Sparsely Grouped Multi-task Generative Adversarial Networks for Facial Attribute Manipulation

Jichao Zhang
Shandong University
zhang163220@gmail.com

Yezhi Shu
Shandong University
shuyz1996@gmail.com

Songhua Xu
Xi'an Jiaotong University
songhua.xu@gmail.com

Gongze Cao
Zhejiang University
asxaqz2@gmail.com

Fan Zhong
Shandong University
zhongfan@sdu.edu.cn

Xueying Qin
Shandong University
qxy@sdu.edu.cn

Abstract

Recently, Image-to-Image Translation (IIT) has achieved great progress in image style transfer and semantic context manipulation for images. However, existing approaches require exhaustively labelling training data, which is labor demanding, difficult to scale up, and hard to adapt to a new domain. To overcome such a key limitation, we propose Sparsely Grouped Generative Adversarial Networks (SG-GAN) as a novel approach that can translate images in sparsely grouped datasets where only a few train samples are labelled. Using a one-input multi-output architecture, SG-GAN is well-suited for tackling multi-task learning and sparsely grouped learning tasks. The new model is able to translate images among multiple groups using only a single trained model. To experimentally validate the advantages of the new model, we apply the proposed method to tackle a series of attribute manipulation tasks for facial images as a case study. Experimental results show that SG-GAN can achieve comparable results with state-of-the-art methods on adequately labelled datasets while attaining a superior image translation quality on sparsely grouped datasets [1].

1. Introduction

Image-to-Image Translation (IIT) aims to learn the mapping from a source image to a target image. Unlike traditional style-transfer methods [8], recent methods leverage Generative Adversarial Networks for learning the end-to-end mapping, e.g. [15]. Supported by the adversarial loss, the discriminator can learn a similarity measure optimized for a specific task rather than using a hand-engineering one, which enables the method to easily adapt to multiple tasks, such as translating color images to edge maps, grayscale images to color images, and labels to street scenes.

Previous IIT methods can be categorized into two broad classes, including *methods that learn from paired training data* and *methods that learn from grouped training data*. The first class of methods that learn from paired data require that each source image in a training set is explicitly associated with a corresponding target image. To collect such training data requires non-trivial labelling efforts, in particular when the data set is sizable. To alleviate this burden in gathering paired training data, the second class of methods that learn from grouped training data are introduced. These methods permit a training dataset to be organized in a way where a group of source images are associated with another group of target images without the need to specify the one-to-one correspondence between the two groups of images, which nevertheless needs human efforts in organizing training data (Left in Fig. 1).

Another limitation of existing image translation approaches is their degraded performance when training data from multiple groups are not balanced. For instance, the age attribute in CelebA [27] is not balanced, where images of younger people are much more than images of older individuals. In terms of unbalanced attributes, the majority group is referred to as MA; while the minority group is referred to as MI. Most of the previous works will suffer this problem [3, 20]. To overcome the problem, a method called ResidualGAN [38] is introduced, which undersamples the MA group to achieve the balance between MA and MI groups before the training step. This undersampling process can easily lose valuable information, whose performance can become particularly problematic when the minority group has a very small size.

To alleviate these problems, we propose a novel sparsely grouped learning method, where only a few of the training

---

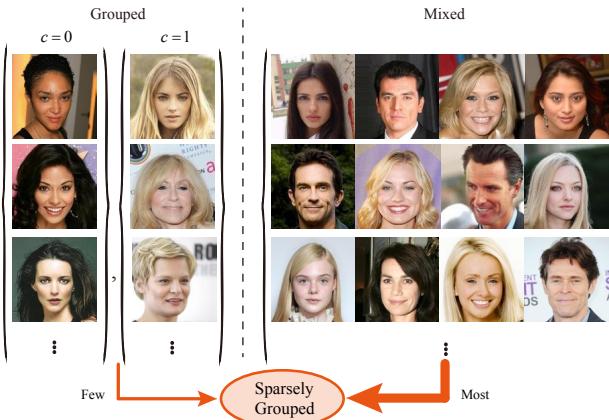[1]Code: https://github.com/zhangqianhui/Sparsely-Grouped-GAN.

Figure 1: A sparsely grouped dataset consists of a most portion of mixed data and a few quantity of group-labelled data. Here, grouped training dataset consists of two groups, one for black hair with $c = 0$ and another for blond hair with $c = 1$.

dataset is grouped while the remaining unlabelled data (i.e. mixed data) as shown in Fig. 1 is used for unsupervised learning to improve the performance of classification and stabilize the training of the adversarial network. Therefore, our method is actually semi-supervised learning. The sparsely grouped learning will work well for unbalanced data in IIT task, as unused data can be fully utilized for unsupervised learning after following up undersampling procedure to balance MA group and MI group. To the best of our knowledge, no previous image translation architecture can be directly applied to learn from sparsely grouped datasets as an off-the-shelf tool. To address the gap, we propose a one-input multiple-output network, called *Sparsely Grouped Generative Adversarial Networks(SG-GAN)*, for learning to translate a diverse set of image attributes among multiple groups.

Overall, the main contributions include:

- We propose SG-GAN, a novel generative adversarial network for tackling image-to-image translation tasks by learning mapping among multiple groups in a sparsely grouped dataset where only a few portion of data points are associated with their group-labels while the group affiliation for the most of data points remains unknown.

- The proposed SG-GAN can generate comparable facial attribute translation results using much fewer group-labelled samples than peer methods; SG-GAN also outperforms state-of-the-art methods for facial attribute manipulation working with severely unbalanced dataset.

- We further introduce an adapted residual image learning component into the proposed SG-GAN to improve

the degree of translation for the targeted image attribute involved in a translation process while preserving other visual attributes unrelated to the translational goal in the generation results.

2. Related Work

Generative Adversarial Networks: In addition to the variational autoencoder(VAE) [19] and PixelCNN [42], Generative Adversarial Networks(GAN) [9] provides a more powerful framework for generating sharp and realistic images [36, 48, 32]. Recently, [16] proposed a new training methodology for GAN to progressively train the generator and discriminator. Their method can generate highly realistic facial images of 1024^2 pixels. With its rapid development and optimization, GAN has been applied to many fields, for example, image in-painting [34, 13], image super-resolution [21], style transfer [22, 50], video prediction [29, 33, 24] and object detection [43].

Researchers have developed a rich collection of techniques to improve both the training stability and diversity of images generated by GAN. To attain more stable training, objective functions of GAN are carefully designed. For example, LS-GAN [28] adopts a least squared loss function in its discriminator to solve the vanishing gradient problem. The Wasserstein GAN(WGAN) proposed in [1] uses the Wasserstein distance instead of the Jensen-Shannon distance to form its objective function, which achieves a more stable training process. For the latter aim, [37] proposed a method to facilitate the convergence of a GAN by adopting virtual batch normalization to replace batch normalization [14].

Image-to-Image translation(IIT): The essence of IIT is to learn the mapping between pairs or groups of images while preserving image characteristics irrelevant to the current translation task. The prior work of IIT using conditional GAN [15] has attained impressive results. Their method applies a supervised learning-based approach onto pairs of IIT images in the training phase, which incurs a major limitation for large-scale application and migration into new domains and tasks. To overcome the aforementioned limitation, a collection of IIT methods based on learning from group-labelled training data was proposed, e.g., [47, 25, 5, 51, 17, 38, 7, 26], to reduce the ground truth group-label acquisition efforts for training data preparation.

In addition to directly learning the mapping from one or multiple source images to one or multiple corresponding target images, another thread of active research endeavors is to conduct disentangled representation learning, the result of which can then be leveraged for facial attribution manipulation in images [20, 6, 2, 39, 45]. Recently, Star-GAN [3] is proposed, which utilizes a GAN-based architecture to learn a series of mappings from a common group of source images to multiple groups of target images using a
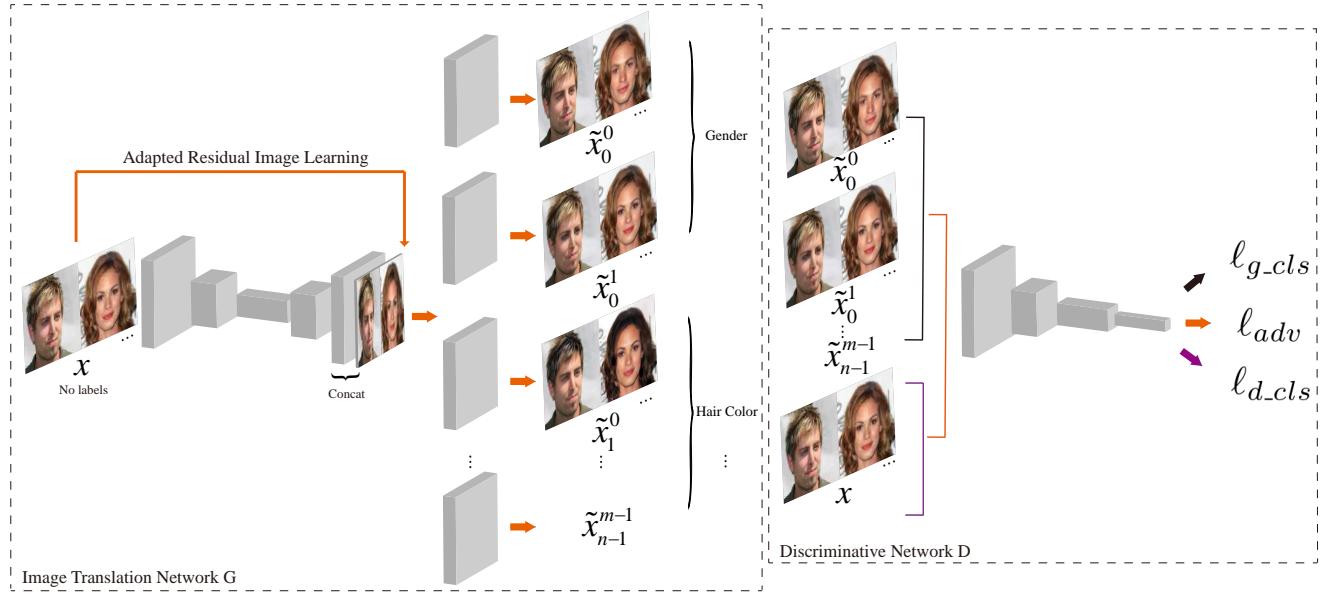
Figure 2: The multi-task learning architecture of SG-GAN. G is a one-input, multi-output network that performs multiple attributes manipulation(e.g., gender(male/female), hair color(black/blond)). Adapted residual image learning just concatenates input x into the middle feature maps of decode network as the input of the next convolution layer. Note that D will be a semi-supervised classifier when trained in the sparsely grouped dataset.

single generative model. In comparison, the new SG-GAN algorithm proposed in this paper requires a much smaller amount of ground truth group-labels in its learning process, thanks to its semi-supervised learning framework.

Facial Attribute Manipulation: Facial attribute manipulation is a special IIT task, which aims at modifying the semantic content of a facial image according to a specified attribute value. [20] proposed a novel model, called VAE/GAN, for facial attribute manipulation but acquires labelled data to compute the visual attribute vectors after training. Another model called Adversarially Learned Inference [6] is based on conditional GAN, which must be embedded with binary attributes when trained for the image semantic translation task. [49] proposed a model called ST-GAN, which can be trained on a mixed dataset to establish relationships between latent codes and generated samples for semantic information discovery. Even though the design approach of their method is inspiring, the quality of its IIT result yet needs to be further improved.

Recently, GAN-based residual image learning has been applied to facial attribute manipulation, the method of which is referred to as ResidualGAN [38]. ResidualGAN attains satisfactory IIT results. In comparison, the proposed SG-GAN can obtain multiple facial attribute manipulation effects using only a single trained model, which also produces more visually realistic IIT results.

Semi-Supervised Learning using GAN: A substantial amount of efforts have been dedicated to conducting semi-supervised learning using GAN, e.g., [37, 4, 40, 46, 31]. For example, the method introduced in [4] improves the per-

formance of semi-supervised image classification in several benchmark datasets. For these semi-supervised GAN models, discriminator D will receive three different sources of data in its training process: real labelled images for supervised learning, real unlabelled images and fake images for unsupervised learning.

3. Methods

3.1. Generative Adversarial Networks

Goodfellow [9] proposes a GAN that consists of one generative model G and another discriminator model D. Its training process can be treated as a minimax game in which D learns to distinguish between real samples and generated samples and at the same time G tries to learn to generate samples $G(z)$ from the random noise z to match the distribution of real samples x and fool D. The objective function of GAN is given as follows:

$$\min_G \max_D \ell(D, G) = \mathbb{E}_x[logD(x)] + \mathbb{E}_z[log(1 - D(G(z)))] \quad (1)$$

3.2. SG-GAN

One-Input Multiple-Output Architecture for Multi-task Learning: Unlike the vanilla GAN model [9], which directly learns the mapping from a noise vector z to images x, the G in IIT task learns the mapping from an input image x to an output image $\tilde{x}$, which can be regarded as an "autoencoder" [12].

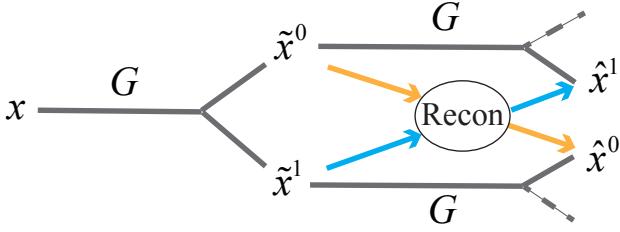The previous methods have two corresponding generator networks with input images from different groups [3, 51].

Figure 3: The reconstruction learning process in SG-GAN ($m = 2$).

However, such architecture with two generators can not work well in the sparsely grouped dataset where most data are mixed while only a few of data is group-labelled. It is hard to learn a good mapping from a source group to a target group using few group-labelled samples. Recently, StarGAN [3] uses a single generator network for manipulating multiple attributes. It is difficult to train their model on a sparsely grouped dataset because of the lack of original domain labels to assess the reconstruction loss.

As illustrated in Fig. 2, we propose a one-input multi-output architecture as generator G, which can use an unlablled data point x as the input and maps it to all existing groups. It is obvious that both labelled and unlabelled images can be used as input to G. The discriminator D is a multi-task learning network, which performs adversarial-learning and classified-learning with multiple facial attributes. The final output of G is $\tilde{x}_j^i, 0 \leq i < m, 0 \leq j < n$, where m is the range of every attribute value and n is the number of facial attributes. Except for adversarial learning, the D network can be regarded as multiple classifiers with outputting n logit vectors where every vector with m-dim is used as input to a softmax function. The output dimensions of G and D will increase as the number of facial attributes participating in the manipulation grows.

Sparsely Grouped Learning: We construct a discriminative network for classification and to distinguish generated samples $\tilde{x}$ and real samples x. The D for classified-learning need to classify input samples by the attribute value and build the objective function for every attribute. Given the real image x with the original group c_j for the j-th facial attribute. The objective function of D for this attribute is the softmax loss:

$$\ell_{d\_cls} = \mathbb{E}_{x,c_j}[-\log(D(c_j|x)], \quad (2)$$

where $D(c_j|x)$ is the softmax probability over this group-label c_j. In the same way, the generated samples $\tilde{x}$ with the target group $\tilde{c}_j$. The softmax objective function for training G is:

$$\ell_{g\_cls} = \mathbb{E}_{\tilde{x},\tilde{c}_j}[-\log(D(\tilde{c}_j|\tilde{x})] \quad (3)$$

Different from the previous GAN model, SG-GAN would output multiple generated samples $\tilde{x}_j^i, i = 0, 1, ..., m - 1$ for the j-th facial attribute. The min-max

adversarial loss for generator G and discriminator D with this facial attribute is:

$$\begin{aligned} \ell_{adv} &= \mathbb{E}_x[logD(x)] \\ &+ \sum_{i=0}^{m-1}[\mathbb{E}_{\tilde{x}_j^i}[log(1 - D(\tilde{x}_j^i))]] \end{aligned} \quad (4)$$

Noted that the D trained with grouped data carries both loss terms ℓ_{d_cls} and ℓ_{adv} in its objective function, while the network trained with mixed data only computes a single loss term ℓ_{adv} in its objective function

Adapted Residual Image Learning: The residual image learning that was proposed by [38] aims to improve the effectiveness of facial attribute manipulation and make the modest modification of the attribute-specific facial area while keeping irrelevant content unchanged. However, this model with the residual image learning, which sums the natural images and outputs of network linearly is hard to generate very realistic images, especially based on the vanilla GAN loss [9]. Through experiments, we found that learned residual images tend to be sparse when using very powerful GAN loss terms [10].

To alleviate these problems, we improve the previous residual image learning method from two aspects. One is to use the concatenation of the feature maps and input x instead of their sums, the other is that we add the new convolution layer to refine the input as the final translation result. As shown in Fig. 2, we name this novel method as adapted residual image learning and adapt it to our architecture.

Reconstruction Loss: By minimizing the adversarial and classification losses, G is trained to generate images that are realistic and classified into the correct target group. However, minimizing these losses does not guarantee that translated images preserve the identity and background content of the input image. To alleviate this problem, the previous methods apply a cycle consistency loss [51, 3] to the generator. This loss does not adapt to our architecture. As shown in Fig. 3, we propose a reconstruction loss for the generator, defined as:

$$\begin{aligned} \ell_{rec} &= \|G(G(x)^0)^1 - G(x)^1\|_1 \\ &+ \|G(G(x)^1)^0 - G(x)^0\|_1, \end{aligned} \quad (5)$$

where $G(x)^i$ means the i-th output of the G. G uses x as input to obtain the translation result $\tilde{x}^0$ and $\tilde{x}^1$. Then, these results will also be new input to obtain new translation result $\hat{x}^1$ and $\hat{x}^0$ of corresponding groups. We adapt this L_1 norm as the loss to make $\tilde{x}^0$ and $\hat{x}^0$, $\tilde{x}^1$ and $\hat{x}^1$ as close as possible. Besides keeping the content of images fixed for image translation among different groups, our reconstruction loss could also keep the content fixed for images translation between the same group.

Overall Objective Function: The objective functions will be different for the mixed data and grouped data, as

4

Figure 4: Facial attribute translation results on CelebA test dataset. The first and second rows show facial attributes manipulation results of the baseline methods, i.e. ResidualGAN and StarGAN; the third row shows results of the proposed model in the condition that all data is grouped; the results are shown in the last two rows when using 5000 and 500 images for every value of the attribute as the grouped data, respectively. G: gender; S: smile; H: hair color.

ℓ_{d_cls} is just for the grouped data with group-labels. Finally, the full objective functions for the specific facial attribute to optimize D is shown as:

$$\ell_D = \begin{cases} \ell_{d\_cls} - \ell_{adv} & \text{Grouped,} \\ -\ell_{adv} & \text{Mixed.} \end{cases} \quad (6)$$

For G, the objective function with this facial attribute is:

$$\ell_G = \ell_{g\_cls} + \ell_{adv} + \alpha \ell_{rec}, \quad (7)$$

where α is a hyper-parameter for reconstruction loss. We use $\alpha = 10$ in all our experiments.

4. Implementation

Network Architecture: Our architecture is similar to previous IIT methods, which have shown impressive results for style transfer and semantic manipulation. Our generator G contains convolution layers with the stride size of two for encoding, some residual blocks for expanding the receptive field, transposed convolution layers for decoding. We use instance normalization [41] for the generator but no normalization for the discriminator. Note that $tanh$ activation function is used for output of the generator. More details about network architecture are shown in the Appendix 7.1.

Training Details: We apply new technique from Wasserstein GAN with gradient penalty (WGAN-GP) [10] to stabilize training process and generate high quality images. We replace Eq. 4 with the new object function of WGAN-GP defined as:

$$\ell_{adv} = \mathbb{E}_x[D(x)] - \sum_{i=0}^{m-1}(\mathbb{E}_{\tilde{x}_j^i}[D(\tilde{x}_j^i)] - \lambda\mathbb{E}(t_j^i)), \quad (8)$$

where t_j^i is a gradient penalty variable for training D and more details can be found in [10]. λ is 10 for all our experiments.

We use the Adam [18] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The size of the training batch is set to 8 for our experiments. Similar to WGAN-GP [10], the generator is updated once after every five updates performed over the discriminator. Our all models are trained with the learning rate of 0.0001 for the first 10000 iterations and the learning rate will be linearly decayed to 0 over the next 10000 iterations.

Coping with Unbalanced Dataset: SG-GAN starts with an undersampling procedure to balance the majority and minority groups. However, rather than discarding data through an undersampling process by existing methods, the proposed method uses a follow-up semi-supervised learning procedure where unsampled data elements are observed
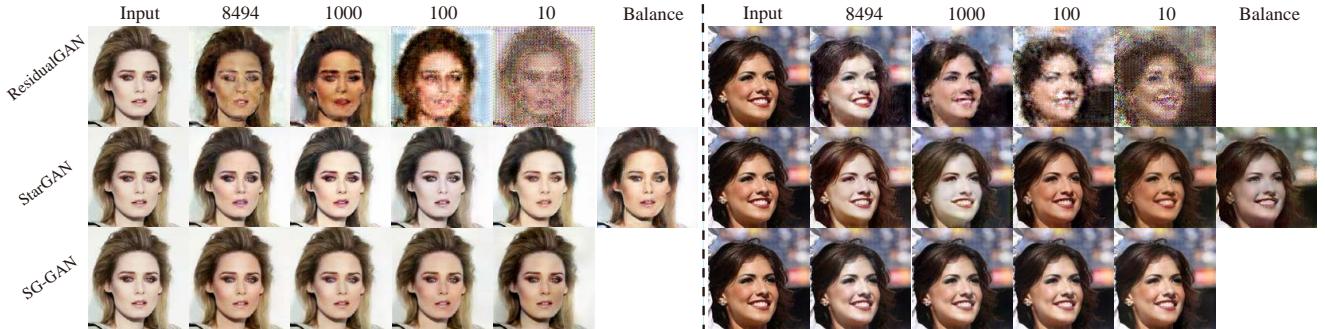
Figure 5: The translation results of the unbalanced facial attribute on CelebA test dataset. Three rows show results using different models: ResidualGAN, StarGAN and SG-GAN, respectively. And the first column shows input images while the next columns are generated using the different number of the MI group for pale skin. Results of the dual task for the same attribute manipulation are shown in the right.

as unlabelled records during a learning process. In this way, all dataset elements are effectively utilized for model training, without overlooking informational clues carried by any one of them.

5. Experiments

We examine our proposed model, SG-GAN, on a sparsely grouped face dataset. In this section, we firstly compare SG-GAN against recent methods on facial attribute manipulation tasks via both qualitative and quantitative evaluation. Next, we analyze the problems of image translation for previous works, demonstrate results that SG-GAN can achieve more realistic and apparent translation results in the unbalanced data. Lastly, we conduct an ablation study and compare the proposed method against several reduced variants to show the effectiveness of the adapted residual image learning component in the method.

5.1. Baseline Models

For multiple facial attributes manipulation, Star-GAN [3] has achieved state-of-the-art results compared with DIAT [23], ICGAN [35], and CycleGAN [51]. Therefore, we compare the proposed model against StarGAN.

We also adopt ResidualGAN [38] as a baseline, which performs attribute transfer with the residual image learning. Both StarGAN and ResidualGAN belong to *Grouped* methods, in which they acquire group-labelled training data. More importantly, ResidualGAN acquires an equal number of images between the MI group and the MA group for the highly unbalanced attribute during its training process. Unlike ResidualGAN, StarGAN only uses group-labelled data, in which it does not adopt specific measures for coping with unbalanced training data. It is worth noting that for multi-attribute manipulation tasks, we must train Residual-GAN many times. In this comparative study, we implement ResidualGAN on our own and use the official code of Star-

GAN [2].

5.2. Dataset

The CelebFaces Attribute Dataset (CelebA) [27] contains 202,599 face images with large pose variations and background clutter. We cropped and scaled each image to 128×128 pixels. 5000 images are randomly selected as the test dataset with the remaining images used as the training dataset. In our experiments, we would select and divide the facial attributes into balanced parts, e.g., gender, hair color (black and blond hair), smile and unbalanced parts, e.g., pale skin. For balanced attributes, we train a model for multiple attributes manipulation. Note that, we train a single model for every unbalanced attribute.

5.3. Baseline Comparison

In this section, we provide comparison results with the baseline methods in balanced attributes transfer tasks and use some metrics to evaluate the translation results from multiple perspectives.

Qualitative evaluation: Fig. 4 shows facial attribute manipulation results generated by SG-GAN for three sparsely grouped datasets, which respectively have 500 samples, 5000 samples for every value in the attribute, and the full dataset group-labelled. To distinguish these models, we denote them as SG-GAN(500), SG-GAN(5000), SG-GAN(All) respectively. As shown in the 3rd to the 5th rows of Fig. 4, the quality of translation results does not noticeably decline when the number of group-labelled attributes is reduced.

In comparison with ResidualGAN, our method attains a higher visual quality of translation results. One important reason is that ResidualGAN uses the vanilla GAN loss which may not provide the stable gradient in its training process. Another is that training a model to perform a fixed translation is prone to overfit [3]. The proposed model us-

---

[2]Please see https://github.com/yunjey/StarGAN

6

ing the WGAN-GP loss term does not suffer from the same problem when applied to translate multiple image attributes.

Furthermore, compared to StarGAN, the proposed model demonstrates its superior advantage in keeping the unrelated content fixed, for example, image background. It can be interpreted that the adapted residual image learning component as adopted by the proposed model can effectively help the information translation process and chooses a good initial point for its training process.

Quantitative Evaluation Protocol: Through qualitative evaluation, the advantages of the proposed method in attaining a high image translation quality and preserving other visual attributes unrelated for the translation task are further demonstrated.

To assess quantitatively the sample quality and degree of translation, we compute the classification accuracy using the ResNet-18 [11], which is the same evaluation network used in StarGAN [3]. We notice that the classification accuracy of translation results concerning the targeted attributes does not consider the preservation of unrelated image attributes in a translation process. Therefore, we not only compute accuracy on targeted attributes but also other unrelated facial attributes to explore whether there is any side effect of accidentally modifying visual attributes unrelated in a translation process. Here we select three attributes: gender, smile and hair color in this experiment.

Additionally, to show the ability keeping the consistency of irrelevant content, for example, image background, we use MS-SSIM [44] to measure the similarity in image background between the translation results and input samples. A higher MS-SSIM value corresponds to a higher similarity between images in human perception. For specifically, we crop the 10×10 top-left corner from both translation results and input samples as the background region.

Quantitative evaluation: As shown in the first column of Table 1, we give the classification accuracy of attribute gender on translation results of attribute gender, which indicates SG-GAN achieves an acceptable degree of translation.

The accuracies of other attributes, smile and hair color, on this translation results are shown in the second and third columns. In the case of attribute smile, SG-GAN(All) achieves the accuracy of 86.87, higher than both baseline models. For hair color, our model and baseline models have the similar accuracy. It indicates that our model and StarGAN have some advantages in generated images with respect to the maintenance of identity in general. Additionally, our model is more capable to keep background fixed, e.g, 72 for SG-GAN(All), 70 for SG-GAN(500), 64 for StarGAN and 63 for ResidualGAN, showing in the last column. More importantly is that SG-GAN(500) and SG-GAN(5000) obtains closed scores compared with SG-GAN(ALL) in most of cases. Table 2 reports scores on

Table 1: Classification accuracy [%] of attribute gender transfer results for different models.

Attribute	Gender	Smile	Hair Color	Background
ResidualGAN	95.49	85.80	95.90	63
StarGAN	96.23	85.99	**99.63**	64
SG-GAN(ALL)	**99.03**	86.87	97.99	**72**
SG-GAN(5000)	93.79	**87.34**	98.42	71
SG-GAN(500)	93.36	84.87	97.90	70
CelebA	99.00	90.22	99.53	100

Table 2: Classification accuracy [%] of attribute hair color translation results for different models.

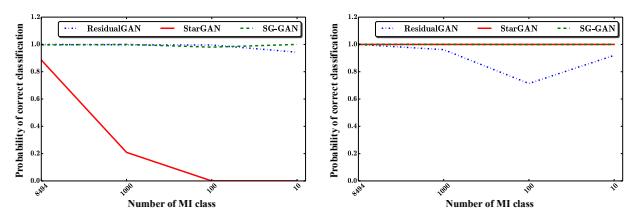| Attribute	Gender	Smile	Hair Color	Background
ResidualGAN	83.15	86.07	93.41	18
StarGAN	**99.43**	88.26	**99.53**	63
SG-GAN(ALL)	91.38	87.42	98.99	**73**
SG-GAN(5000)	89.58	87.36	95.72	72
SG-GAN(500)	89.11	**88.68**	88.91	69
CelebA	99.00	90.22	99.53	100

Figure 6: Probability of correct classification for the samples shown in Figure 5 in different models. As shown in the left, it is obvious that StarGAN would mistake the sample from MI group as MA group when the unbalance between MI and MA group is very serious.

translating hair color as the targeted attribute, which produce the similar conclusion.

5.4. Translation experiments on unbalanced attributes

In this subsection, we will show the translation results on unbalanced attributes (i.e. pale skin). As shown in Fig. 5, our method and baseline methods give out translation results for attribute pale skin on the training dataset with the different number of the MI group. Concerning the pale skin attribute on the CelebA dataset, the numbers of positive and negative labels in the training dataset are 8494 and 189106 respectively.

For ResidualGAN, requiring the same number of MI group and MA group, its translation results lack quality and visibility, worse and worse with the decrease in the number of MI group. Obviously, ResidualGAN could not generate realistic samples when just using 10 images for single group (1-th row of Fig. 5).

StarGAN does not require the balance between the number of images between the MA group and MI group. As shown in the 2-th row of Fig. 5, StarGAN achieves excessive translation results from MA group to MI group of pale skin, but not obvious translation conversely. As shown in Fig. 6, the reason of this phenomenon is that StarGAN

	Input	Gender	Hair Color	Pale Skin
StarGAN				
SG-GAN				

(a)
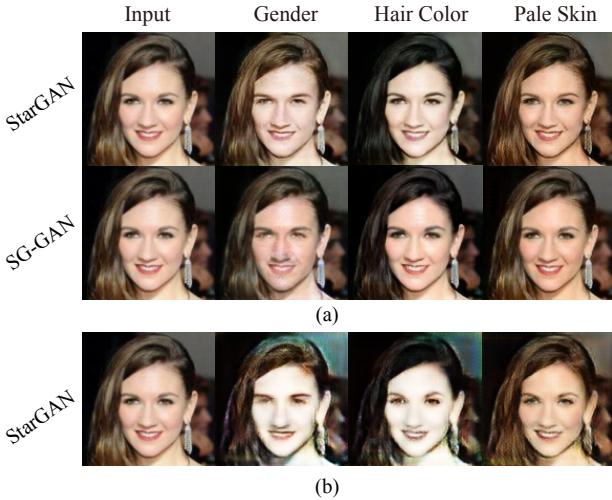
StarGAN				

(b)

Figure 7: More comparison between StarGAN and SG-GAN(ALL): (a) both are trained on three attributes, balanced: gender, hair color; unbalanced: pale skin. (b) The facial attribute manipulation results of StarGAN, when it is trained on a sparsely grouped dataset without using the reconstruction loss.
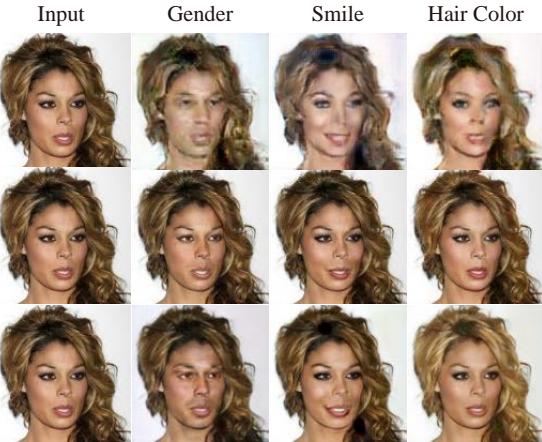

Figure 8: The validation of effectiveness for adapted residual image learning. First row is the multiple attributes manipulation results without residual learning; 2rd is the results with the original residual image learning method; 3rd is the results with the proposed adapted residual image learning.

does not have fine classifier after training on the unbalanced dataset and mistakes the samples of MI group for MA group. StarGAN is based on conditional GAN [30] and it requires the group-label of the target domain c when images translate from the source domain to the target domain. However, the translation results from the the MI group to the MA group could not be obvious when G of StarGAN thinks some samples from the MI group have the same domain with the MA group. The more obvious translation results have been shown in Fig. 5 (See 5th column) when Star-

Table 3: Classification accuracy of translated results on SG-GAN(500) using residual image learning or adapted residual image learning.

Method	Gender	Smile	Hair Color
Residual Image Learning	18.77	23.67	21.35
Adapted Residual Image Learning	**93.26**	**82.67**	**88.91**

GAN has been trained in the balanced dataset, which contains 8494 images from MA group and 8494 images from MI group.

Similar to ResidualGAN, SG-GAN requires the balance between MI group and MA group. But the difference is that SG-GAN, which works well on sparsely grouped datasets, can make full use of unlabelled data to stabilize the training of networks. When the MI group is very small, SG-GAN still generates very high-quality samples. Compared with StarGAN, SG-GAN can be trained to obtain a fine classifier. As shown in the 3rd row of Fig. 5, SG-GAN could achieve high-quality results and very obvious translation effect, whether it is translating the attribute of pale skin from the MI group to MA group or in the reverse direction. We notice that image translation results of SG-GAN are also satisfactory when the number of MI group are very small, e.g. having only 10 data points in our experiments.

Because of requiring the multi-attribute group-labels as the target domains, StarGAN would suffer the serious problems that the manipulation results for the single attribute is easy to be affected by other attributes, especially when trained on the unbalanced attribute. As shown in Fig. 7(a), the attribute pale skin affects the manipulation results of other attributes. When trained on the same dataset, however, SG-GAN does not suffer from this problem. As mentioned above, it is hard for StarGAN to be trained on the sparsely grouped dataset, because of the lack of original domain labels to assess the reconstruction loss. As shown in Fig. 7(b), StarGAN without the reconstruction loss could not obtain the high-visual quality translation results when trained on the sparsely grouped dataset. These comparison results show the superiority of our architecture.

5.5. Ablation Study

In this section, we validate the effectiveness of the adapted residual image learning component in the proposed method by comparing translation results generated by SG-GAN with several of its variants. As shown in Fig. 8, without residual image learning, SG-GAN can only generate low-quality images that appear blurry and fail to preserve visual characteristics uninvolved in the translation task. In comparison with the original residual image learning method [38], the adapted residual image learning component improves the degree of images translation. For the original residual learning, the summation between the real images and the network output would be the final translation result. However, the learned residual image tends to be-

come sparse and the output becomes close to the input when the adversarial loss has sufficient capacity. The proposed method solves this by concatenating the real images and the feature maps of network D instead of summing them up linearly. As shown in the Table 3, the quantitative evaluation validates the effectiveness of this design.

6. Conclusion

We have proposed a new model, i.e. SG-GAN, to perform multiple facial attributes manipulation with one-input multi-output architecture, which is well suited for tackling multi-task learning and sparsely grouped learning tasks. Tightly coupled with this learning architecture, an adaptive residual image learning paradigm has been shown to enhance the performance of the new method in image translation. As consistently demonstrated results reported in the paper, the proposed method is able to attain comparable image translation results as multiple state-of-the-art peer methods while having access to significantly fewer training group-labels. Moreover, the experiments show that SG-GAN is able to consistently achieve more apparent translation results over the dataset for unbalanced attributes. It is interesting and valuable to apply SG-GAN for more IIT tasks, for example, age progression and regression of facial image.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[2] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

[4] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov. Good semi-supervised learning that requires a bad gan. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 6513–6523. Curran Associates, Inc., 2017.

[5] H. Dong, P. Neekhara, C. Wu, and Y. Guo. Unsupervised image-to-image translation with generative adversarial networks. *arXiv preprint arXiv:1701.02676*, 2017.

[6] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. C. Courville. Adversarially learned inference. *CoRR*, abs/1606.00704, 2016.

[7] Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, and L. Carin. Triangle generative adversarial networks. *CoRR*, abs/1709.06548, 2017.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. 313:504–7, 08 2006.

[13] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, 2017.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[20] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[21] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[22] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[23] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *CoRR*, abs/1610.05586, 2016.

[24] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua. Attentive moment retrieval in videos. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24. ACM, 2018.

[25] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017.

[26] S. Liu, Y. Sun, D. Zhu, R. Bao, W. Wang, X. Shu, and S. Yan. Face aging with contextual generative adversarial nets. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 82–90, New York, NY, USA, 2017. ACM.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[28] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[29] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[30] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[31] T. Miyato, A. M. Dai, and I. Goodfellow. Adversarial Training Methods for Semi-Supervised Text Classification. *ArXiv e-prints*, May 2016.

[32] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3510–3520. IEEE, 2017.

[33] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 1789–1798, New York, NY, USA, 2017. ACM.

[34] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[35] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible Conditional GANs for image editing. In *NIPS Workshop on Adversarial Training*, 2016.

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

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[38] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1225–1233. IEEE, 2017.

[39] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages –. IEEE, 2017.

[40] J. T. Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *CoRR*, abs/1511.06390, 2015.

[41] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. 07 2016.

[42] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[43] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[45] T. Xiao, J. Hong, and J. Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. *arXiv preprint arXiv:1803.10562*, 2018.

[46] Z. Yang, J. Hu, R. Salakhutdinov, and W. W. Cohen. Semisupervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1040–1050, 2017.

[47] W. Yin, Y. Fu, Y. Ma, Y.-G. Jiang, T. Xiang, and X. Xue. Learning to generate and edit hairstyles. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 1627–1635, New York, NY, USA, 2017. ACM.

[48] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[49] J. Zhang, F. Zhong, G. Cao, and X. Qin. St-gan: Unsupervised facial image semantic transformation using generative adversarial networks. In M.-L. Zhang and Y.-K. Noh, editors, *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 248–263. PMLR, 15–17 Nov 2017.

[50] Y. Zhao, B. Deng, J. Huang, H. Lu, and X.-S. Hua. Stylized adversarial autoencoder for image generation. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 244–251, New York, NY, USA, 2017. ACM.

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

7. Appendix

7.1. Network Architecture

The network architecture of SG-GAN are shown in Table 4 and Table 5. Because SG-GAN is a multi-task learning framework, it can manipulate multiple attributes at the same time, we only choose an attribute to show the architecture. Here are some notations should be noted. n_c: channel of results. n_t: range of value for the attribute. h: height of input images. w: width of input images. C: channels of images. K: size of the kernel. S: size of the stride. P: padding method. D: the scale of resize.

Part	Input Shape	Operation	Output Shape
encoder	(h, w, n_c)	CONV-(C64, K7×7, S1×1,P_{same}), ReLU, Instance Normal	$(h, w, 64)$
	$(h, w, 64)$	CONV-(C128, K4×4, S2×2,P_{same}), ReLU, Instance Normal	$(\frac{h}{2}, \frac{w}{2}, 128)$
	$(\frac{h}{2}, \frac{w}{2}, 128)$	CONV-(C256, K4×4, S2×2,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
bottleneck	$(\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block:CONV-(C256,K3×3,S1×1,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
	$(\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block:CONV-(C256,K3×3,S1×1,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
	$(\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block:CONV-(C256,K3×3,S1×1,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
	$(\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block:CONV-(C256,K3×3,S1×1,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
	$(\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block:CONV-(C256,K3×3,S1×1,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
	$(\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block:CONV-(C256,K3×3,S1×1,P_{same}), ReLU, Instance Normal	$(\frac{h}{4}, \frac{w}{4}, 256)$
decoder	$(\frac{h}{4}, \frac{w}{4}, 256)$	DECONV-(C128,K4×4,S2×2,P_{same}), ReLU, Instance Normal	$(\frac{h}{2}, \frac{w}{2}, 128)$
	$(\frac{h}{2}, \frac{w}{2}, 128)$	DECONV-(C64,K4×4,S2×2,P_{same}), ReLU, Instance Normal	$(h, w, 64)$
	$(h, w, 64)$	CONCAT	$(h, w, 64 + 3)$
	$(h, w, 64 + 3)$	CONV-(C(n_c),K7×7,S1×1,P_{same})	(h, w, n_c)

Table 4: Generator architecture

Part	Input Shape	Operation	output
discriminator	(h, w, n_c)	CONV-(C64, K5×5, S2×2,P_{same}), Leaky ReLU	$(\frac{h}{2}, \frac{w}{2}, 64)$
	$(\frac{h}{2}, \frac{w}{2}, 64)$	CONV-(C128, K5×5, S2×2,P_{same}), Leaky ReLU	$(\frac{h}{4}, \frac{w}{4}, 128)$
	$(\frac{h}{4}, \frac{w}{4}, 128)$	CONV-(C256, K5×5, S2×2,P_{same}), Leaky ReLU	$(\frac{h}{8}, \frac{w}{8}, 256)$
	$(\frac{h}{8}, \frac{w}{8}, 256)$	CONV-(C512, K5×5, S2×2,P_{same}), Leaky ReLU	$(\frac{h}{16}, \frac{w}{16}, 512)$
	$(\frac{h}{16}, \frac{w}{16}, 512)$	CONV-(C512, K5×5, S2×2,P_{same}), Leaky ReLU	$(\frac{h}{32}, \frac{w}{32}, 512)$
	$(\frac{h}{32}, \frac{w}{32}, 512)$	CONV-(C1024, K5×5, S2×2,P_{same}), Leaky ReLU	$(\frac{h}{64}, \frac{w}{64}, 1024)$
D_{cls}	$(\frac{h}{64}, \frac{w}{64}, 1024)$	CONV-(C(n_t), K2×2, S1×1,P_{valid})	$(\frac{h}{128}, \frac{w}{128}, n_t)$
D_{adv}	$(\frac{h}{64}, \frac{w}{64}, 1024)$	CONV-(1, K3×3, S1×1,P_{same})	$(\frac{h}{64}, \frac{w}{64}, 1)$

Table 5: Discriminator architecture

7.2. Additional Qualitative Results

Input	Gender	Smile	Hair Color	Lipstick	G+S	G+H	S+H

Figure 9: Single and multiple attribute translation results on CelebA using method SG-GAN(ALL).

| Input | Gender | Smile | Hair Color | Lipstick | G+S | G+H | S+H |

13
Figure 10: Single and multiple attribute translation results on CelebA using method SG-GAN(500).