Generative Adversarial Network with Spatial Attention for Face Attribute Editing

Gang Zhang[1,2], Meina Kan[1,3], Shiguang Shan[1,3(✉)], and Xilin Chen[1]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, CAS, Beijing 100190, China
gang.zhang@vipl.ict.ac.cn,{kanmeina,sgshan,xlchen}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China

Abstract. Face attribute editing aims at editing the face image with the given attribute. Most existing works employ Generative Adversarial Network (GAN) to operate face attribute editing. However, these methods inevitably change the attribute-irrelevant regions, as shown in Fig. 1. Therefore, we introduce the spatial attention mechanism into GAN framework (referred to as SaGAN), to only alter the attribute-specific region and keep the rest unchanged. Our approach SaGAN consists of a generator and a discriminator. The generator contains an attribute manipulation network (AMN) to edit the face image, and a spatial attention network (SAN) to localize the attribute-specific region which restricts the alternation of AMN within this region. The discriminator endeavors to distinguish the generated images from the real ones, and classify the face attribute. Experiments demonstrate that our approach can achieve promising visual results, and keep those attribute-irrelevant regions unchanged. Besides, our approach can benefit the face recognition by data augmentation.

Keywords: Face attribute editing · GAN · Spatial attention
Data augmentation

1 Introduction

Face attribute editing is the task that alters the face image towards a given attribute. It has been widely used in facial animation, art, entertainment, and face expression recognition [1–4] and has drawn increasing attentions in recent years. The desired result of face attribute editing (e.g. expression editing or removing/wearing eyeglasses etc.) is that the attribute-specific region is altered to the given attribute while the rest irrelevant region keeps unchanged.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01231-1_26) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11210, pp. 422–437, 2018.
https://doi.org/10.1007/978-3-030-01231-1_26

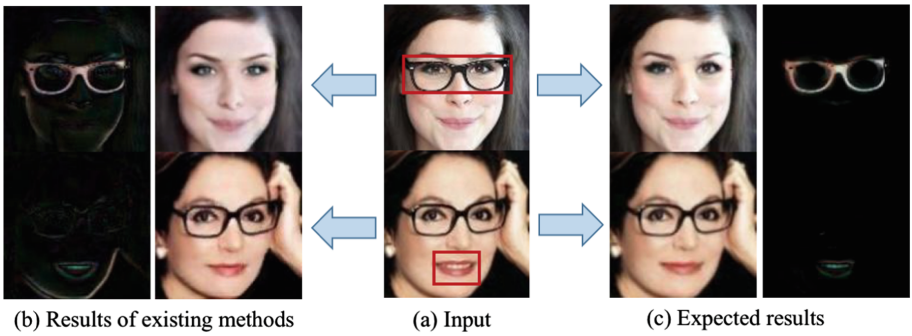

Fig. 1. Illustration of Face attribute editing. (a) Shows the input face images, and the attributes to be edited are eyeglass and mouth close, with the corresponding attribute-specific region marked in red boxes respectively. (b) Shows the residual images and the attribute edited images from existing methods, where the whole image is altered although the attribute to be edited is local. (c) Shows the expected edited images and the expected residual images respectively, where only those attribute-specific region should be altered and the rest keep unchanged. The residual images are defined as the differences between the input face images and the edited face images. (color figure online)

In early years, face attribute editing is treated as a regression problem by using paired training samples, such as face frontalization or face eyeglasses removal. Zhu *et al.* [5] proposed a face frontalization method, which takes as input a face image to regress the desired frontal face image by minimizing the pixel-wise image reconstruction loss. To remove the eyeglasses from a face image, Zhang *et al.* [6] trained a model of multi-variable linear regression with training samples collected from face images with eyeglasses and their corresponding face images without eyeglasses. The performance of these methods heavily depends on the paired training datas, which are however quite difficult to acquire.

Recently, Generative Adversarial Network (GAN), proposed by Goodfellow *et al.* [7], has achieved great progress in image generation [8–10], image super resolution [11] and neural style transfer [12,13]. Face attribute editing also benefits a lot from GAN, which treats the face attribute editing as an unpaired image-to-image translation task. The conventional GAN framework consists of a generator and a discriminator. The discriminator learns to distinguish the generated images from the real ones, while the generator manages to fool the discriminator to produce photo-realistic images. The GAN approaches take the original face image as input and generate the edited face image with the given attribute. An extension for specific generation is conditional GANs (cGANs) [14], which allows to generate specific images given a conditional signal. Furthermore, IcGAN [15] introduces an encoder to the cGANs forming an invertible conditional GANs (IcGAN) for face attribute editing, which maps the input face image into a latent representation and an attribute vector. The face image with new attributes is generated with the altered attributes vector as the condition.

For better generation in the absence of paired samples, dual learning has been introduced into GAN-based methods [12]. In [12], an effective unpaired image translation method CycleGAN is proposed by coupling the generation and its inverse mapping under a cycle consistency loss. CycleGAN is used in a wide range of applications, including style transfer, object transfiguration, attributes transfer and photo enhancement. A recent work, StarGAN [16], also adopts cycle consistency loss, but differently the generator of StarGAN takes an image and a domain manipulation vector as input, which allows to translate images between multiple domains using only a single model with promising results.

However, all above methods directly operate on the whole image, and thus inevitably change the rest attribute-irrelevant region besides the attribute-specific region. To avoid change the whole images, Shen *et al.* [17] models the face attribute editing as learning a sparse residual image, which is defined as the difference between the input face image and the desired manipulated image. This method is referred to as ResGAN in this work for short. Compared with operating on the whole image, learning the residual image avoids changing the attribute-irrelevant region by restraining most regions of the residual image as zero. This work is quite insightful to enforce the manipulation mainly concentrate on local areas especially for those local attributes. However, the location and the appearance of target attributes are modeled in single sparse residual image which is actually hard for a favorable optimization than modeling them separately, and this can be seen from the Fig. 4 in [17], where the response of the residual image scattered the whole image although the strong response of the residual image mainly concentrate on the local areas, even for the eyeglass attribute.

Inspired by the ResGAN [17], in this work we introduce the spatial attention mechanism into GAN for more accurate face attribute editing. Spatial attention mechanism allows one to select those prior part and ignore the rest for further faster or more accurate processing, which has performed successfully in image classification [18–20], and semantic segmentation [21], etc. For face attribute editing, spatial attention mechanism can be used to restrict the manipulation only within the attribute-specific regions. The proposed GAN with spatial attention (referred to as SaGAN) consists of a generator and a discriminator. The generator aims at generating face images with target attribute for an input image. The generator is made up of two networks, an attribute manipulation network (AMN) to edit the face image with the given attribute and a spatial attention network (SAN) to localize the attribute-specific region which restricts the alternation of AMN within this region. As adversary of the generator, the discriminator distinguishes the generated images from the real ones, and classifies the face attribute. Compared with the ones operating on the whole image or learning a sparse residual image, the proposed SaGAN can precisely localize the attribute-specific region for editing by utilizing the spatial attention mechanism. Experiments demonstrate that the proposed SaGAN achieves promising visual results and further benefits the face recognition by data augmentation.

In brief, our contribution can be summarized in three-folds:

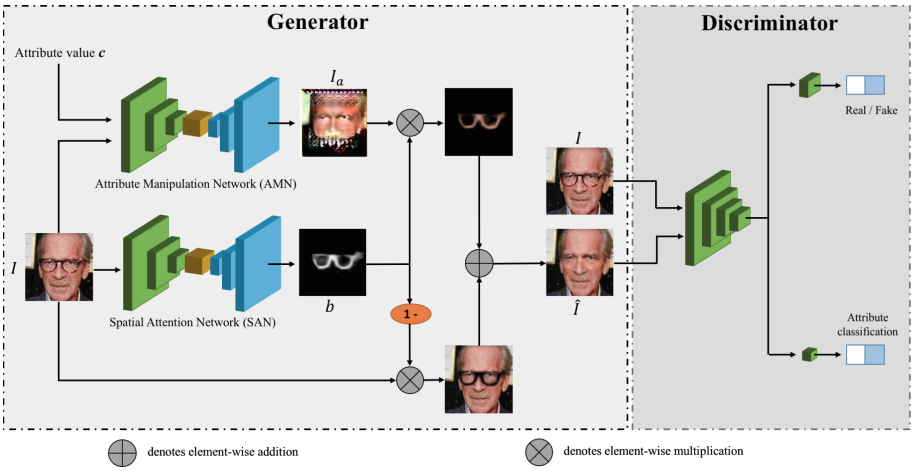

Fig. 2. An overview of our proposed SaGAN, consisting of a generator G and a discriminator D. G consists of an attribute manipulation network (AMN) to edit the face image with the given attribute, and a spatial attention network (SAN) to localize the attribute-specific region which restricts the alternation of AMN within this region. D learns to distinguish the generated images from the real ones, and classify the face attribute.

- The spatial attention is introduced to the GAN framework, forming an end -to-end generative model for face attribute editing (referred to as SaGAN), which can only alter those attribute-specific region and keep the rest irrelevant region remain the same.
- The proposed SaGAN adopts single generator with attribute as conditional signal rather than two dual ones for two inverse face attribute editing.
- The proposed SaGAN achieves quite promising results especially for those local attributes with the attribute-irrelevant details well preserved. Besides, our approach also benefits the face recognition by data augmentation.

2 Generative Adversarial Network with Spatial Attention

In this section, we will first describe the details of the generative adversarial network with spatial attention (SaGAN) method; and then give a detailed discussion about the difference from the existing methods.

An overview of SaGAN is shown in Fig. 2. For a given input image I and an attribute value c, the goal of face attribute editing is to translate I into an new image $\hat{I}$, which should be realistic, with attribute c and look the same as the input image excluding the attribute-specific region. The SaGAN consists of a generator G and a discriminator D in adversarial manner.

2.1 Discriminator

The discriminator D, as adversary of generator, has two objectives, one to distinguish the generated images from the real ones, and another to classify the attributes of the generated and real images, as shown in Fig. 2. The two classifiers are both designed as a CNN with softmax function, denoted as D_{src} and D_{cls} respectively. Generally, the two networks can share the first few convolutional layers followed by distinct fully-connected layers for different classifications.

The output of real/fake classifier $D_{src}(I)$ means the probability of an image I to be a real one, and that of attribute classifier $D_{cls}(c|I)$ means the probability of an image I with the attribute c. Here, $c \in \{0, 1\}$ is a binary indicator of with or without an attribute. The input images can be real ones or generated ones.

The loss for optimizing the real/fake classifier is formulated as a standard cross-entropy loss as below:

$$\mathcal{L}_{src}^{D} = \mathbb{E}_{I}[\log D_{src}(I)] + \mathbb{E}_{\hat{I}}[\log(1 - D_{src}(\hat{I}))], \tag{1}$$

where I is the real image and $\hat{I}$ is the generated image. Similarly, the loss for optimizing the attribute classifier is also formulated as a standard cross-entropy loss as below:

$$\mathcal{L}_{cls}^{D} = \mathbb{E}_{I,c^g}[-\log D_{cls}(c^g|I)], \tag{2}$$

where c^g is the ground truth attribute label of the real image I.

Finally, the overall loss function for discriminator D is formulated as follows:

$$\min_{D_{src},D_{cls}} \mathcal{L}_{D} = \mathcal{L}_{src}^{D} + \mathcal{L}_{cls}^{D}. \tag{3}$$

By minimizing Eq. (3), the obtained discriminator D can well separate the real images from those fake ones, and correctly predict the probability that an image I is with the attribute c.

2.2 Generator

The generator G endeavors to translate an input face image I into an edited face image $\hat{I}$ conditioned on an attribute value c, formulated as follows:

$$\hat{I} = G(I, c), \tag{4}$$

G contains two modules, an attribute manipulation network (AMN) denoted as F_m and a spatial attention network (SAN) denoted as F_a. AMN focuses on how to manipulate and SAN focuses on where to manipulate.

The attribute manipulation network takes a face image I and an attribute value c as input, and outputs an edited face image I_a, which is formulated as

$$I_a = F_m(I, c). \tag{5}$$

The spatial attention network takes the face image I as input, and predict a spatial attention mask b, which is used to restrict the alternation of AMN within this region, formulated as below:

$$b = F_a(I), \tag{6}$$

Ideally, the attribute-specific region of b should be 1, and the rest regions should be 0. In practice, the values may be any continuous number between 0 and 1 after the optimization. Therefore, those regions with non-zeros attention values are all regarded as attribute-specific region, and the rest with zero attention values are regarded as attribute-irrelevant region.

Guided by the attention mask, in the final edited face image $\hat{I}$, the attribute-specific regions are manipulated towards the target attribute while the rest regions remain the same, formulated as below:

$$\hat{I} = G(I, c) = I_a \cdot b + I \cdot (1 - b), \tag{7}$$

A favorable attribute edited image should be realistic, correctly with target attribute c, and also with modest manipulation, i.e. keep those attribute-irrelevant regions unchanged. So naturally, three kinds of losses are needed to ensure the achieving of these goals.

Firstly, to make the edited face image $\hat{I}$ photo-realistic, an adversarial loss is designed to confuse the real/fake classifier following most GAN-based methods:

$$\mathcal{L}_{src}^{G} = \mathbb{E}_{\hat{I}}[-\log D_{src}(\hat{I})]. \tag{8}$$

Secondly, to make $\hat{I}$ be correctly with target attribute c, an attribute classification loss is designed to enforce the attribute prediction of $\hat{I}$ from the attribute classifier approximates the target value c as below:

$$\mathcal{L}_{cls}^{G} = \mathbb{E}_{\hat{I}}[-\log D_{cls}(c|\hat{I})]. \tag{9}$$

Last but not least, to keep the attribute-irrelevant region unchanged, a reconstruction loss is employed similar as CycleGAN [12] and StarGAN [16], which is formulated as follows:

$$\mathcal{L}_{rec}^{G} = \lambda_1 \mathbb{E}_{I,c,c^g}[(\|I - G(G(I, c), c^g)\|_1] + \lambda_2 \mathbb{E}_{I,c^g}[(\|I - G(I, c^g)\|_1], \tag{10}$$

where c^g is the original attribute of input image I, λ_1 and λ_2 are two balance parameters. The first term is dual reconstruction loss. In this loss, when an attribute edited image $\hat{I} = G(I, c)$ is translated back to the image $G(G(I, c), c^g)$ with the original attribute c^g, it is expected to be the same as the original image I. The second term is identity reconstruction loss, which guarantees that an input image I is not modified when edited by its own attribute label c^g. Here, the L1 norm is adopted for more clear reconstruction.

Finally, the overall objective function to optimize G is achieved as below:

$$\min_{F_m, F_a} \mathcal{L}_G = \mathcal{L}_{adv}^{G} + \mathcal{L}_{cls}^{G} + \mathcal{L}_{rec}^{G}. \tag{11}$$

For the whole SaGAN network, the generator G and the discriminator D can be easily optimized in an adversarial way, following most existing GAN-based and CNN-based methods.

2.3 Discussions

Differences with CycleGAN [12]**.** In terms of loss function, CycleGAN and our SaGAN are similar as they both adopt the adversarial loss, the dual reconstruction loss and the identity reconstruction loss, but they differ in the way of generating the attribute editing images. The CycleGAN operates on the whole image to produce an edited image and couples the counter editing of an attribute as a cycle architecture. Differently, our SaGAN introduces spatial attention mechanism to enforce the attribute manipulation only within the attribute-specific regions for more precise attribute editing, and achieves two counter editing via single model but with different conditional signal.

Differences with StarGAN [16]**.** Again, the most significant difference between StarGAN and our SaGAN is that StarGAN operates on the whole image while our SaGAN only focuses on the attribute-specific region. An advantage of StarGAN is that it can edit multiple attributes with one model, while our SaGAN can only edit one attribute which will be our future work.

Differences with ResGAN [17]**.** ResGAN and our SaGAN are the only two methods that aims at manipulating modest region, i.e. attribute-specific region, while keeping the rest remain unchanged. They are different in how to achieve this goal. ResGAN models the manipulation of attribute-specific region as a sparse residual image, which determines the attribute-specific region via the sparsity constraint. The sparsity degree depends on a control parameter but not the attribute itself. Differently, our SaGAN determines the attribute-specific region via an attention mask predicted from the spatial attention network, which is adaptive to the attribute, and thus more accurate than that from the simple sparsity constraint. Besides, ResGAN employs two generators for the counter editing of one attribute, while our SaGAN adopts a single generator but with different conditional signal.

3 Implementation Details

Optimization. To optimize the adversarial real/fake classification more stably, in all experiments the objectives in Eqs. (1) and (8) is optimized by using WGAN-GP [22], reformulated as

$$\mathcal{L}_{src}^{D} = -\mathbb{E}_I[D_{src}(I)] + \mathbb{E}_{\hat{I}}[D_{src}(\hat{I})] + \lambda_{gp}\mathbb{E}_{\tilde{I}}[(\|\nabla_{\tilde{I}} D_{src}(\tilde{I})\|_2 - 1)^2], \tag{12}$$

$$\mathcal{L}_{src}^{G} = -\mathbb{E}_{\hat{I}}[D_{src}(\hat{I})], \tag{13}$$

while $\tilde{I}$ is sampled uniformly along a straight line between the edited images $\hat{I}$ and the real images I. λ_{gp} is the coefficient of the gradient penalty which is empirically set as $\lambda_{gp} = 10$.

Network Architecture. The detailed architectures of our SaGAN are shown in Tables 1 and 2. For the generator, the two networks of AMN and SAN share

Table 1. The network architecture of generator G. I, O, K, P, and S denote the number of input channel, the number of output channel, kernel size, padding size and stride size respectively, and IN denotes the instance normalization.

Layer	Attribute Manipulation Network (AMN)	Spatial Attention Network (SAN)
L1	Conv(I4, O32, K7, P3, S1), IN, ReLU	Conv(I3, O32, K7, P3, S1), IN, ReLU
L2	Conv(I32, O64, K4, P1, S2), IN, ReLU	Conv(I32, O64, K4, P1, S2), IN, ReLU
L3	Conv(I64, O128, K4, P1, S2), IN, ReLU	Conv(I64, O128, K4, P1, S2), IN, ReLU
L4	Conv(I128, O256, K4, P1, S2), IN, ReLU	Conv(I128, O256, K4, P1, S2), IN, ReLU
L5	Residual Block(I256, O256, K3, P1, S1)	Residual Block(I256, O256, K3, P1, S1)
L6	Residual Block(I256, O256, K3, P1, S1)	Residual Block(I256, O256, K3, P1, S1)
L7	Residual Block(I256, O256, K3, P1, S1)	Residual Block(I256, O256, K3, P1, S1)
L8	Residual Block(I256, O256, K3, P1, S1)	Residual Block(I256, O256, K3, P1, S1)
L9	Deconv(I256, O128, K4, P1, S2), IN, ReLU	Deconv(I256, O128, K4, P1, S2), IN, ReLU
L10	Deconv(I128, O64, K4, P1, S2), IN, ReLU	Deconv(I128, O64, K4, P1, S2), IN, ReLU
L11	Deconv(I64, O32, K4, P1, S2), IN, ReLU	Deconv(I64, O32, K4, P1, S2), IN, ReLU
L12	Conv(I32, O3, K7, P3, S1), Tanh	Conv(I32, O1, K7, P3, S1), Sigmoid

Table 2. The network architecture of discriminator D. I, O, K, P, and S denote the number of input channel, the number of output channel, kernel size, padding size and stride size respectively, and IN denotes the instance normalization.

Layer	Discriminator	
L1	Conv(I3, O32, K4, P1, S2), Leaky ReLU	
L2	Conv(I32, O64, K4, P1, S2), Leaky ReLU	
L3	Conv(I64, O128, K4, P1, S2), Leaky ReLU	
L4	Conv(I128, O256, K4, P1, S2), Leaky ReLU	
L5	Conv(I256, O512, K4, P1, S2), Leaky ReLU	
L6	Conv(I512, O1024, K4, P1, S2), Leaky ReLU	
L7	*src*: CONV(I2014, O1, K3, P1, S1)	*cls*: CONV(I1024, O1, K2, P0, S1), Sigmoid

the same network architecture except slight difference in the input and output: (1) AMN takes as input a four-channel tensor, consisting of an input image and a given attribute value, while SAN just takes as input the input image. (2) AMN outputs a three-channel RGB image, while SAN outputs a single channel attention mask image. (3) AMN uses $Tanh$ as the activation function for the output layer as the input image has been normalized to $[-1, 1]$ like most existing GAN methods, while SAN adopts $Sigmoid$ as the attention is within $[0, 1]$. For the discriminator, the same architecture as PatchGAN [12,23] is used considering its promising performance.

Training Settings. The parameters of all models are randomly initialized according to the normal distribution with mean as 0 and standard deviation as 0.02. During optimization of SaGAN, Adam [24] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate $lr = 0.0002$ is adopted as the optimizer. For all of our experiments, We set $\lambda_1 = 20$ and $\lambda_2 = 100$ in Eq. (10). And the batch size is set to 16. The generator is updated once, while the discriminator is updated three times.

4 Experiments

In this section, we firstly illustrate the datasets used for experiments; and then compare our SaGAN against recent methods on face attribute editing in terms of visual performance; and finally demonstrate that our SaGAN can benefit the face recognition by data augmentation.

4.1 Datasets

The CelebA [25] dataset contains 202,599 face images of 10,177 celebrities. Each face image is annotated with 40 binary attributes. The official aligned and cropped version of CelebA are used, and all images are resized to 128×128. The 8,177 people with the most samples are used for training and the rest 2,000 people for testing. In summary, the training data contains 191,649 images, and the testing data contains 10,950 images for evaluation of both face attribute editing and face verification. Besides, LFW [26] dataset is also used for testing the generalization of the proposed SaGAN. Four attributes are used as exemplars for editing, including *eyeglasses*, *mouth_slightly_open*, *smiling* and *no_beard*.

4.2 Visual Comparison on Face Attribute Editing

We first investigate the results of attribute editing and the attention mask generated by SaGAN. Then, we compare the proposed SaGAN with the state-of-the-art methods including CycleGAN [12], StarGAN [16] and ResGAN [17] on face attribute editing. All these methods are trained with the same training data. They are tested on both CelebA and LFW.

Investigation of SAN. The spatial attention network (SAN), aiming at localizing the attribute-specific region which restricts the face attribute editing within this region, plays an important role in the proposed SaGAN. Therefore, we visualize the corresponding spatial attention masks to figure out how SAN contributes to the performance for face attribute editing. As can be seen in Fig. 3, the spatial attention masks mainly concentrate on the attribute-specific regions, and those attribute-irrelevant regions are successfully suppressed. This helps to keep the attribute-irrelevant regions unchanged. For local attribute such as eyeglass, the spatial attention mask only have response around the eyes, while for the attribute that may involve the movement of global face such as mouth open and smiling, the spatial attention have response on larger or even the whole face area. This illustrates that the spatial attention network can adaptively and effectively determine the attribute-specific regions according to the attribute to edit.

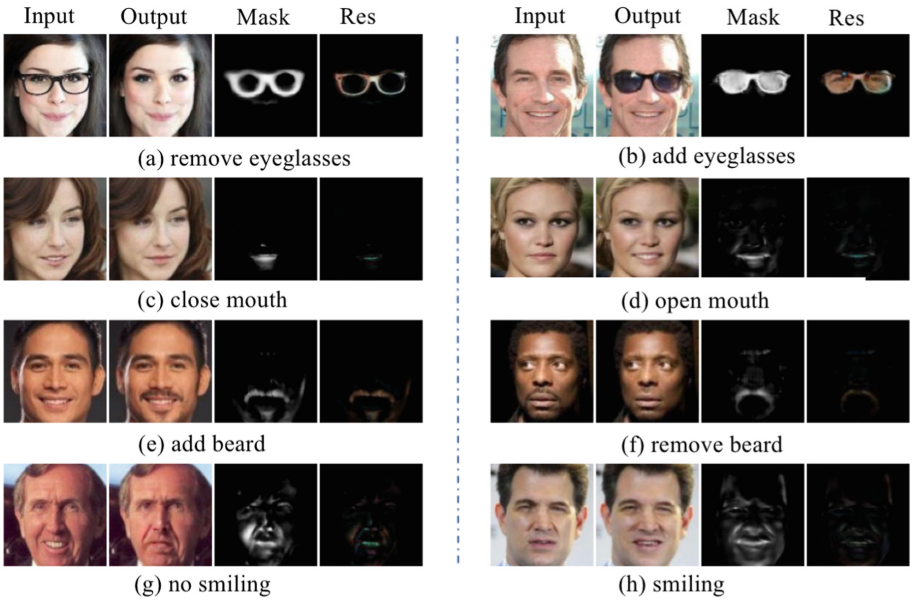

Fig. 3. Face attribute editing of our SaGAN on the CelebA dataset. "Mask" represents the spatial attention mask generated by SAN, while "Res" denotes the residual images.

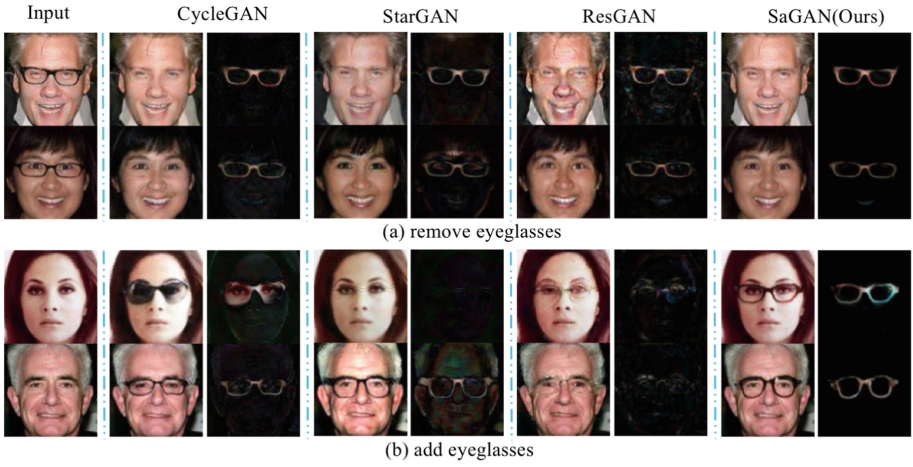

Fig. 4. Face attribute *eyeglasses* editing from different methods on the CelebA dataset.

Visual Results on CelebA. Figures 4 and 5 show the editing results on CelebA dataset for face attribute *eyeglasses* and *mouth_slightly_open* respectively. Compared with CycleGAN and StarGAN, ResGAN and our SaGAN preserves most attribute-irrelevant regions unchanged which is preferable. However, there are some artifacts on the attribute-specific regions from ResGAN especially

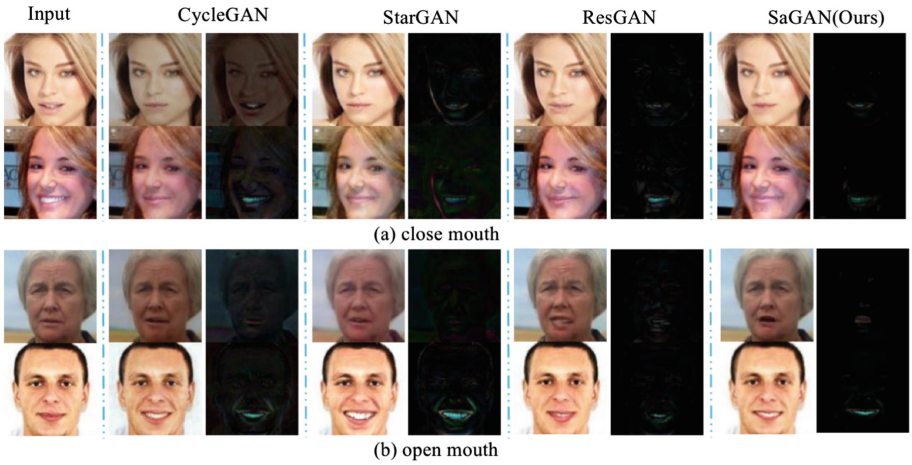

Fig. 5. Face attribute *mouth_slightly_open* editing from different methods on the CelebA dataset.

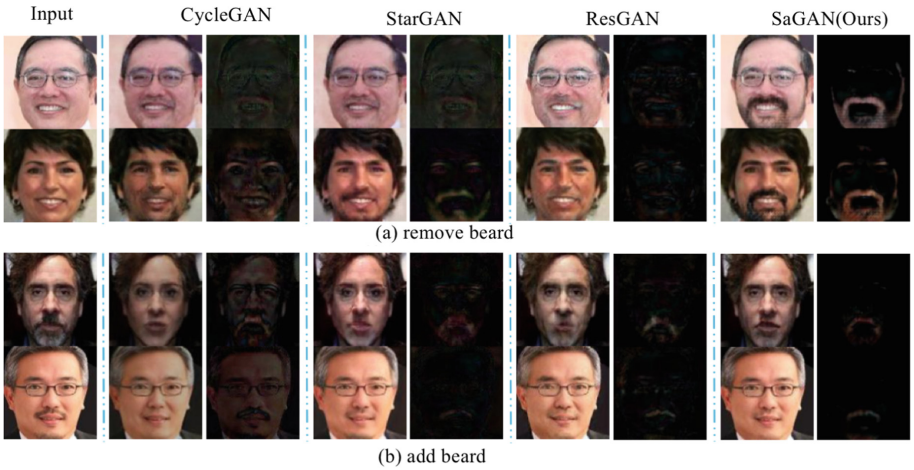

Fig. 6. Face attribute *no_beard* editing from different methods on the CelebA dataset.

on the eyeglass attribute. By contrast, our SaGAN achieves favorable manipulation on the attribute-specific region and preserve the rest irrelevant regions unchanged as well. The reason lies in that the generator of SaGAN contains a spatial attention module SAN for explicitly attribute-specific region detection, which makes the attribute manipulation network only concentrate on how to manipulate regardless of where to manipulate. Figure 6 shows the editing results of *no_beard*, and all methods inevitably change the gender of the input face as *no_beard* is correlated with gender (e.g. no woman has beards). Even so, SaGAN modifies the images modestly, e.g. preserves the most regions beyond cheek and

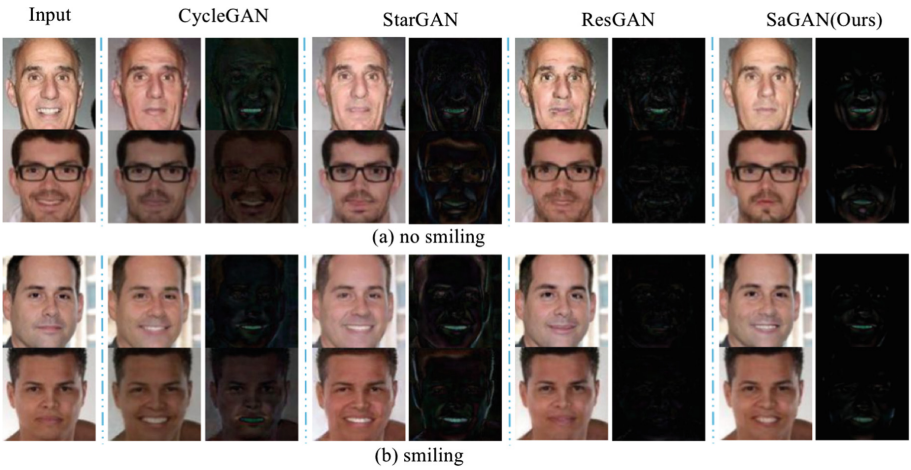

Fig. 7. Face attribute *smile* editing from different methods on the CelebA dataset.

jaw. Figure 7 shows the results of global face attribute *smile*. Not surprisingly, SaGAN achieves better visual quality again demonstrating the effectiveness of the proposed method.

Visual Results on LFW. To investigate the generalization capability of SaGAN, the model trained on CelebA is further evaluated on the LFW dataset as shown in Fig. 9. As can be seen, all methods of CycleGAN, StarGAN and ResGAN degenerate on this dataset with those distorted results in Fig. 9, e.g. CycleGAN changes a male image to a female one after removing the beard. Surprisingly, SaGAN performs almost as good as on the CelebA, illustrating the robustness of our proposed method.

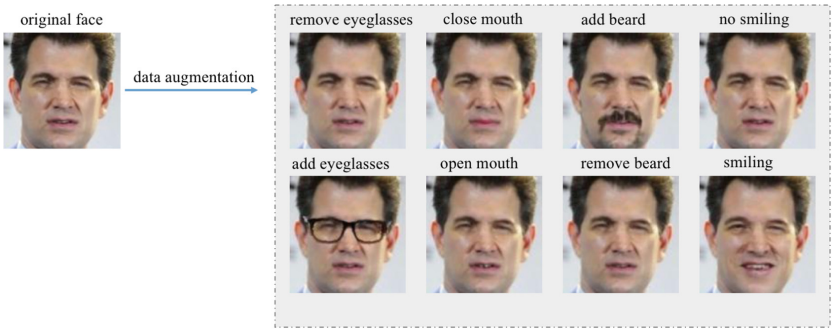

Fig. 8. Data augmentation on CelebA dataset.

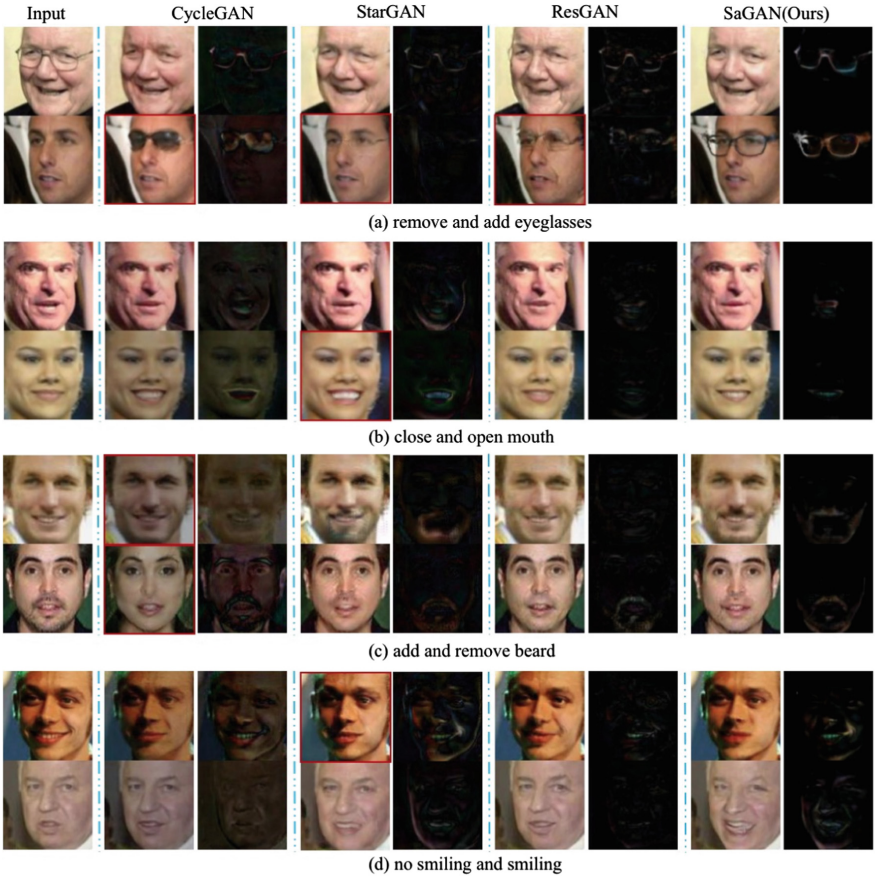

Fig. 9. Face attribute editing from different methods on the LFW dataset.

4.3 Comparison on face recognition

Seeing the favorable visual editing results, a natural idea is whether it is beneficial for face recognition by such as data augmentation. To investigate this, we augment each training sample by modifying the attribute. As shown in Fig. 8, for each attribute, a single training sample is augmented into three samples, e.g. the original face image and the two face images with adding and removing eyeglasses respectively. Actually, a face image edited by its own attribute looks almost the same as the original one, and the reason of augmenting with its original attribute is just for simplicity without the need of classifying the attribute of an image. The ResNet-18 [27] is used as the face recognition model. The testing is conducted on the test sets of CelebA and LFW which are the same as that for face attribute editing. On CelebA, one face image is randomly selected as target and the rest as query. On LFW, the standard protocol is employed. On both datasets, the performance is reported in terms of ROC curves.

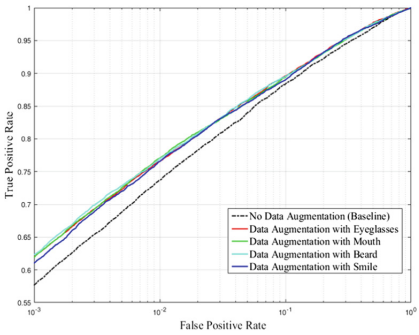

Fig. 10. Face verification on CelebA.

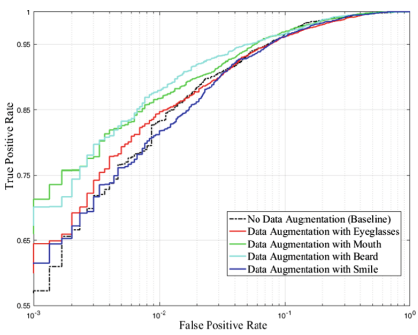

Fig. 11. Face verification on LFW.

Figure 10 shows the face verification results evaluated on CelebA. As can be observed, for each attribute, the model with data augmentation performs better than the baseline model without data augmentation, as the augmentation with accurately attribute editing images from our SaGAN enriches the variations of the training data leading to more robust model. The face verification results evaluated on LFW are shown in Fig. 11. As can be seen, the model with data augmentation with all face attributes expect *smile* are much better than the baseline model without data augmentation similar as that on the CelebA, demonstrating the benefits of our SaGAN for face verification. One possible reason for the slightly worse performance of augmentation with *smile* is that the smile faces in test data are few and the augmentation with *smile* makes the model biased to smile leading to performance degeneration.

5 Conclusions and Future Works

This work introduces the spatial attention mechanism into the GAN framework, forming a SaGAN method for more accurate face attribute editing. This kind of spatial attention mechanism ensures the manipulation of attributes only within the attribute-specific regions while keep the rest irrelevant regions unchanged. Experiments on CelebA and LFW, demonstrate that the proposed SaGAN performs better than the existing face attribute editing methods benefitted from the spatial attention mechanism. Besides, the proposed SaGAN can also benefit the face recognition through data augmentation. In the future, we will try to apply the proposed SaGAN to the general image editing tasks.

Acknowledgement. This work was partially supported by National Key Research and Development Program of China Grant 2017YFA0700804 Natural Science Foundation of China under contracts Nos. 61390511, 61772496 and 61532018.

References

1. Ichim, A.E., Bouaziz, S., Pauly, M.: Dynamic 3D avatar creation from hand-held video input. ACM Trans. Graph. (ToG) **34**(4), 45 (2015)
2. Antipov, G., Baccouche, M., Dugelay, J.L.: Face aging with conditional generative adversarial networks. arXiv preprint arXiv:1702.01983 (2017)
3. Wang, W., et al.: Recurrent face aging. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Ding, H., Sricharan, K., Chellappa, R.: ExprGAN: facial expression editing with controllable expression intensity. In: The Association for the Advance of Artificial Intelligence (AAAI) (2018)
5. Zhu, Z., Luo, P., Wang, X., Tang, X.: Recover canonical-view faces in the wild with deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
6. Zhang, Z., Peng, Y.: Eyeglasses removal from facial image based on MVLR. The Era of Interactive Media. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-3501-3_9
7. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS) (2014)
8. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
9. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems (NIPS) (2016)
10. Berthelot, D., Schumm, T., Metz, L.: BEGAN: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)
11. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802 (2016)
12. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
13. Zhang, X., Yu, F.X., Chang, S.F., Wang, S.: Deep transfer network: unsupervised domain adaptation. CoRR, abs/1503.00591 (2015)
14. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
15. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355 (2016)
16. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. arXiv preprint arXiv:1711.09020 (2017)
17. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)
19. Wang, F., et al.: Residual attention network for image classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

21. Huang, Q., et al.: Semantic segmentation with reverse attention. In: British Machine Vision Conference (BMVC) (2017)
22. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems (NIPS) (2017)
23. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
24. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
26. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07–49, University of Massachusetts, Amherst (2007)
27. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38