Conditional CycleGAN for Attribute Guided Face Image Generation

Yongyi Lu
HKUST
yluaw@cse.ust.hk

Yu-Wing Tai
Tencent
yuwingtai@tencent.com

Chi-Keung Tang
HKUST
cktang@cse.ust.hk

Abstract

State-of-the-art techniques in Generative Adversarial Networks (GANs) such as cycleGAN is able to learn the mapping of one image domain X to another image domain Y using unpaired image data. We extend the cycleGAN to *Conditional* cycleGAN such that the mapping from X to Y is subjected to attribute condition Z. Using face image generation as an application example, where X is a low resolution face image, Y is a high resolution face image, and Z is a set of attributes related to facial appearance (e.g. gender, hair color, smile), we present our method to incorporate Z into the network, such that the hallucinated high resolution face image Y' not only satisfies the low resolution constrain inherent in X, but also the attribute condition prescribed by Z. Using face feature vector extracted from face verification network as Z, we demonstrate the efficacy of our approach on identity-preserving face image super-resolution. Our approach is general and applicable to high-quality face image generation where specific facial attributes can be controlled easily in the automatically generated results.

1 Introduction

We are interested in realistic face image generation where facial attributes can be fully controlled in the automatic generation process. For example, in the ill-posed problem of single-image super-resolution (SISR), there exist many high resolution images that can generate the same low resolution input, so it is essential to preserve the person's identity in SISR for face image super-resolution. For face image super-resolution. In image-based cosmetic transfer, only the cosmetic style should be transferred from the source face to the target face, where the target face's feature should be preserved.

Existing Generative Adversarial Networks (GANs) [3] are able to generate highly realistic images. In particular, for SISR, the SRGAN [8] has produced impressive results with upscaling factor up to 4. While the hallucinated facial features and details look very realistic, they may not correspond to any real person's face. Recently, the cycleGAN [18] was proposed to address the image-to-image translation problem using unpaired image data, and it has produced many state-of-the-art results to date. In this paper, we capitalize on the cycleGAN, and propose the *conditional* cycleGAN where the face image result is generated subjected to input face attribute condition. Specifically, let X be a low resolution face image, Y be a high resolution face image, and Z be face attributes. Conditional cycleGAN incorporates Z into the network such that the hallucinated high resolution face image Y' satisfies not only the low resolution constrain from X, but also the attribute condition given by Z. We are particularly interested in SISR because it is a highly under-constrained problem, and being able to provide additional conditional constrain not only enhances the generated SR results, but also allows deterministic complex controls to achieve the desired results. This added advantage allows a range of face generation applications to be generalized under the same SISR framework.

With the proposed attribute-guided approach to face image generation, where the input consists of a low resolution face image and a set of face attributes during inference, we demonstrate the efficacy of our approach on *identity-preserving* face image super-resolution. By simply altering

the attribute condition, our approach can be directly applied to generate high-quality face images that simultaneously preserve the constraints given in the low resolution input while transferring facial features (e.g. gender, hair color, emotion, sun-glasses) prescribed by the input face attributes. As will be demonstrated in our new and significant results, while the automatically generated face images are not perfect, the artifacts can arguably be edited away easily, which is in contrast with conventional photo-editing where a lot of human intervention is required in interactive image matting and compositing, warping and blending to produce realistic outputs.

2 Related Work

Recent state-of-the-art image generation techniques have leveraged the deep convolutional neural networks (CNNs). For example, in SISR, a deep recursive CNN for SISR was proposed in [7]. Learning upscaling filters have improved accuracy and speed [2, 14, 15]. A deep CNN approach was proposed in [1] using bicubic interpolation. The ESPCN [14] performs SR by replacing the deconvolution layer in lieu of upscaling layer. However, many existing CNN-based networks still generate blurry images. The SRGAN [8] uses the Euclidean distance between the feature maps extracted from the VGGNet to replace the MSE loss which cannot preserve texture details. The SRGAN has improved the perceptual quality of generated SR images. A deep residual network (ResNet) was proposed in [8] that produces good results for upscaling factors up to 4. In [6] both the perceptual/feature loss and pixel loss are used in training SISR.

Existing GANs [3] have generated many state-of-the-art results in automatic image generation. The key of their success lies in the adversarial loss which forces the generated images to be indistinguishable from real images. This is achieved by two competing neural networks, the generator and the discriminator. In particular, the DCGAN [12] incorporates deep convolutional neural networks into GANs, and has generated some of the most impressive realistic images to date. GANs are however notoriously difficult to train: GANs are formulated as a minimax "game" between two networks. In practice, it is hard to keep the generator and discriminator in balance, where the optimization can oscillate between solutions which may easily cause the generator to collapse. Among different techniques, including the conditional GAN [5] proposed to address this problem, enforcing forward-backward consistency has emerged to be one of the most effective ways to train GAN.

Forward-backward consistency has been enforced in computer vision algorithms such as image registration, shape matching, co-segmentation, to name a few. In the realm of image generation using deep learning, using unpaired training data, the CycleGAN [18] was proposed to learn image-to-image translation from a source domain X to a target domain Y. In addition to the standard GAN loss respectively for X and Y, a pair of cycle consistency losses (forward and backward) was formulated using L1 reconstruction loss. For forward cycle consistency, given $x \in X$ the image translation cycle should reproduce x. Backward cycle consistency is similar. In this paper, we propose conditional cycleGAN for face image generation so that the image generation process can preserve (or transfer) facial identity, where the results can be controlled by various input attributes. Preserving facial identity has also been explored in synthesizing the corresponding frontal face image from a single side-view face image [4], where the identity preserving loss was defined based on the activations of the last two layers of the Light CNN [16]. In multi-view image generation from a single view [17], a condition image (e.g. frontal view) was used to constrain the generated multiple views in their coarse-to-fine framework. However, facial identity was not explicitly preserved in their results and thus many of the generated faces look smeared, although as the first generated results of multiple views from single images, the pertinent results already look quite impressive.

While our conditional cycleGAN is an image-to-image translation framework, the IcGAN [11] factorizes an input image into a latent representation z and conditional information y using their respective trained encoders. By changing y into y', the generator network then combines the same z and new y' to generate an image that satisfies the new constraints encoded in y'. We are inspired by their best conditional positioning, that is, where y' should be concatenated among all of the convolutional layers. For SISR, in addition, z should represent the embedding for an (unconstrained) high resolution image, where the generator can combine with the identity feature y to generate the superresolved result. In [9] the authors proposed to learn the dense correspondence between a pair of input source and reference, so that visual attributes can be swapped or transferred between them. As we will show in our experiments, their work can be regarded as a special case for our

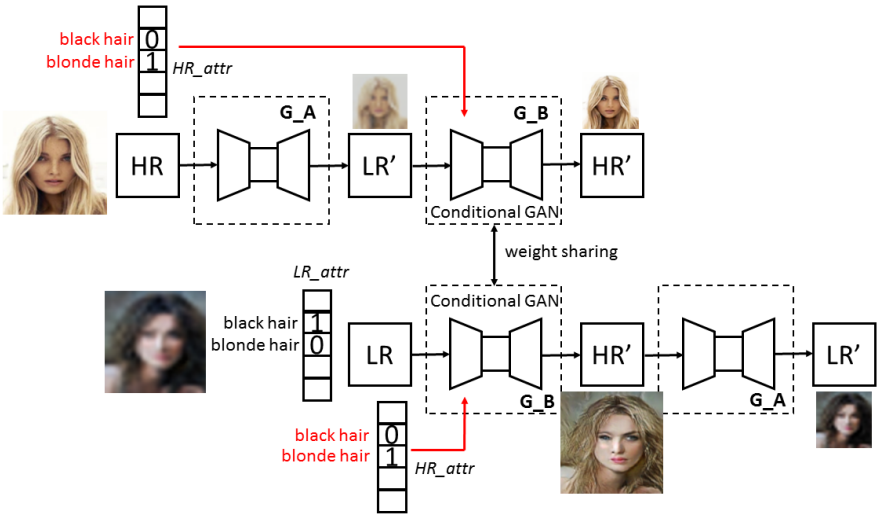

Figure 1: Our Conditional CycleGAN for attribute guided face super-resolution. On top of the original cycleGAN, we embed an additional attribute vector, and utilize conditional GAN to train a generator G_B to generate high resolution face image given the low resolution face image and the attribute vector as inputs.

identity-preserving face super-resolution where in our case the respective inputs are downsampled with identities given by swapping original (high resolution) images.

3 Conditional CycleGAN

3.1 CycleGAN

A Generative Adversarial Network [3] (GAN) consists of two neural networks, a generator $G_{X\to Y}$ and a discriminator D_Y, which are iteratively trained in a two-player minimax game manner. The adversarial loss $\mathcal{L}(G_{X\to Y}, D_Y)$ is defined as

$$\mathcal{L}(G_{X\to Y}, D_Y) = \min_{\Theta_g}\max_{\Theta_d}\{\mathbb{E}_y[\log D_Y(y)] + \mathbb{E}_x[\log(1 - D_Y(G_{X\to Y}(x)))]\} \quad (1)$$

where Θ_g and Θ_d are respectively the parameters of the generator $G_{X\to Y}$ and discriminator D_Y, and $x \in X$ and $y \in Y$ denotes the training data in source and target domain respectively.

In cycleGAN, X and Y are two different image representations, and the cycleGAN learns the translation $X \to Y$ and $Y \to X$ simultaneously. Different from "pix2pix" [5], training data in cycleGAN is unpaired. Thus, they introduce Cycle Consistency to enforce forward-backward consistency which can be considered as "pseudo" pairs of training data. With the Cycle Consistency, the loss function of cycleGAN is defined as:

$$\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \mathcal{L}(G_{X\to Y}, D_Y) + \mathcal{L}(G_{Y\to X}, D_X) + \lambda\mathcal{L}_c(G_{X\to Y}, G_{Y\to X}) \quad (2)$$

where $\mathcal{L}_c(G_{X\to Y}, G_{Y\to X}) = ||G_{Y\to X}(G_{X\to Y}(x)) - x||_1 + ||G_{X\to Y}(G_{Y\to X}(y)) - y||_1$ is the Cycle Consistency Loss. In our implementation, we follow the network architecture of cycleGAN to train our conditional cycleGAN except for the modifications described in the next subsections.

3.2 Attribute Guided Conditional CycleGAN

We are interested in face image generation guided by user-input facial attributes to control the high-resolution results. To include conditional constrain into the cycleGAN network, the adversarial loss is modified to include the conditional feature vector as part of the input of the generator and discriminator as

$$\mathcal{L}(G_{(X,Z)\to Y}, D_Y) = \min_{\Theta_g}\max_{\Theta_d}\left\{\mathbb{E}_{y,z}[\log D_Y(y,z)] + \mathbb{E}_{x,z}[\log(1 - D_Y(G_{(X,Z)\to Y}(x,z),z))\right\} \quad (3)$$

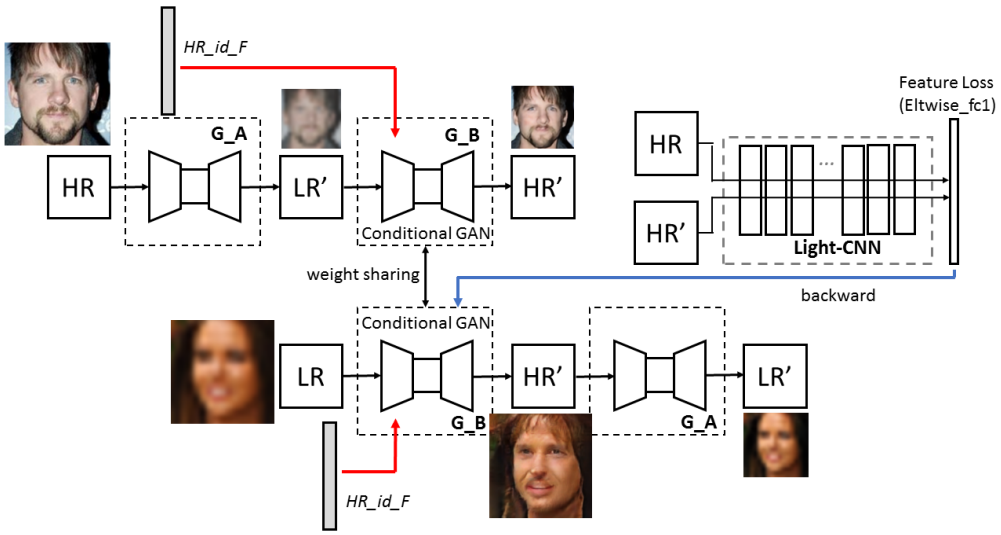

Figure 2: Our Conditional CycleGAN for identity preserving face super-resolution. Different from attribute guided face super-resolution, we include additional face verification loss into the training process. The network for computing the face verification loss is pretrained.

In our implementation, the conditional feature vector is first resized (using replicate) to match the image size of the input image which is downsampled into a (very) low resolution image, with the intensity value of each feature map equal to the value of each column of the feature vector. Hence, for 18-dimensional feature vector, we have 18 homogeneous feature maps after resizing. The resized feature vector is then concatenated with the $conv1$ layer of the generator network to propagate the inference of feature vector to the generated images. In the discriminator network, the resized feature vector is also concatenated with the $conv1$ layer.

In order to train the conditional GAN network, only the correct pair of groundtruth high resolution face image and feature vector are treated as positive examples. The generated high resolution face image matched with groundtruth feature vector, and the groundtruth high resolution face image matched with randomly sampled feature vector are both treated as negative examples. We refer readers to IcGAN [11] for further details of the training of conditional GAN.

With the conditional adversarial loss, we modify the cycleGAN network as illustrated in Figure 1. The conditional feature vector is provided by the CelebA dataset [10], and we follow IcGAN [11] to pick the same 18 attributes as our conditional feature vector. Note that the conditional feature vector is associated with the high resolution face image, instead of the low resolution face image. In each "pair" of training iteration, the same conditional feature vector is used to generate the high resolution face image. Hence, the generated intermediate high resolution face image in the lower branch of Figure 1 would have different attributes from its actual high resolution image. This is on purpose because the conditional discriminator network would enforce the generator network to utilize the information from the conditional feature vector. If the conditional feature vector always receives the correct attributes, the generator network would learn to skip the information in the conditional feature vector, since some of the attributes can be found in the low resolution face image.

3.3 Identity Preserving Conditional CycleGAN

To demonstrate the efficacy of our conditional cycleGAN guided by control attributes, we use identity preserving face image super-resolution as an example. We utilize the feature vector from Light-CNN [16] as the conditional feature vector. The identity feature vector is a 256-D vector. Compared with another state-of-the-art FaceNet [13], which returns a 1792-D face feature vector for each face image, the 256-D representation of light-CNN obtains state-of-the-art results while it has fewer parameters and runs faster.

In our initial implementation, we follow the same architecture and training strategy to train the conditional cycleGAN for identity preserving face super-resolution. However, we found that the

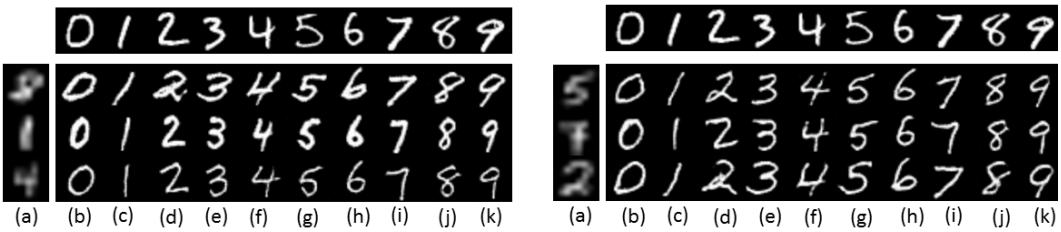

Figure 3: From the low resolution digit images (a), we can generate high resolution digit images (b) to (k) subject to the conditional constrain from the digit class label in the first row.

trained network does not generate good results. We believe this is because the discriminator network is trained from scratch, and the trained discriminator network is not as powerful as the light-CNN which was trained from million pairs of face images. Thus, we include additional face verification loss in parallel with the discriminator network as illustrated in Figure 2. The verification errors from the light-CNN network is back propagated concurrently with the errors from the discriminator network. After the inclusion of the face verification loss, we are able to generate high quality high resolution face images matching the identity given by the conditional feature vector. As shown in the running example in Figure 2, the lady's face is changed to a man's face whose identify is given by the light-CNN feature.

4 Experiments

4.1 Datasets

We use two image datasets, MNIST (for sanity check) and CelebA [10] (for face image generation) to evaluate our method. The MNIST is a digit dataset of 60,000 training and 10,000 testing images. Each image is a 28×28 black and white digit image with the class label from 0 to 9. The CelebA is a face dataset of 202,599 face images, with 40 different attribute labels where each label is a binary value. We use the aligned and cropped version, with 182K images for training and 20K for testing. To generate low resolution images, we downsampled the images in both dataset by a factor of 8, and we separate the images such that the high resolution and low resolution training images are non-overlapping.

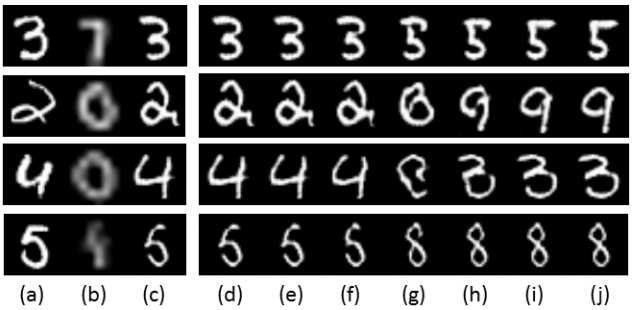

Figure 4: Interpolation results of digits. (a) HR digit images; (b) LR digit images; (c) generated digit face images; (d) to (j) interpolated results. We interpolate between the respective binary vectors of the source and destination digits.

4.2 MNIST

We first evaluate the performance of our method on MNIST dataset. The conditional feature vector is the class label of digits. As shown in Figure 3, our method can generate high resolution (HR) digit images from the low resolution (LR) inputs. Note that the generated high resolution digit follows the given class label when there is conflict between the low resolution image and feature vector. This is desirable, since the conditional constraint consumes large weights during the training. This sanity check also verifies that we can impose conditional constraint into the cycleGAN network.

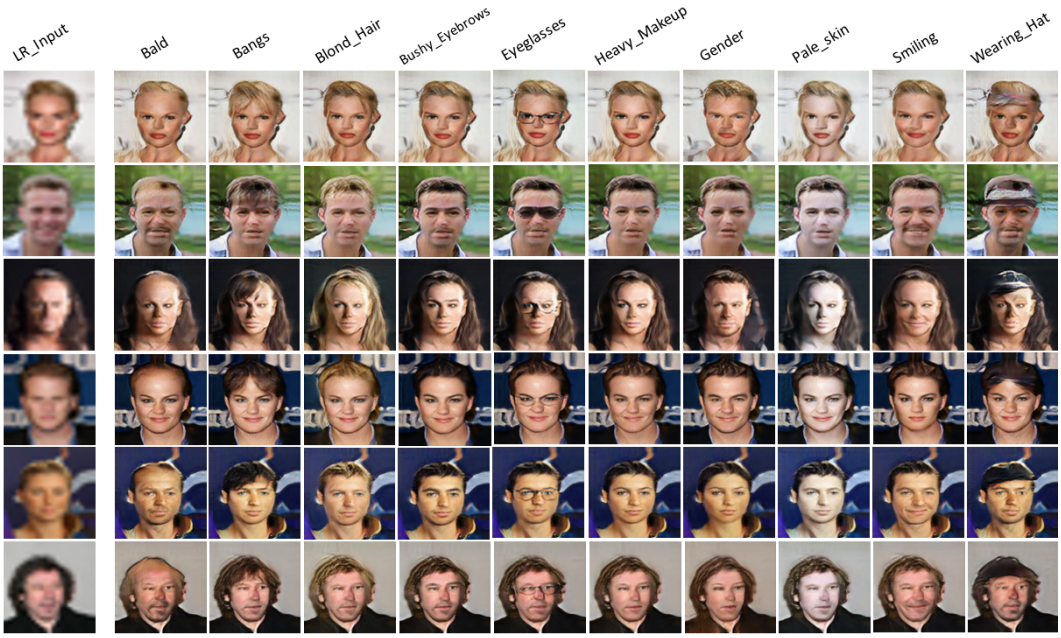

Figure 5: We flip one attribute label per each generated high resolution face images, given the low resolution face inputs. The 10 labels are: Bald, Bangs, Blond_Hair, Bushy_Eyebrows, Eyeglasses, Heavy_Makeup, Male, Pale_Skin, Smiling, Wearing_Hat.

Apart from the label changes based on the HR inputs, we also find that the generated HR images share some common properties with the LR inputs, such as the orientation and thickness of the digits presented in the input images. We can see from Figure 3 the outputs share the same orientation with the LR input '8' which is tilted to the right. In the next row the outputs adopt the thickness of the input, that is, the relatively thick stroke presented by the LR '1'. The reason is that we enforce cycle consistency in our loss function Eq. (2), and the generated HR images cannot 'wander' too far from the LR inputs. This is a good indicator of the ability of our trained generator: allow a certain degree of freedom in changing labels based on the HR images presented as identity attribute, while preserving the essential appearance presented by the LR inputs.

Apart from generating high resolution digit images from the low resolution inputs, we also perform linear interpolation between two high resolution images (as identity features) to show our model is able to learn the digit representation. Specifically, we interpolate between the respective binary vectors of the two digits. Sample results are shown in Figure 4.

4.3 Attribute Guided Face SR

Figure 5 shows sample results for attribute guided face generation using SISR. Recall the condition is encoded as a 18-D vector. The 10 results shown in the figure are generated with one attribute label flipped in their corresponding condition vector in conditional cycleGAN. Note the inherent difficulty associated with some of the labels (e.g. BALD, WEARING_HAT) where the generated results may not be totally convincing for these particular examples. On the other hand, the generated results conditioned on attributes such as BANGS, HEAVE_MAKEUP, MALE, MOUTH_SLIGHTLY_OPEN, PALE_SKIN are quite convincing.

4.4 Identity Preserving Face SR

Figure 6 shows sample SR results where the identity face features are from *different* persons. There are two interesting points to note: First, the generated high resolution results in (e) bears high resemblance to (a) from which the respective identity feature vectors are computed using Light-CNN. The unique identity features transfer well from (a) to (c), such as eye glasses in the third row and gender change in the fourth row. Second, the facial expression of the generated high resolution images adopt the facial expression in the *low* resolution inputs. Specifically, refer to the example in

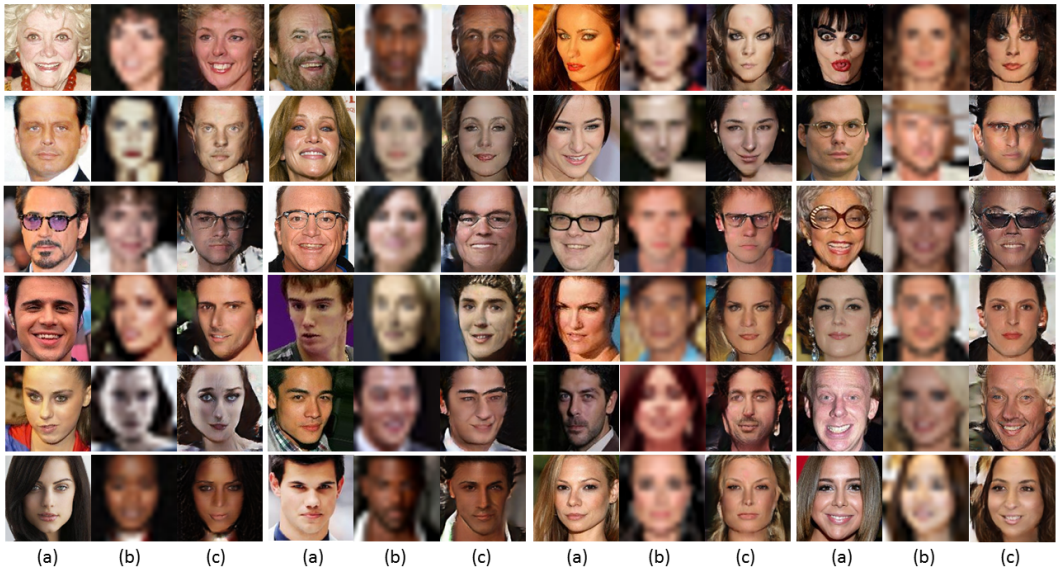

Figure 6: Identity preserving face super-resolution. (a) HR face images; (b) LR face images; (c) generated HR face images from (b). The conditional identity vector is computed from (a), and the generated HR image in (c) is enforced to have the same identity as (a).

the last row on the left side of Figure 6, where in (a) the mouth is closed while in (d) the smiling mouth is open with teeth showing. The generated high resolution image in (e) preserves the identity in (a) while the mouth is open with teeth showing. Facial expression is *not* part of the identity, and the original facial expression in the low resolution input, that is, smiling with teeth showing, should be preserved in the high resolution output. Another relevant example is the skin tone which is demonstrated in the first two examples in the last row of the results.

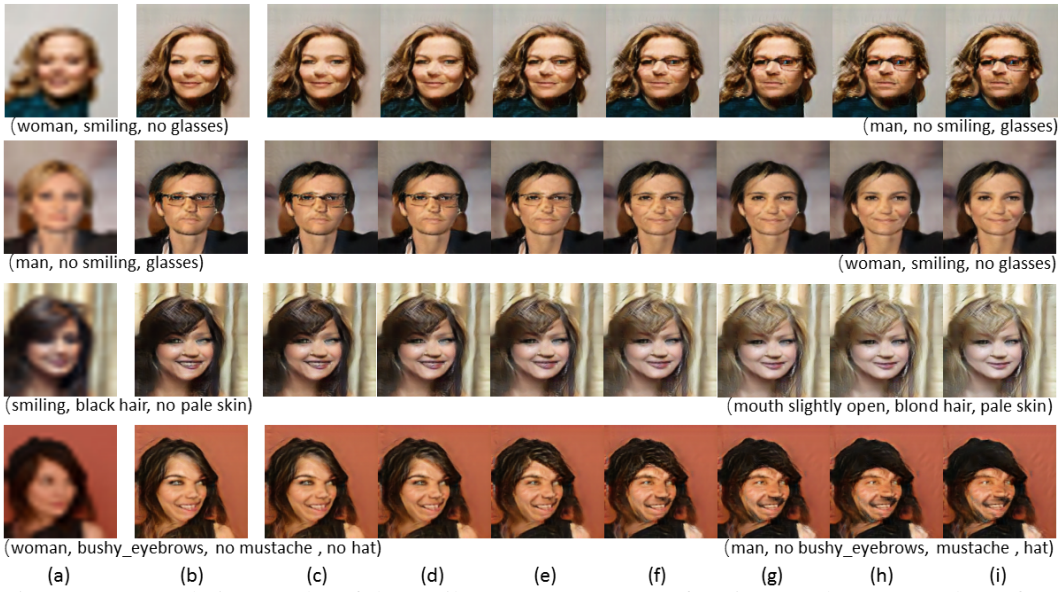

Figure 7: Interpolation results of the attribute vectors. (a) LR face inputs; (b) generated HR face images; (c) to (i) interpolated results. Attributes of source and destination are shown in text.

4.5 Interpolating Attribute Vector

The conditional attribute information learned by the generator can be further explored beyond generating images by flipping the attribute labels (Recall the condition is encoded as a 18-D vector). In this section we linearly interpolate between two different attribute vectors to generate the interpolated

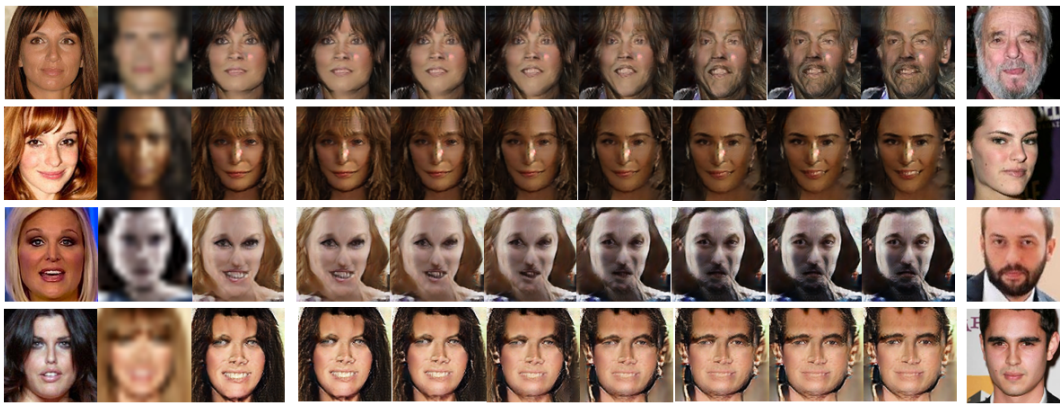

Figure 8: Interpolating results of the identity feature vectors. (a) HR face inputs (b) LR face inputs; (c) generated HR face images. We randomly sample another HR image (k) and interpolate between identity features of (a) and (k). Results are shown in (d) to (j).

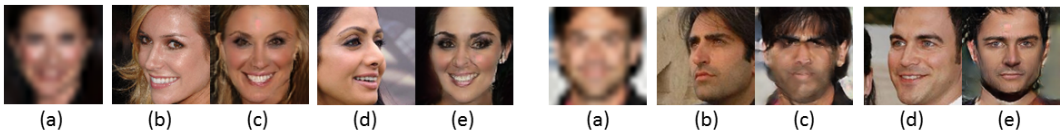

Figure 9: Generated frontal face results. Given a LR template (a), our method can generate corresponding frontal faces from different side faces, e.g., (b) to (c), (d) to (e).

faces. As Figure 7 shown, all the interpolated faces are visually plausible, with smooth transition among them. This further convinces us that the model is generalizing the face representation properly instead of just directly memorizing the training samples.

4.6 Interpolating Identity Vector

Like interpolating the attribute vectors, we also experiment with interpolating the 256-D identity feature vectors under our identity-preserving conditional model. We randomly sample two HR face images and interpolate between the two identity features. Results are shown in Figure 8, which indicates that our model generalizes the face representation properly given the conditional feature vectors.

4.7 Frontal Face Generation

Another application of our model consists of generating images of frontal faces from face images in other orientations. By simply providing a low resolution frontal face image and adopting our Identity Preserving Conditional CycleGAN model, we can generate the corresponding high resolution frontal face images given side-face images as HR face attributes. Figure 9 shows sample results on our frontal face image generation.

5 Conclusion

We have presented the Conditional CycleGAN for attribute guided face image generation where the processing pipeline is similar to SISR. Our technical contribution consists of the conditional cycleGAN to guide the super-resolution process via easy user input of complex attributes for generating high quality results. In the attribute guided conditional cycleGAN, the adversarial loss is modified to include a conditional feature vector as parts of the inputs to the generator and discriminator networks. We utilize the feature vector from light-CNN in identity-preserving conditional cycleGAN. We have presented the first and significant results on identity-preserving face super-resolution and attribute-guided face image generation involving transfer of gender, hair color and emotion. While the application example focuses on face in this paper, our conditional cycleGAN is general and can be easily extended to other applications, which is the focus of our future work.

References

[1] C. Dong, C. C. Loy, K. He, and X. Tang. *Learning a Deep Convolutional Network for Image Super-Resolution*, pages 184–199. Springer International Publishing, Cham, 2014.

[2] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. *CoRR*, abs/1608.00367, 2016.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680. 2014.

[4] R. Huang, S. Zhang, T. Li, and R. He. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. *ArXiv e-prints*, Apr. 2017.

[5] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[6] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[7] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016.

[8] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[9] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual Attribute Transfer through Deep Image Analogy. *ArXiv e-prints*, May 2017.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *CoRR*, abs/1411.7766, 2014.

[11] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible Conditional GANs for image editing. *ArXiv e-prints*, Nov. 2016.

[12] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[13] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, June 2015.

[14] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016.

[15] Y. Wang, L. Wang, H. Wang, and P. Li. End-to-end image super-resolution via deep and shallow convolutional networks. *CoRR*, abs/1607.07680, 2016.

[16] X. Wu, R. He, and Z. Sun. A lightened CNN for deep face representation. *CoRR*, abs/1511.02683, 2015.

[17] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, and J. Feng. Multi-View Image Generation from a Single-View. *ArXiv e-prints*, Apr. 2017.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.