arXiv:1804.06655, (2018)
3 Ma, H., Celik, T.: Fer-net: Facial expression recognition using densely connected convolutional network. Electron. Lett. 55(4), 184–186 (2019)
4 LeCun, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE 86(11), 2278–2324 (1998)
5 Krizhevsky, I.A., Hinton, E.: Imagenet classification with deep convolutional neural networks. In: Conference on Neural Information Processing Systems, pp. 1097–1105. NeurIPS, La Jolla, CA (2013)
6 Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, (2014)
7 Girshick, R.B. et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. IEEE, Piscataway, NJ (2014)
8 Goodfellow, I.J., et al.: Generative adversarial nets. Advances in Neural Information Processing Systems 27, 2672–2680 (2014)
9 Ma, W. et al.: Face image illumination processing based on generative adversarial nets. In: 2018 24th International Conference on Pattern Recognition, pp. 2558–2563. IEEE, Piscataway, NJ (2018)
10 Ma, W. et al.: Face image illumination processing based on gan with dual triplet loss. In: Chinese Conference on Pattern Recognition and Computer Vision, pp. 150–161. Springer, Berlin (2018)
11 Han, X., et al.: Asymmetric joint gans for normalizing face illumination from a single image. IEEE Trans. Multimedia 22(6), 1619–1633 (2019)
12 Zhang, Y., Hu, C., Lu, X.: Il-gan: Illumination-invariant representation learning for single sample face recognition. J. Visual Commun. Image Represent. 59, 501–513 (2019)
13 Ling, S. et al.: Illumination normalization of face image. In: Twelfth International Conference on Digital Image Processing, pp. 11519–11557. SPIE, Bellingham, WA (2020)
14 Ling, S., et al.: A high-performance face illumination processing method via multi-stage feature maps. Sensors 20(17), 4869 (2020)
15 Zhu, J.-Y. et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, pp. 2242–2251. IEEE, Piscataway, NJ (2017)
16 Fu, Y. et al.: Edit: Exemplar-domain aware image-to-image translation. arXiv:1911.10520, (2019)
17 Isola, P. et al.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5967–5976. IEEE, Piscataway, NJ (2017)
18 Lin, Y., et al.: An identity-preserve model for face sketch-photo synthesis. IEEE Signal Process. Lett. 27, 1095–1099 (2020)
19 He, K. et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–777. IEEE, Piscataway, NJ (2016)
20 Wu, X., et al.: A light cmn for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. 13(11), 2884–2896 (2018)
21 Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022, (2016)
22 Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612 (2004)
23 Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 1398–1402. IEEE, Piscataway, NJ (2003)
24 Zhang, L., et al.: Fsim: A feature similarity index for image quality assessment. IEEE Trans. Image Process. 20, 2378–2386 (2011)
25 Tang, Y., Salakhutdinov, R., Hinton, G.E.: Deep Lambertian networks. In: Proceedings of the 29th International Conference on Machine Learning, pp. 1419–1426. Omnipress, Madison, WI (2012)
26 Gao, S., et al.: Single sample face recognition via learning deep supervised autoencoders. IEEE Trans. Inf. Forensics Secur. 10, 2108–2118 (2015)
27 Zhang, R. et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 586–595. IEEE, Piscataway, NJ (2018)
28 Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer, Berlin (2016)
29 Chen, L.-C., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2018)