Copy and Paste GAN: Face Hallucination from Shaded Thumbnails – Supplementary Materials

Yang Zhang[1,2,3], Ivor W. Tsang[3], Yawei Luo[4], Changhui Hu[1,2,5], Xiaobo Lu[1,2]*, Xin Yu[3,6]

[1] School of Automation, Southeast University, China

[2] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, China

[3] Centre for Artificial Intelligence, University of Technology Sydney, Australia

[4] School of Computer Science and Technology, Huazhong University of Science and Technology, China

[5] School of Automation, Nanjing University of Posts and Telecommunications, China

[6] Australian Centre for Robotic Vision, Australian National University, Australia

Abstract

In this supplementary material, we first present the details about the loss functions used in CPGAN. Then, more qualitative results on Multi-PIE (indoor) and CelebA (in the wild) datasets are performed. Meanwhile, we present the effect of our proposed illumination compensation loss on the face hallucination. Finally, we elaborate the training details of our RaIN model.

1. Loss Function

To train our CPGAN framework, we propose an illumination compensation loss (L_{ic}) together with an intensity similarity loss (L_{mse}), an identity similarity loss (L_{id}) [16], a structure similarity loss (L_h) [2] and an adversarial loss (L_{adv}) [3].

1.1. Illumination Compensation Loss:

CPGAN not only recovers UI-HR face images but also compensates for the non-uniform illumination. Inspired by the style loss in AdaIN [6], we propose the illumination compensation loss L_{ic}. The basic idea is to constrain the illumination characteristics of the reconstructed UI-HR face that is close to the guided UI-HR one in the latent subspace:

$$\begin{aligned} L_{ic} =& \mathbb{E}_{(\hat{h}_i,g_i)\sim p(\hat{h},g)}\{\sum_{j=1}^{L}\left\|\mu\left(\varphi_j(\hat{h}_i)\right)-\mu\left(\varphi_j(g_i)\right)\right\|_2 \\ &+\sum_{j=1}^{L}\left\|\sigma\left(\varphi_j(\hat{h}_i)\right)-\sigma\left(\varphi_j(g_i)\right)\right\|_2\}, \end{aligned} \tag{1}$$

where g_i represents the guided UI-HR image, $\hat{h}_i$ represents the generated UI-HR image, $p(\hat{h},g)$ represents their joint distribution. Each $\varphi_j(\cdot)$ denotes the output of relu1-1, relu2-1, relu3-1, relu4-1 layer in a pre-trained VGG-19 model [17], respectively. Here, μ and σ are the mean and variance for each feature channel.

1.2. Intensity Similarity Loss

To enforce the generated UI-HR images $\hat{h}_i$ to approximate to the ground truth (GT) images h_i in intensity level, pixel-wise Mean Square Error (MSE) loss L_{mse} is introduced:

$$\begin{aligned} L_{mse} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)}\left\|\hat{h}_i-h_i\right\|_F^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)}\left\|C_t\left(l_i\right)-h_i\right\|_F^2, \end{aligned} \tag{2}$$

where t is the parameters of the upsampling network C in CPGAN. l_i represents the input NI-LR face image. $p(\hat{h},h)$ represents the joint distribution of the generated UI-HR face $\hat{h}_i$ and the corresponding UI-HR GT image h_i, respectively. $p(l,h)$ represents the joint distribution of the input NI-LR and GT images in the training dataset. Frankly, the MSE loss leads to high peak signal-to-noise ratio (PSNR) values. However, only employing the MSE loss in feed-forward optimization is insufficient to capture the high-frequency features, resulting in overly smooth facial details.

*Corresponding author (xblu2013@126.com).

This work was done when Yang Zhang (zhangyang201703@126.com) was a visiting student at University of Technology Sydney.

1.3. Identity Similarity Loss

Identity similarity is one of the most important parts for face hallucination [16]. We adopt Resnet50 [5] network with pre-trained parameters to extract the feature maps of high-level facial features, enabling the identity preserving ability for CPGAN:

$$\begin{aligned} L_{id} &= \mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)} \left\| \Phi\left(\hat{h}_i\right) - \Phi\left(h_i\right) \right\|_F^2 \\ &= \mathbb{E}_{(l_i,h_i)\sim p(l,h)} \left\| \Phi\left(C_t\left(l_i\right)\right) - \Phi\left(h_i\right) \right\|_F^2 . \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ represents the extracted feature maps from the AveragePooling layer in Resnet50 [5].

1.4. Structure Similarity Loss

To constrain the structural consistency between the generated UI-HR image and the GT image, the structure similarity loss [19] is also introduced to our framework:

$$L_h = \mathbb{E}_{(l_i,h_i)\sim p(l,h)} \frac{1}{P} \sum_{k=1}^{P} \left\| H^k\left(h_i\right) - H^k\left(\tilde{C}_t\left(l_i\right)\right) \right\|_2^2, \tag{4}$$

where $H^k(\cdot)$ represents the heatmap corresponding to the k-th landmark. $H^k\left(\tilde{C}_t\left(l_i\right)\right)$ represents the k-th predicted heatmap, which is estimated by adopting the stacked hourglass module [14] on the intermediate upsampled features in CPGAN. $H^k\left(h_i\right)$ denotes the k-th ground-truth heatmap obtained by running a Face Alignment Network (FAN) [1] on the GT image.

1.5. Adversarial Loss

The adversarial loss is introduced to encourage the generated UI-HR face images to reside in the manifold of the GT ones. Thus, the loss function of discriminator is:

$$L_D = -\mathbb{E}_{(\hat{h}_i,h_i)\sim p(\hat{h},h)} \left[\log D_d\left(h_i\right) + \log\left(1 - D_d\left(\hat{h}_i\right)\right) \right], \tag{5}$$

where D and d represent the discriminative network and its parameters. The goal of the discriminative network is to distinguish the hallucinated UI-HR faces from the GTs.

However, the upsampling network is designed to produce realistic UI-HR images to fool the discriminative network. Thus, the corresponding adversarial loss is:

$$\begin{aligned} L_{adv} &= -\mathbb{E}_{\hat{h}_i\sim p(\hat{h})} \log\left(D_d\left(\hat{h}_i\right)\right) \\ &= -\mathbb{E}_{l_i\sim p(l)} \log\left(D_d\left(C_t\left(l_i\right)\right)\right). \end{aligned} \tag{6}$$

During training, we minimize the loss L_D and update the parameters d for the discriminative network. For training upsampling network, we minimize the loss L_{adv} to update the parameters t.

1.6. Total Loss Function

The overall loss function is the weighted sum of the above terms:

$$L_G = \alpha L_{mse} + \beta L_{id} + \gamma L_h + \zeta L_{ic} + \psi L_{adv}. \tag{7}$$

2. Experiments

Table 1. Ablation study on the training loss

		Multi-PIE		CelebA	
		PSNR	SSIM	PSNR	SSIM
w/o L_{ic}	L_G^-	21.948	0.689	21.003	0.516
	$L_G^\dagger$	22.813	0.694	21.674	0.552
	$L_G^\ddagger$	23.022	0.701	21.853	0.588
	$L_G^\star$	22.624	0.692	21.241	0.543
w/ L_{ic}	L_G^-	22.943	0.693	23.262	0.699
	$L_G^\dagger$	23.036	0.718	24.103	0.731
	$L_G^\ddagger$	25.104	0.782	24.948	0.755
	$L_G^\star$	24.639	0.778	23.972	0.723

2.1. Experiment Setting

CPGAN is trained and tested on the Multi-PIE dataset [4] (indoor) and the CelebFaces Attributes dataset (CelebA) [12] (in the wild). For the face images in Multi-PIE dataset, all face regions are cropped by the keypoint location*. For the CelebA images, the face regions are cropped by the Landmarks Annotations †. The NI-LR/UI-HR face pairs from various illumination conditions and identities are required for this framework. For each identity, the face image with uniform illumination is served as the GT image. During the training and testing processes, the external guided UI-HR images are *randomly* selected from the UI-HR ones.

Our model is implemented with Pytorch. During training, the ADAM optimizer [7] is adopted to optimize CPGAN. We set parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The learning rate is set to 10^{-3}, and multiplied by 0.99 after each epoch. The trade off weights $\alpha, \beta, \gamma, \zeta, \psi$ in overall loss function are set to 1, 10^{-3}, 10^{-2}, 1, 10^{-2}, respectively. Similar to [21], spectral normalization [13] is introduced for both our upsampling network and discriminator to stabilize the training of CPGAN. The training codes and details will be released on our website.

2.2. Additional Qualitative Results

Additional qualitative results are presented in Fig.1 (Multi-PIE) and Fig.2 (CelebA), which justify the su-

*https://github.com/HRLTY/TP-GAN

†http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

perior performance of our CPGAN over the competing methods.

2.3. Additional Quantitative Results

Table 1 reports the performance of different CPGAN variants (trained with different loss combinations) on the hallucinated results. As indicated in Table 1, our proposed illumination compensation loss (L_{ic}) also improves the quantitative results. With a slight abuse of notation, we denote:

- L_G^-: L_{mse},
- $L_G^\dagger$: L_{mse}+L_{id},
- $L_G^\ddagger$: L_{mse}+L_{id}+L_h,
- $L_G^\star$: L_{mse}+L_{id}+L_h+L_{adv}.

2.4. Comparisons with SoA on Face Recognition

We employ a state-of-the-art pretrained face recognition model (SphereFaceNet [11]) to conduct face recognition experiments on the Multi-PIE database. The NI-LR images of 50 different individuals under 20 illumination conditions are chosen and hallucinated by the compared techniques. The UI-HR image with normal illumination of each individual is selected as the ground truth to construct the gallery set. For each testing image, we extract the deep features (SphereFace) [11] from the output of the FC_1 layer in SphereFaceNet model. Tab. 2 indicates the face recognition results of the compared methods.

Table 2. Testing accuracies of different methods on face recognition experiment

SR method	Accuracy	
	IN+FH	FH+IN
Bicubic	48.12%	49.84%
SRGAN [9]	50.68%	51.09%
TDAE [20]	52.26%	52.97%
FHC [18]	65.29%	
NI-LR	61.21%	
UI-HR	98.13%	
CPGAN	84.36%	

3. Data Generation based on RaIN

3.1. Training Settings

We first train the RaIN model using MS-COCO[‡] [10] and WikiArt[§] [15] for the content and style images, respectively. Both datasets consist of approximately 80,000 images. To encode the photo-realistic facial details, we incorporate the fine-tuning procedure on the Multi-PIE dataset [4]. During training and fine-tuning processes, we adopt the ADAM optimizer [7] using a batch size of 8 content-style image pairs. Other settings are similar to [6]. Inspired by [6, 8], we incorporate the content, style, reconstruction (for feature statistics (μ and σ) of the f_s) and Kullback-Leibler divergence (for feature statistics (μ and σ) of the f_s) losses to optimize our RaIN model.

[‡]http://cocodataset.org/home

[§]https://www.wikiart.org/

3.2. Training Details

Based on the trained RaIN model, we feed the content image (the face image with uniform illumination) along with a random noise to generate the stylized images with a different illumination condition. Fig.3 performs some generated NI-HR face samples.

For training our CPGAN, we resize the generated stylized face samples (with various illumination conditions) to 128×128 pixels and then apply 2D transforms, including rotations, translations, scaling and downsampling, to obtain the NI-LR images of 16×16 pixels. For the UI-HR images, we resize the corresponding content ones to 128×128 pixels. In this way, we generated sufficient NI-LR/UI-HR face pairs for training and testing of our CPGAN model.

To capture uneven illumination style, the color tone of the generated NI-HR samples by our RaIN model is slightly different from the face images in Multi-PIE dataset. When we downsample the generated NI-HR faces, this color jittering can be largely reduced (see Fig. 4). Benefiting from the guided HR image, CPGAN achieves natural color tune similar to the guided image; while faithfully hallucinating the NI-LR face images with both finer details and global shapes. Therefore, our CPGAN achieves superior performance in comparison with the state of the arts.

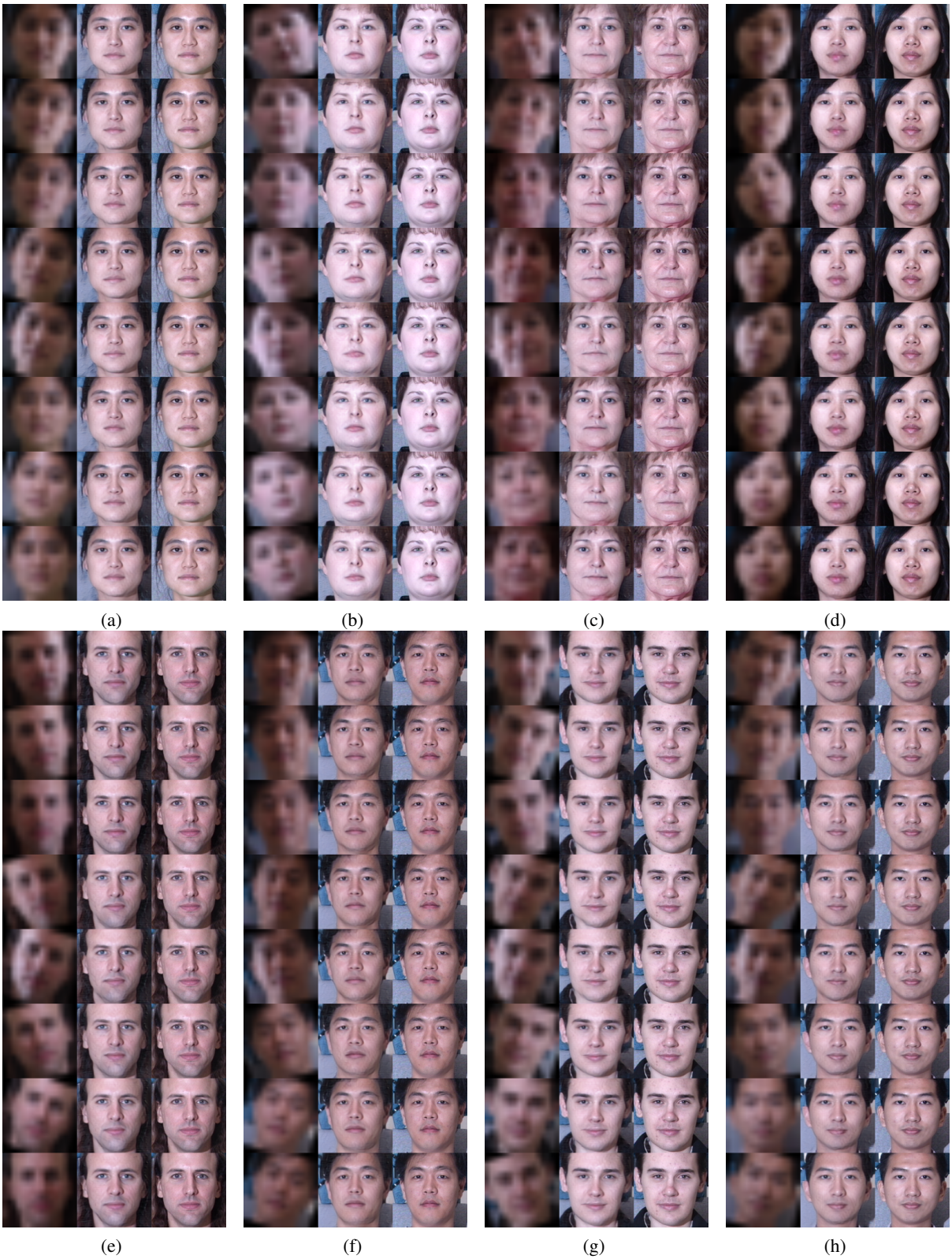

Figure 1. Hallucinated results for the samples in Multi-PIE dataset. In each subfigure, the images in the first and second columns are the input NI-LRs (under various illuminations) and the hallucinated UI-HRs. The images in the third column are the corresponding GTs.

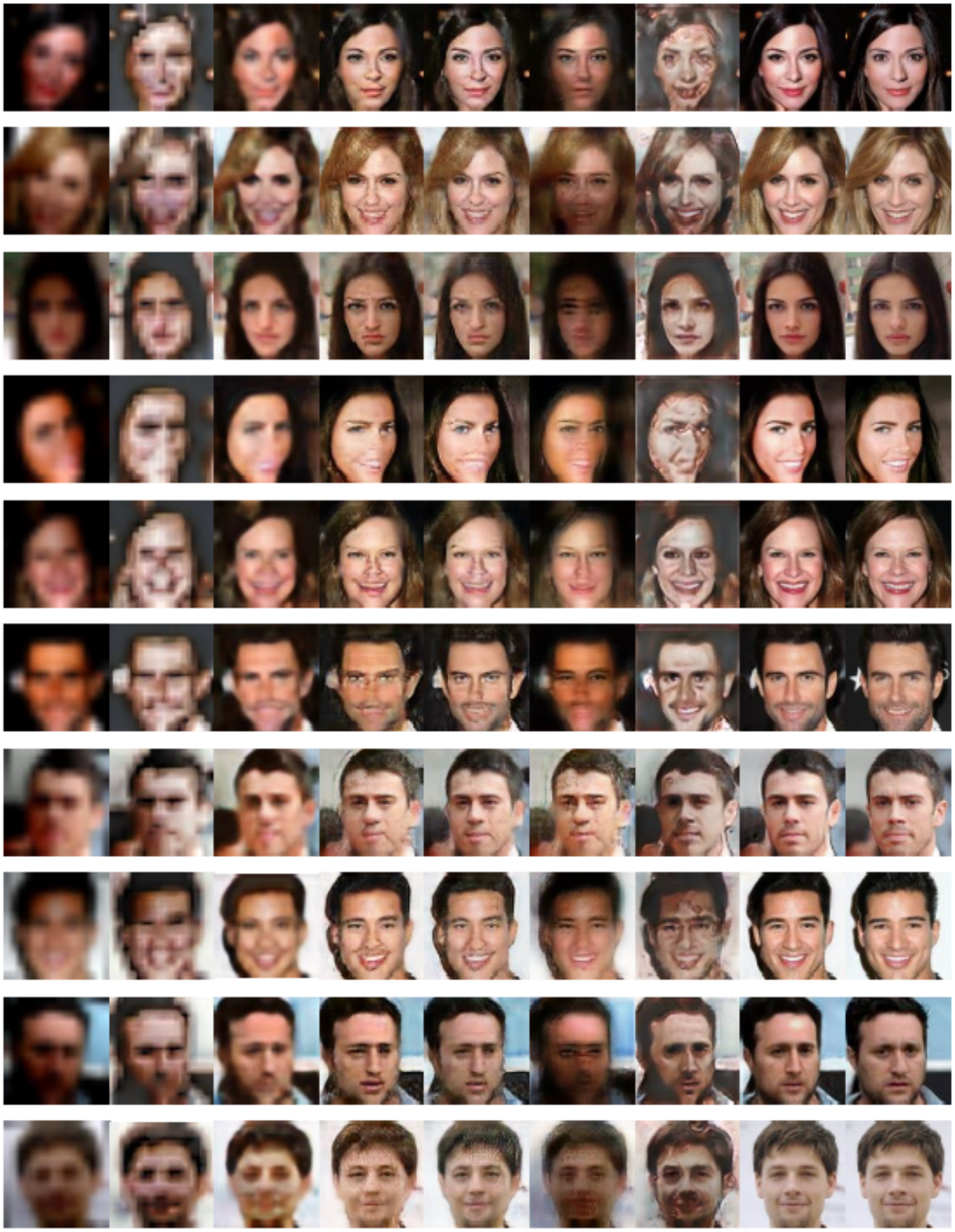

Figure 2. Hallucinated results for the samples in CelebA dataset. Comparison with state-of-the-art methods. Columns: (a) Unaligned NI-LR inputs. (b) Bicubic interpolation + CycleGAN [22]. (c) SRGAN [9]. (d) TDAE [20]. (e) FHC [18]. (f) CycleGAN [22] + SRGAN [9]. (g) TDAE [20] + CycleGAN [22]. (h) Ours. (i) GTs.

Figure 3. The NI-HR face samples generated by RaIN model.

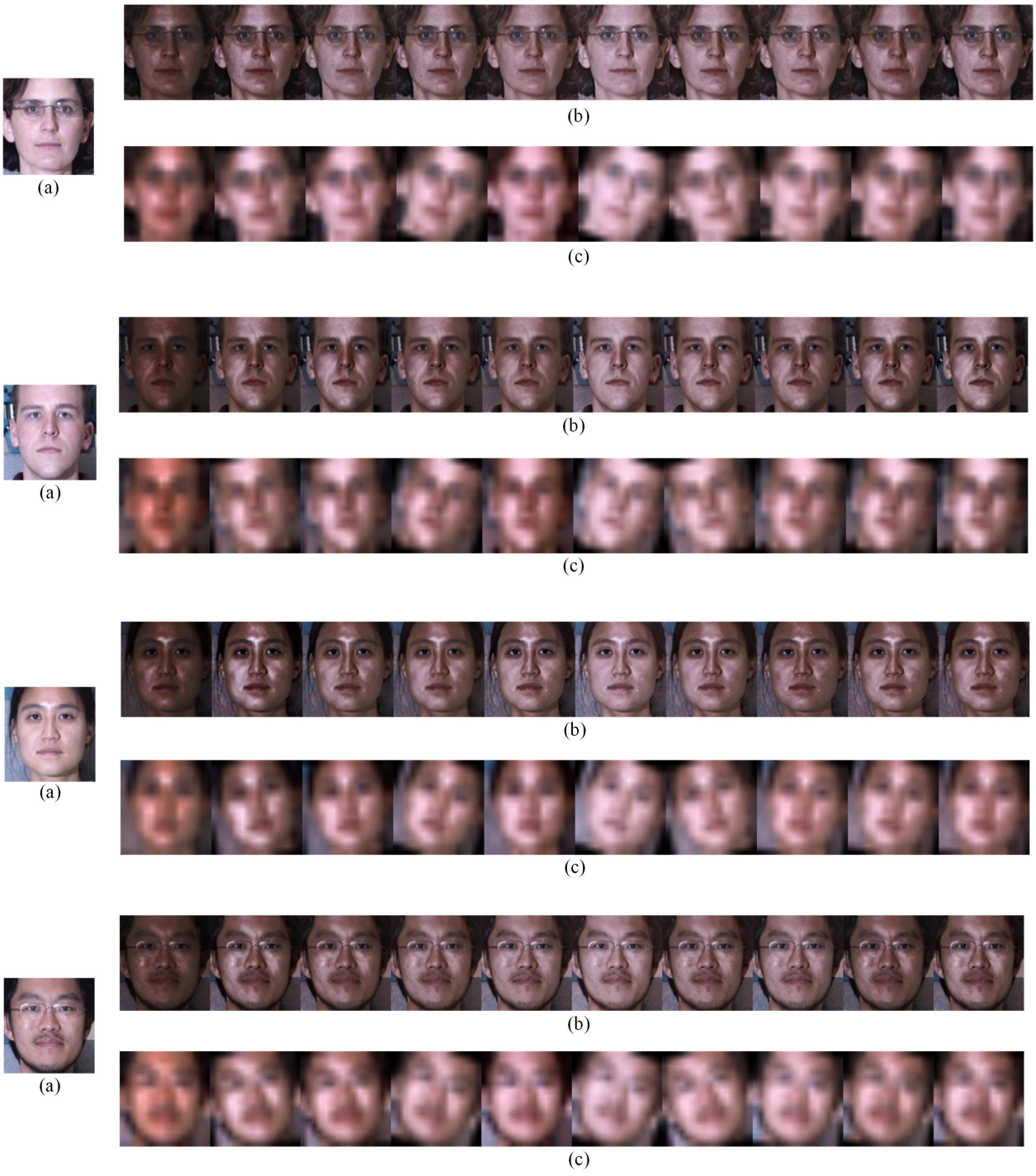

Figure 4. Illustration of the generated NI-LR/UI-HR face pairs. (a) UI-HR image (Original UI-HR face in Multi-PIE). (b) The NI-HR face images generated by RaIN. (c) Corresponding NI-LR faces. (Spatially transformed and downsampled version of (b)).

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[2] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[4] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[11] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[13] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[14] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016.

[15] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011.

[16] Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, and Piotr Koniusz. Identity-preserving face recovery from stylized portraits. *International Journal of Computer Vision*, 127(6-7):863–883, 2019.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, pages 217–233, 2018.

[19] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 908–917, 2018.

[20] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI*, 2017.

[21] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[22] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.