ASYMMETRIC CYCLEGAN FOR UNPAIRED NIR-TO-RGB FACE IMAGE TRANSLATION

Hao Dou[12], *Chen Chen*[12], *Xiyuan Hu*[12]*, *Silong Peng*[123]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Visystem Co.Ltd, Beijing, China

ABSTRACT

Translating near-infrared (NIR) face into color (RGB) face, is helpful to improve the visual effect of images and the performance of face recognition. The model for unpaired image-to-image translation is suitable for this task due to the high cost of pixel-matched data. Because of the complexity difference between NIR and RGB image domains, the complexity inequality in bidirectional NIR-RGB translations is significant. We analyze the limitation of the original CycleGAN in asymmetric translation tasks, and propose an Asymmetric CycleGAN model with U-net-like generators of unequal sizes to adapt to the asymmetric need in NIR-RGB translations. The edge-retain loss between NIR and the generated RGB images is also introduced to enhance face visual quality. The qualitative visual evaluation and quantitative evaluation with face ID and skin color criteria show that our model achieves great improvements compared with state-of-the-art methods on three public datasets and a newly proposed dataset.

Index Terms— NIR-to-RGB, Image-to-Image Translation, Asymmetric CycleGAN

1. INTRODUCTION

In the field of public security, near-infrared (NIR) face images are often used for face recognition [1, 2] and retrieval. Compared with RGB face images, NIR face images are not friendly to human vision and have worse performance of face recognition. Translating NIR face images into RGB face images and reconstructing facial skin color, can greatly improve the visual effect and the performance of face recognition.

Deep learning based methods can be applied to the task of NIR-to-RGB face image translation. Many pixel-level supervised models such as colorization [3, 4, 5, 6], pix2pix [7], has achieved successes in grey or NIR image colorization, translations of day-to-night and maps-to-scenes, etc. However, the NIR-RGB image pairs for human face are of great difficulty to collect, since the pixel-matched NIR-RGB face data costs more than unpaired data. Therefore, unpaired translation models are more suitable for the task of NIR-to-RGB face image translation. Based on a pair of Generative Adversarial Networks [8], CycleGAN [9] has been a popular unpaired image-to-image translation model. By introducing the cycle-consistency loss, CycleGAN can synchronously implement the bidirectional image-to-image translations. Compared to models such as DistanceGAN [10], UINT [11], MUNIT [12], CycleGAN is more robust and easier to train.

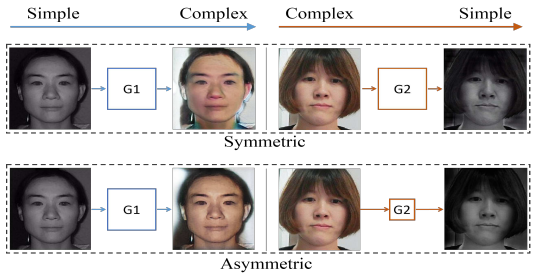

Fig. 1. An example for bidirectional unpaired translations of Symmetric CycleGAN and Asymmetric CycleGAN.

As for the unpaired image-to-image translation, when the complexities of the two domains are significantly different, the complexity of translation task with the direction from simple to complex is usually much lower than that with the other direction, and vice versa. The uneven bidirectional translation tasks might learn more or less content or texture (e.g. sketch-photo), or change color space and increase or decrease information channel (e.g. Grey-RGB). Therefore, these translation tasks with directions of significant information ascending or descending are defined as *asymmetric translation* tasks, to distinguish the translation between domains with generally even complexity (namely *symmetric translation*).

The two generators are symmetric with the same structure and size in original CycleGAN, which performs well in symmetric translation tasks, called *Symmetric CycleGAN*. When the translation tasks are asymmetric, two generators of the same size are easily trained unevenly after the similar process of optimization. Fig 1 illustrates an example of the bidirectional translations. After the sufficient training, Symmetric CycleGAN could learn a good face translation from RG-

*Corresponding author;
This work was supported by the National Nature Science Foundation of China (Grant No. 61571438).

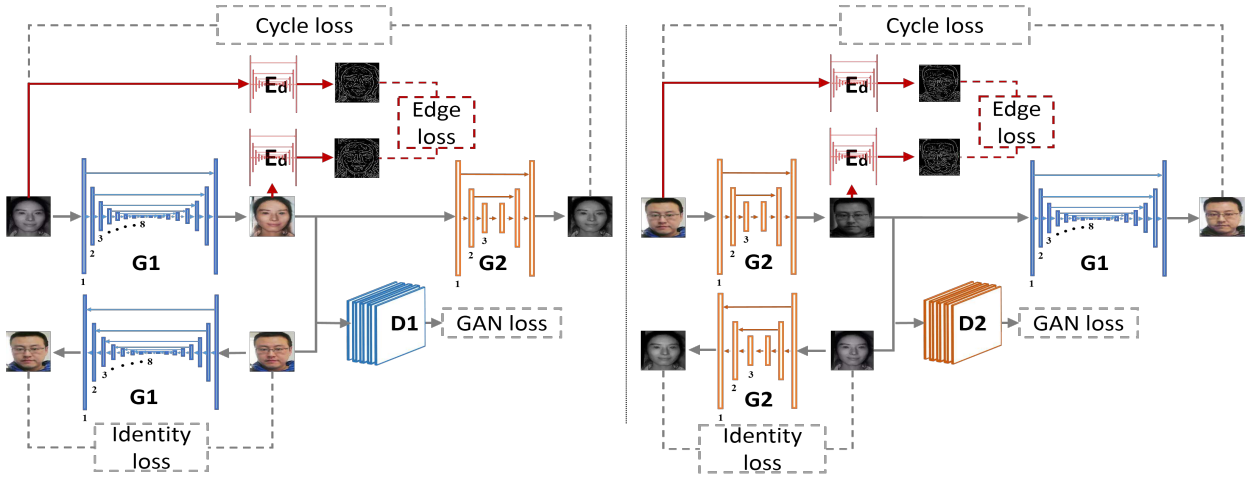

Fig. 2. The structure of Asymmetric CycleGAN model. G_1 (NIR-RGB) and G_2 (RGB-NIR) denote two generators with different size; E_d represents the network with fixed weight to extract the image edge for the proposed edge loss computation.

B to NIR (complex-to-simple), but the resulting RGB face image generated from NIR (simple-to-complex) is not well learned. Based on this, it will be reasonable to make the network complexity and translation complexity inosculate well. We can use the complex network for the simple-to-complex translation and the simple network for the complex-to-simple translation. Therefore we designed a CycleGAN with different size of generators to adapt to this asymmetric translation, called *Asymmetric CycleGAN*. In the example of Fig 1, the proposed Asymmetric CycleGAN manages to match the asymmetric translation tasks with the corresponding generators of uneven sizes, thus compared with symmetric CycleGAN, Asymmetric CycleGAN with more reasonable network settings achieves significant improvement.

Compared with outstanding performance in natural image translation tasks, for human face image applications, CycleGAN based methods have weakness in small edge remaining and limitation in generating proper facial details. Therefore, we also introduce an additional edge loss to make the generated color image retain the necessary edge details from the input image as much as possible, which can greatly improve the face image quality.

We have two main contributions in this paper. Firstly, we propose an efficient Asymmetric CycleGAN to improve the asymmetric translations between NIR and RGB face images. The idea of asymmetric setting can be easily and effectively apply to all asymmetric translation tasks. Secondly, an edge loss term is introduced to keep necessary facial details, which is helpful to enhance the fineness of generated face images.

2. PROPOSED METHOD

For asymmetric translations between NIR and RGB face images, we use networks of different depth as different generators to adapt to the asymmetric translations. An extra regularization term on image edge is introduced to improve the facial quality.

2.1. Model structure of Asymmetric CycleGAN

As shown in Fig 2, the basic structure of our model will be illustrated by the three parts, i.e. the generators, the edge detector and the discriminators.

G_1 and G_2 are constructed by different size of U-net [13]. Compared to the residual blocks [14] in original CycleGAN, U-net consists of more sampling layers and has the ability to extract more precise features. U-net enables features to be transmitted across connections, as a result, the common features of input image can be reused and generation quality can be improved. U-net includes K down-sampling operations. The maximum of down-sampling operations in U-net will be 8 when the size of input images is 256×256. Since the feature extracted by three down-sampling convolution layers can basically contain the shallow information of images [15], which is suitable for the dimension-reduced image translation process, we set K=3 for the generator G_2. To match the complexity of the dimension-ascending image translation, we set K=8 for the generator G_1. Besides, the transposed convolution[16] of U-net is changed to the combination of up-sampling and convolution, to avoid the checkerboard problem[17].

E_d is used to denote the pre-trained U-net for edge detection [18]. We obtain the edge data of images by Canny [19] operator, and then the U-net in Pix2pix model is trained with face-to-edge image pairs. During the training of our model, E_d is fixed and just extracts the edge for the edge-retain loss.

In consistent with CycleGAN, we use PatchGAN as the discriminators, i.e. D_1 and D_2 shown in Fig 2.

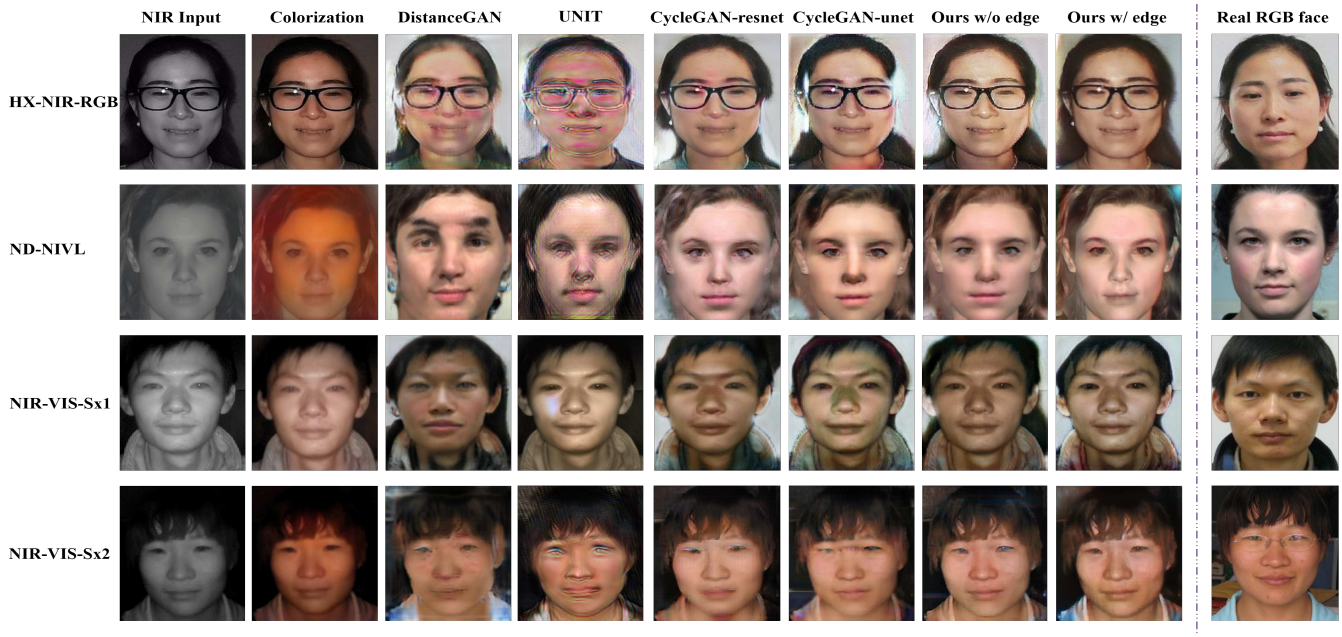

Fig. 3. Comparison of related methods on four datasets. From left to right: input, a pre-trained colorization model[5], DistanceGAN [10], UNIT, CycleGAN using residual blocks, CycleGAN using U-net, the proposed Asymmetric CycleGAN without the edge-retain loss, Asymmetric CycleGAN with the edge-retain loss, and the non pixel-matched groundtruths.

2.2. Edge-retain oriented facial loss

The loss function of our proposed model consists of the edge-retain loss and original CycleGAN loss. In the following text, X and Y are used to denote the domains of NIR face and RGB face, respectively.

Edge-retain loss In order to enhance the visual appearance especially for further face recognition process, the facial edge of input images should be maintained in the translation processes. We use the pre-trained edge detector E_d (described in section 2.1) to predict the edges of both input images and the generated images. The well-detected edges can be regard as the prior knowledge guiding better facial image generation. We compute the $L1$ distance of the edges between input images and generated images used as a regularization, named *edge-retain loss*. We formulate the edge-retain loss as,

$$L_{Edge}(G_1;G_2;X;Y) = E_{x\sim p_{data(x)}}[\|E_d(G_1(x)) - E_d(x)\|_1] + E_{y\sim p_{data(y)}}[\|E_d(G_2(y)) - E_d(y)\|_1]. \quad (1)$$

Original CycleGAN loss The CycleGAN loss includes a couple of GAN losses, cycle-consistency loss and identity loss [9, 20]. The LSGAN [21] is applied as the GAN loss. We thus formulate the original CycleGAN loss as,

$$L_{CycleGAN}(G_1;G_2;D_1;D_2;X;Y) = L_{LSGAN_1} + L_{LSGAN_2} + \lambda_{cyc}L_{Cyc} + \lambda_I L_{Identity}, \quad (2)$$

where L_{LSGAN_1} and L_{LSGAN_2} denote the LSGAN losses, L_{Cyc} denotes the cycle-consistency loss, $L_{Identity}$ denotes the identity loss, and λ_{cyc} and λ_I denote the weight of cycle-consist loss and identity loss.

The total loss The total loss is the combination of original CycleGAN loss and the proposed edge-retain loss, which can be formulated as follows:

$$L_{Total}(G_1;G_2;D_1;D_2;X;Y) = L_{CycleGAN} + \lambda_E L_{Edge}, \quad (3)$$

where λ_E is the parameter for the trade-off between the original CycleGAN loss and edge-retain loss.

3. EXPERIMENTS

We have conducted the experiments on four datasets and evaluate the performance with visual effect and objective metrics.

3.1. Datasets

NIR images are sensitive to the shooting conditions, e.g. illumination, imaging devices. In order to verify the performance for data on different conditions, we perform the proposed method on a newly proposed dataset HX-NIR-RGB, and three public datasets: ND-NIVL [22] , NIR-VIS-Sx1 and NIR-VIS-Sx2. We randomly select 75% of images as training sets and others as testing sets for each dataset.

HX-NIR-RGB We record two videos captured indoors, one

using a NIR camera and the other using a RGB camera. We select the clear frames with complete face in the videos as the images in this dataset. This dataset consists of 90 NIR images and 74 RGB images.

ND-NIVL Images in this dataset are taken with high resolutions on good illumination. After eliminating the images with incomplete face and high similarity, we utilize 1333 NIR images and 1098 RGB images.

NIR-VIS-Sx1 and NIR-VIS-Sx2 The two datasets are both from NIR-VIS-2.0 [23] dataset. We divide the NIR-VIS-2.0 into two different sub datasets according to the shooting conditions and face identities. There are 283 NIR / 283 RGB images in NIR-VIS-Sx1 while 172 NIR / 172 RGB images in NIR-VIS-Sx2.

3.2. Implementation settings

The learning policy and hyperparameters in our model are similar to the original CycleGAN. We use Adam [24] as the optimizer with the learning rate of 0.0002. The weight of the edge-retain loss is set to 2.5.

3.3. Qualitative visual evaluation

Compared with the related methods, our model has achieved the better results in terms of color and texture and reconstructed more natural face, as shown in Fig 3. Original CycleGAN have weaker performance, reflected in unnormal color on HX-NIR-RGB, superfluous texture on NIR-VIS-Sx1 and blurred edge on ND-NIVL and NIR-VIS-Sx2. Other translation methods, i.e. colorization, DistanceGAN and UNIT, also cannot generate the satisfactory results. By the means of the edge-retain loss, our model has generated clear facial edge and texture compared to the model without the edge-retain loss. The edges of input images and generated images are also displayed as follows, which illustrates the necessary edges have been retained sucessfully.

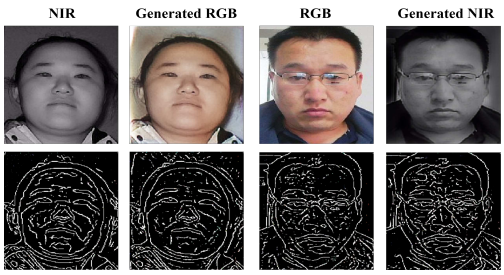

Fig. 4. Detected edges of input images and generated images.

3.4. Extended quantitative evaluation

The generated face and groundtruths are expected to have the same identification and the similar facial skin color. Therefore we apply the following two criteria for quantitative evaluations.

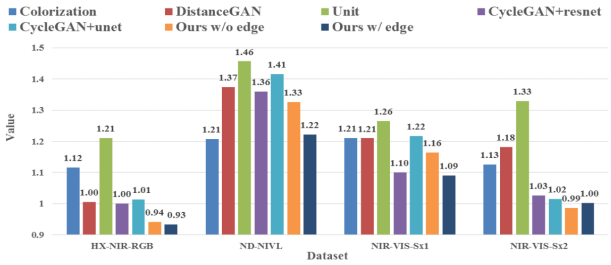

Fig. 5. Comparison on Face ID distance.

Face ID criterion A VGG [25] network pre-trained for face recognition [26] is used to extract the feature of face images. We compute the mean $L1$ distance of features between generated images and non pixel-matched groundtruths. As shown in Fig5, on HX-NIR-RGB, NIR-VIS-Sx1 and NIR-VIS-Sx2, our model obtains the minimum value for this criterion. On ND-NIVL, the value of our model with edge-retain loss is very approximate to the minimum.

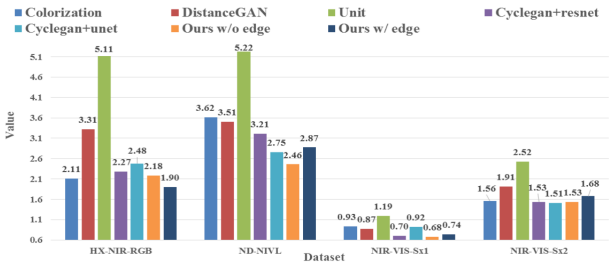

Fig. 6. Comparison on skin color difference.

Facial skin color criterion We use a facial mask to get the local facial skin, and compute the mean color difference between the generated images and non pixel-matched groundtruths in HSV space [27]. As shown in Fig 6, our model can generate the face images with the closest skin color to the groundtruths on HX-NIR-RGB and ND-NIVL, and obtain the comparable results to the CycleGAN on NIR-VIS-Sx1 and NIR-VIS-Sx2.

4. CONCLUSION

In this paper, we propose a method named Asymmetric CycleGAN to deal with the task of unpaired NIR-to-RGB face image translation. We use different size of generators to adapt to the asymmetric translations and introduce the edge-retain loss for face images to enhance the generated image quality. Our model shows the improvement and obtains good performance on this task. In further research, we will verify the performance of the proposed Asymmetric CycleGAN on general asymmetric translation tasks.

5. REFERENCES

[1] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[2] H Ran, W Xiang, S Zhenan, et al., "Wasserstein cnn: Learning invariant features for nir-vis face recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, pp. 1–1.

[3] Zezhou Cheng, Qingxiong Yang, and Bin Sheng, "Deep colorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[4] Matthias Limmer and Hendrik P. A. Lensch, "Infrared colorization using deep convolutional neural networks," in *IEEE International Conference on Machine Learning and Applications*, 2016, pp. 61–68.

[5] Gustav Larsson et al., "Learning representations for automatic colorization," in *European Conference on Computer Vision*, 2016.

[6] Matthias Limmer and Hendrik P. A. Lensch, "Improved ir-colorization using adversarial training and estuary networks," in *British Machine Vision Conference*, 2017.

[7] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *arXiv preprint*, 2017.

[8] Ian J. Goodfellow et al., "Generative adversarial nets," in *International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[9] Jun Yan Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *arXiv preprint*, 2017.

[10] Sagie Benaim and Lior Wolf, "One-sided unsupervised domain mapping," in *Advances in neural information processing systems*, 2017.

[11] Ming-Yu Liu, Breuel Thomas, and Jan Kautz., "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017.

[12] Xun Huang, Ming Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *arXiv preprint*, 2018.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014.

[16] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *International Conference on Computer Vision*, 2011, pp. 2018–2025.

[17] Augustus Odena, , et al., "Deconvolution and checkerboard artifacts," in *Distill*, 2016.

[18] E. Nadernejad et al., "Edge detection techniques: evaluations and comparisons," in *Applied Mathematical Sciences*, 2008, pp. 1507–1520.

[19] J. Canny, "A computational approach to edge detection," 1986, IEEE Computer Society.

[20] Yaniv Taigman, Adam Polyak, and Lior Wolf, "Unsupervised cross-domain image generation," in *arXiv preprint*, 2016.

[21] Xudong Mao et al., "Least squares generative adversarial networks," in *International Conference on Computer Vision*, 2014.

[22] J Bernhard, J Barr, K. W Bowyer, and P Flynn, "Near-ir to visible light face matching: Effectiveness of preprocessing options for commercial matchers," in *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2015.

[23] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao, "The casia nir-vis 2.0 face database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.

[25] Simonyan Karen and Zisserman Andrew, "Very deep convolutional networks for large-scale image recognition," in *arXiv preprint*, 2014.

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[27] Alvy Ray Smith, "Color gamut transform pairs," 1978, vol. 12, pp. 12–19.