Translate the Facial Regions You Like Using Self-Adaptive Region Translation

Wenshuang Liu[1,2,3], **Wenting Chen**[1,2,3], **Zhanjia Yang**[1,2,3], **Linlin Shen**[1,2,3*]
[1] Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2] Shenzhen Institute of Artificial Intelligence & Robotics for Society
[3] Guangdong Key Laboratory of Intelligent Information Processing
{liuwenshuang2018, chenwenting2017, yangzhanjia2019}@email.szu.edu.cn, llshen@szu.edu.cn

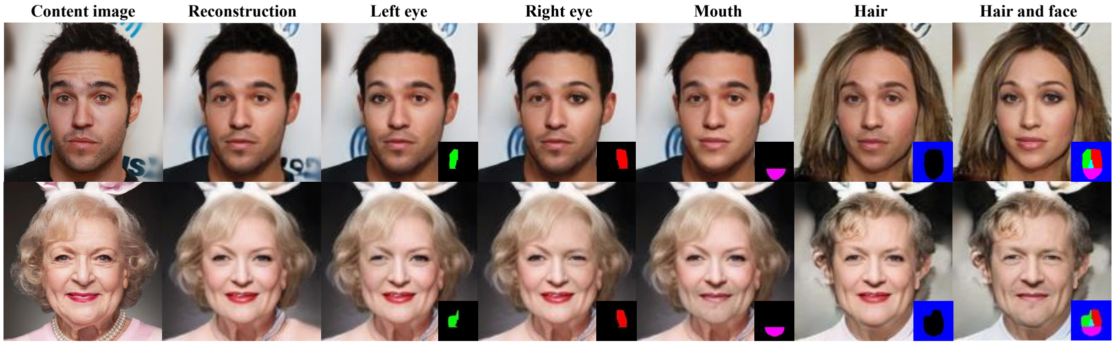


Figure 1: Eyes, mouth, hair and both hair&face translation for example faces in the CelebAMask-HQ dataset.

Abstract

With the progression of Generative Adversarial Networks (GANs), image translation methods has achieved increasingly remarkable performance. However, most available methods can only achieve image level translation, which is unable to precisely control the regions to be translated. In this paper, we propose a novel self-adaptive region translation network (SART) for region-level translation, which uses region-adaptive instance normalization (RIN) and a region matching loss (RML) for this task. We first encode the style and content image for each region with style and content encoder. To translate both shape and texture of the target region, we inject region-adaptive style features into the decoder by RIN. To ensure independent translation among different regions, RML is proposed to measure the similarity between the non-translated/translated regions of content and translated images. Extensive experiments on three publicly available datasets, i.e. Morph, RaFD and CelebAMask-HQ, suggest that our approach demonstrate obvious improvement over state-of-the-art methods like StarGAN, SEAN and FUNIT. Our approach has further advantages in precise control of the regions to be translated. As a result, region level expression changes and step-by-step make-up can be achieved. The video demo is available at (https://youtu.be/DvIdmcR2LEc).

Introduction

Generative Adversarial Networks (GANs) has been widely used for image generation, image super-resolution (Karras et al. 2017; Karras, Laine, and Aila 2019) and image translation (Zhu et al. 2017a; Choi et al. 2018, 2020; Liu et al. 2019; Jiang et al. 2020; Deng et al. 2020). Though achieved remarkably success in transferring complex appearances across different domains, most image translation methods can only translate faces at image level. As shown in Fig.2(a), for image level translations approaches like StarGAN(Choi et al. 2018), AttGAN(He et al. 2019) and FUNIT(Liu et al. 2019), the same translation is usually applied to different face regions and they can only transfer the whole image from domain A to B. When image level approaches change the attribute of a specific region, e.g. hair style and skin tone, using conditional attribute labels, other regions could easily be changed as well, during the image level translation. In contrast, our region level approach can adaptively translate the styles of different regions to various domains like B, C, D and E, without affecting other regions.

Recently, some region-level translation methods have been proposed. While Jiang and Deng (Jiang et al. 2020; Deng et al. 2020) use unsupervised attention to solve this task, Zhu (Zhu et al. 2020) applied semantic image synthesis to generate different regions. However, the former fails to precisely control the translated region and distinguish the boundaries between translated and untranslated regions as masks are not used to enforce constraints on the generation of images. Fig. 2(b) shows the differences between our approach and semantic image synthesis methods like SPADE (Park et al. 2019) and SEAN (Zhu et al. 2020). Given a face mask with specified regions like eyes, nose, mouth, face and

[*] Corresponding Author: Linlin Shen.

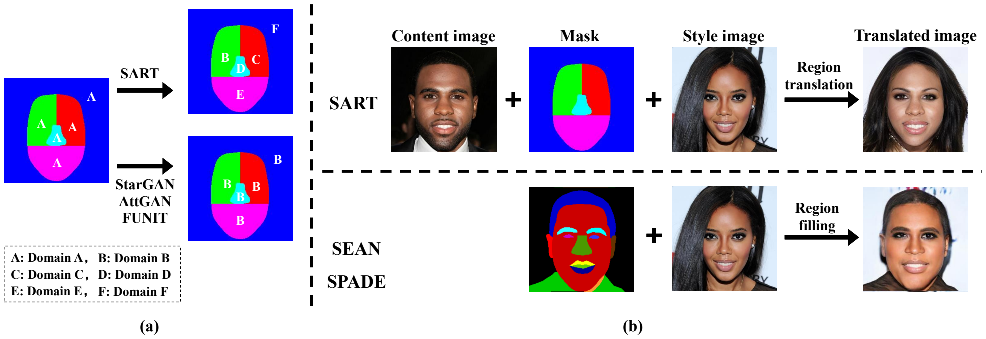


Figure 2: Differences between SART and other GAN based methods.

hair, SPADE and SEAN mainly generate the content of these regions according to the styles of the style face. As shown in the 2^{nd} row of Fig. 2(b), the regions of the generated face have exactly the same shape with that of the mask. For example, although the style image consists of a lady's face with long hair, the hair of the generated face is short, due to the constraint of the hair region in the face mask. In contrast, our approach adaptively translate the eyes, nose, mouth and hair of a man to the style of lady's face. While the face still looks similar to the man in the content image, the eyes, nose, mouth and hair now look much more like a beautiful lady, due to the generation of eye shadow, lip stick and long hair.

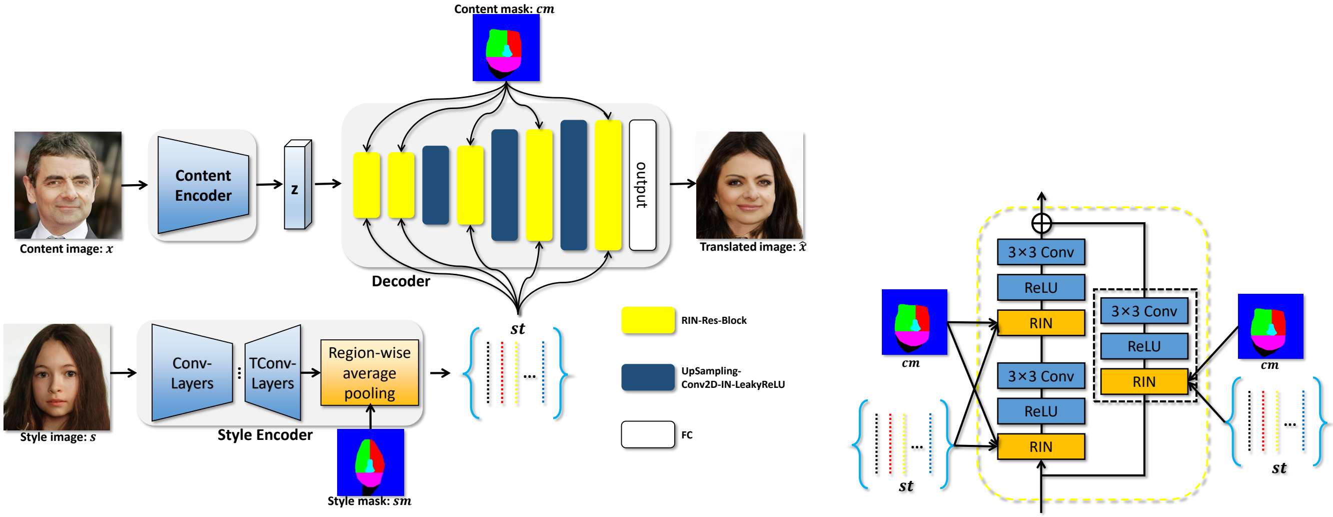
We propose in this paper a self-adaptive region translation network (SART), for region level face translation. Our work mainly presents two contributions: region-adaptive instance normalization (RIN) and region matching loss (RML). To translate different regions of a face into various styles, RIN firstly extracts the feature map and style code of each region from the content and style images, respectively, and then adaptively inject the styles into the feature maps of corresponding regions. To make sure that the regions are translated separately, i.e. the translation of certain regions does not affect other regions, we propose RML to measure the similarity between the non-translated/translated regions of content and translated images. Our approach is extensively tested on three publicly available datasets, i.e. Morph (Ricanek and Tesafaye 2006), RaFD (Langner et al. 2010) and CelebAMask-HQ (Lee et al. 2020; Karras et al. 2017; Liu et al. 2015). The results are quantitatively evaluated using metrics like Accuracy, FID (Frechet Inception Distance) and LPIPS (Learned Perceptual Image Patch Similarity). Both visual and quantitative results suggest that our approach demonstrate a large improvement over state-of-the-art methods like StarGAN, SEAN and FUNIT. The idea of our work can be summarized as below:

- We propose a novel self-adaptive region translation network (SART) for region level translation, which can change the styles of target regions and keep the styles of other regions at the same time.
- We propose a region-adaptive instance normalization (RIN) module to adaptively translate the styles (both shape and texture) of target regions for a given face.
- We propose a region matching loss (RML) to ensure region level translation, such that the translation of certain regions does not affect other regions.
- Experimental results quantitatively and qualitatively proves that SART achieves the-start-of-art performance on three publicly available datasets, i.e. Morph , RaFD and CelebAMask-HQ.

Related Work

Image-to-Image Translation. Image-to-image translation is an umbrella concept that can be used to describe many problems in computer vision and computer graphics. As Generative Adversarial Networks (GANs) becoming increasingly mature in image generation, super resolution (Karras et al. 2017; Karras, Laine, and Aila 2019) and text-to-image generation (Xu et al. 2018; Hong et al. 2018), it also achieves remarkable performance in image translation (Chen, Shen, and Lai 2019; Chen et al. 2020; W. Liu 2020). As a milestone, Isola et al. (Isola et al. 2017) first showed that conditional GANs can be used as a general solution to various image-to-image translation problems. Since then, their method has been extended by several works to scenarios including unsupervised learning (Liu, Breuel, and Kautz 2017; Zhu et al. 2017a), few-shot learning (Liu et al. 2019), high resolution image synthesis (Wang et al. 2018), multimodal image synthesis (Zhu et al. 2017b; Huang et al. 2018) and multi-domain image synthesis (Choi et al. 2018, 2020; Chen et al. 2019). However, as illustrated in Fig. 2 (a), all the methods mentioned above can only translate at image level and are not suitable to separately transfer each region from domain A to other domains.

Region level translation. Region level translation can be mainly classified into three categories: label-guided methods, attention based methods and mask-guided methods. Though designed for image level translation, StarGAN(Choi et al. 2018) and AttGAN(He et al. 2019) can transfer face regions related to attributes specified by the targeting labels. While region information is not available in the labels, these approaches generally translate the whole face and related regions based on the translations learned from the training images collected from different domains. Such a translation is not regionally separated and thus can not be precisely controlled. For attention based methods, r-FACE (Deng et al. 2020) takes an example-guided attention mod-



(a) The generator of the proposed SART.

(b) RIN Res-Block.

Figure 3: Detailed architecture of our proposed SART.

ule to fuse attention features and the target face component features extracted from the reference image. PSGAN (Jiang et al. 2020) proposes an Attentive Makeup Morphing module that adaptively morphs the makeup matrices to source images, which can robustly transfer pose and expression. Nonetheless, the boundaries of the translated image are always blurry and changed unexpectedly. Thus, geometry-guided methods are introduced to address this problem by providing semantic segmentation mask. To further translate the texture of style image for each region, SEAN encodes the style of each region to the spatially varying normalization parameters. Then, per region style can be transferred by applying these parameters. However, these works can only fill texture of a given style into the semantic face masks and can not transfer both the shape and texture, as demonstrated in Fig. 2 (b). As content image is not involved, only the texture of the style image can be presented.

Method

In this paper, we focus on region level face translation. The proposed framework, named self-adaptive region translation (SART), aims to individually translate the styles of different regions in a content image. In the following sections, we will show the details of system framework, RIN and RML.

System Framework

As shown in Fig. 3(a), our generator network architecture consists of content encoder, style encoder and decoder. Our generator takes four input images (content image x, content mask cm, style image s and style mask sm) and outputs a translated image $\hat{x}$. The process of generation can be represented as:

$$\begin{aligned} z &= CE(x) \\ st &= SE(s, sm) \\ \hat{x} &= De(z, cm, st) \end{aligned} \tag{1}$$

where CE is the content encoder, SE is the style encoder, De is the decoder, st is the style tensor encoded by style encoder, cm is the content mask with R regions, $\hat{x}$ is the translated image.

Style encoder. As shown in Fig. 3(a), our encoder employs a bottleneck structure to remove the information irrelevant to styles from the style image. The feature map extracted by $TConv-Layers$ (transposed convolution) will be passed through a region-wise average pooling module to get style tensor st. Each vector in st corresponds to one region in style mask. In implementation, we first transform style mask into one-hot tensor where each channel represents a region. Take a channel representing hair region for example, while the values of pixels in hair region are set as 1, others are set to 0. A set of R style feature maps can then be obtained by element-wise multiplication between feature map and different one-hot channels. Finally, we use a global average pooling to get style tensor st, which consists of style information of the R regions.

Decoder. As shown in Fig. 3(a), the decoder is composed of five $RIN-Res$ blocks, three $UpSampling$ blocks and one $Fully-Connected$ layer. As shown in Fig. 3(b), our proposed $RIN-Res$ block consists of three convolutional layers, three $ReLU$ layers and three RIN blocks. Each RIN residual block takes three inputs: content feature maps, per-region style tensor st and content mask. Note that the input content mask is downsampled to the same height and width of the feature maps at the beginning of each RIN block.

Discriminator. Our discriminator is a PatchGAN discriminator (Isola et al. 2017), which utilizes the Leaky ReLU nonlinearity and employs no normalization. The discriminator consists of one convolutional layer followed by 10 activation first residual blocks (Mescheder, Geiger, and Nowozin 2018).

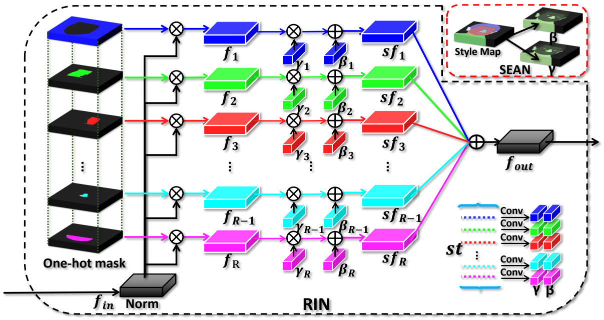


Figure 4: Region-adaptive Instance Normalization.

Region-adaptive Instance Normalization

As shown in the red box of Fig. 4, the style injection branch of SEAN applies the same tensors of gamma and beta to adjust the weight and bias of content feature map for normalization, however, our RIN applies adaptive gamma and beta for different regions.

Given a style tensor st encoding R region styles, the segmentation mask of content image cm and input feature map f_{in}, our RIN tries to translate each of the R regions in content image to the corresponding style specified in the tensor st, by region-adaptive instance normalization. As shown in Fig. 4, we fist multiply (element-wise) feature map f_{in} with the one-hot masks (channels) to get per-region feature maps $\{f_1, f_2, \ldots, f_R\}$, which are then modulated by the adaptive normalization parameters learned from the style tensor st. Let f_{in} denote the input feature map of the current RIN block in a deep convolutional network for a batch of N samples, H, W and C be the height, width and channel numbers of the feature map, the style feature map of the ith region at site $f_i^{n,c,y,x}$ $(n \in N, c \in C, y \in H, x \in W)$ can be represented as:

$$f_i^{n,c,y,x} = \frac{f_{in}^{n,c,y,x} - \mu^c}{\sigma^c} \times cm[i] \tag{2}$$

where $f_{in}^{n,c,y,x}$ denote the feature map at the site before normalization, $cm[i]$ denotes the one-hot mask corresponding to the ith region, μ^c and σ^c are the mean and standard deviation of the feature map in channel c.

After getting the per-region feature map of content image, with the same operation as AdaIN (Huang and Belongie 2017), we do the element-wise calculation between the per-region feature map and its corresponding adaptive regional modulation parameters γ and β extracted by st:

$$sf_i^{n,c,y,x} = f_i^{n,c,y,x} \times (1 + \gamma_i) + \beta_i \tag{3}$$

where $sf_i^{n,c,y,x}$ denotes style feature map for the ith region, γ_i and β_i are the adaptive modulation parameters learned from the ith channel of st.

By now, the per-region feature maps have been all injected with per-region styles encoded from style image, using our region-adaptive instance normalization. Finally, the R modulated per-region feature maps are added together to get the output feature map:

$$f_{out}^{n,c,y,x} = \sum_i sf_i^{n,c,y,x} \tag{4}$$

Region Matching Loss

As shown in Fig. 5, we first use a content image x and a style image s to generate a face $\hat{x}$ presenting similar expression with s, which can be represented as:

$$\begin{aligned} st_t &= SE(s, sm) \\ \hat{x} &= De(CE(x), st_t, cm) \end{aligned} \tag{5}$$

where st_t is the style tensor encoded from the R regions of style image s and $\hat{x}$ is the result image where all R regions have been translated to the per-region styles encoded in st_t. In the second task, we only translate the style of ith region of the content image x, by replacing the ith channel of its style tensor, with that of st_t:

$$\begin{aligned} st_r &= SE(x, cm) \\ st_r[i] &= st_t[i] \\ \hat{r} &= De(CE(x), st_r, cm) \end{aligned} \tag{6}$$

where $st_r[i]$ and $st_t[i]$ are the ith channel of style tensor st_r and st_t, respectively, which encode the style of the ith region of x and s, $\hat{r}$ represent the result image by translating the style of ith region of content image x.

Given a content image x and the fully translated image $\hat{x}$ and partially translated $\hat{r}$, we design a region matching loss to measure the similarity between the ith regions of $\hat{x}$ and $\hat{r}$, and the similarity between other regions of x and $\hat{r}$:

$$\mathcal{L}_{RM} = E_{x,\hat{r},\hat{x}}[||\hat{r} - \hat{x}||_1^1 \times cm[i] + ||\hat{r} - x||_1^1 \times \sum_{j \neq i} cm[j]] \tag{7}$$

where $cm[i]$ and $cm[j]$ represent the one-hot mask corresponding to the ith and jth regions, respectively.

The Full Objective Functions

The proposed SART was trained by solving a minimax optimization problem given by

$$\begin{aligned} \min_D \max_G \mathcal{L}_{GAN}(D, G) + \lambda_R \mathcal{L}_R(G) + \\ \lambda_{FM} \mathcal{L}_{FM}(G) + \lambda_{RM} \mathcal{L}_{RM}(G) \end{aligned} \tag{8}$$

where $\mathcal{L}_{GAN}$, $\mathcal{L}_R$, $\mathcal{L}_{FM}$ and $\mathcal{L}_{RM}$ are the GAN loss, the content image reconstruction loss, the feature matching loss and the region matching loss, respectively. The GAN loss is a conditional one given by

$$\begin{aligned} \mathcal{L}_{GAN}(D, G) = E_x[-log D^{c_x}(x)] + \\ E_{x,\{y_1,\ldots,y_k\}}[log(1 - D^{c_y}(\hat{x})] \end{aligned} \tag{9}$$

The content reconstruction loss helps G learn a translation model. Specifically, when using the same image for both the input content image and the input style image, the loss encourages G to generate an output image identical to the input

$$\mathcal{L}_R(G) = E_x[||x - G(x, cm, \{x, cm\})||_1^1] \tag{10}$$

The feature matching loss regularizes the training. We first construct a feature extractor, referred to as D_f, by removing the last (prediction) layer from D. We then use D_f to extract features from the translation output $\hat{x}$ and the style image $\{y_1, ..., y_k\}$ and minimize

$$\mathcal{L}_{FM}(G) = E_{x,\{y_1,\ldots,y_k\}}[||D_f(\hat{x}) - \sum_k \frac{D_f(y_k)}{k}||_1^1] \tag{11}$$

The GAN loss, the content reconstruction loss and the feature matching loss are the same as that of FUNIT.

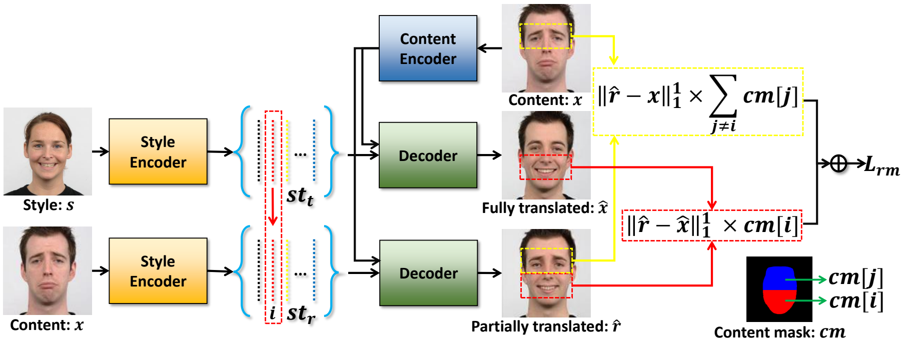

Content Encoder
Style Encoder
Style Encoder
Decoder
Decoder
Style: s
Content: x
Content: x
st_t
st_r
i
...
Fully translated: $\hat{x}$
Partially translated: $\hat{r}$
$\|\hat{r} - x\|_1^1 \times \sum_{j \neq i} cm[j]$
$\|\hat{r} - \hat{x}\|_1^1 \times cm[i]$
$\oplus \rightarrow L_{rm}$
$cm[j]$
$cm[i]$
Content mask: cm

Figure 5: The overview of region matching loss.

Experiment

Our proposed SART was evaluated on three challenging datasets, i.e. Morph, RaFD, CelebAMaskHQ. A wide range of quantitative metrics including FID, Accuracy and LPIPS were evaluated among different models; Qualitatively, the examples of synthesized images are shown for visual inspection.

Datasets

Morph. The Morph dataset (Ricanek and Tesafaye 2006) is a large-scale public longitudinal face dataset, collected in indoor office environment with variations in age, pose, expression and lighting conditions. It has two subsets: Album1 and Album2. Album 2 contains 55,134 images of 13,000 individuals with age label ranging from 16 to 77 years old. We divide the images into a training set with 50020 images and a test set with 4,925 images. The images are separated into five groups with ages of 11-20, 21-30, 31-40, 41-50 and 50+.

RaFD. The RaFD dataset (Langner et al. 2010) is a high-quality face database, containing a total of 67 models with 8,040 pictures displaying 8 emotional expressions, i.e., angry, fearful, disgusted, contempt, happy, surprise, sad and neutral. Each expression consists of three different gaze directions and was simultaneously photographed from different angles using five cameras. We divide the images into a training set with 4,320 images and a test set with 504 images.

CelebAMask-HQ. The CelebAMask-HQ dataset (Lee et al. 2020; Karras et al. 2017; Liu et al. 2015) containing 30,000 segmentation masks for the CelebAHQ face image dataset. We divide the images into a training set with 25,000 images and a test set with 5,000 images.

Metrics

In the training stage of three datasets above, we train different GAN models with their training set. Note that all the baselines are trained with the batch size of 4, the image size of 128×128 and the maximum iteration of 100,000. We perform all traning runs on NVIDIA DGX with one Tesla V100 GPU using Pytorch 1.1.0 and cuDNN 7.4.2.

In the test stage, we evaluate performance of different models on their test set using three metrics as follows:

Accuracy. Three classifiers (Resnet-18) (He et al. 2016) trained using three training sets of different datasets are used to test accuracy of translation. If the synthetic face of target class is correctly classified by the classifier, we decide such translation as a successful one.

FID. Calculated as the Frechet inception distance (Heusel et al. 2017) between two feature distribution of the generated and real images, FID score has been shown to correlate well with human judgement of visual quality. We use the ImageNet-pretrained Inception-V3 (Szegedy et al. 2016) classifiers as feature extractor. For each test image from a source domain, we translate it into a target domain using 10 style images randomly sampled from the test set of the target domain. We then compute the FID between the translated images and training images in the target domain. We compute the FIDs for every pair of image domains and report the average score.

LPIPS. Learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018) measures the diversity of the generated images using the L1 distance between features extracted from the pretrained AlexNet (Krizhevsky, Sutskever, and Hinton 2012). For each test image, we translate its style with reference to 10 style images randomly sampled from the target domain. The L1 distances between each pair of translated image and the style image are then averaged as the LPIPS of the test image. Finally, we report the average of the LPIPS values over all test images. Note that LPIPS is not available for StarGAN, as it does not require any style image for face translation.

Results on the Morph Dataset

Firstly, the SART is evaluated on Morph dataset to assess region level age attribute translation. Note that the per-region styles are encoded using 10 style images randomly sampled from the test set of the target age groups. Fig. 6 shows the translation results of an example face of a 25 years old man. In the first row, the hair of the young man is translated to the styles of different age groups (long black to short white), with fixed face regions. In the second row, the face of the young man is translated to the styles of different age groups (appearance of wrinkles), with fixed hair style. In the third row, both hair and face are translated. One can visually observe that our SART can well control the regions to be trans-

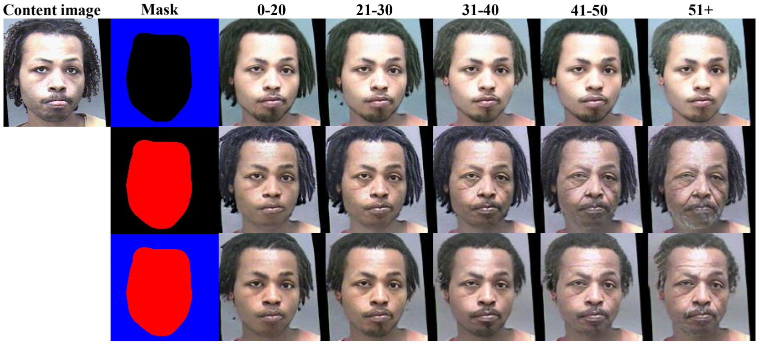


Figure 6: Translate hair or/and face into styles of different age groups for an example face in the Morph dataset.

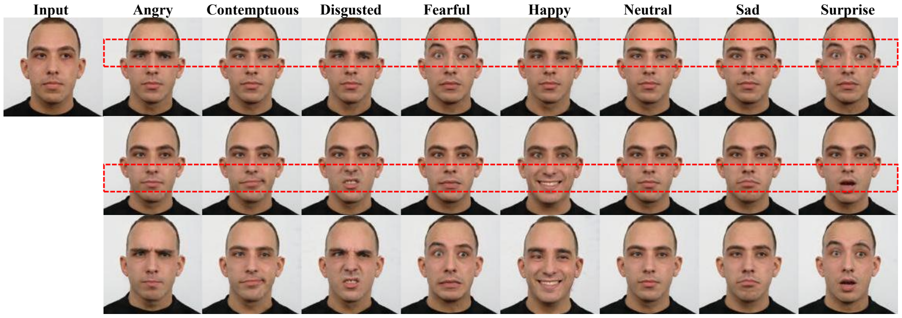


Figure 7: Translate eyes or/and mouth into different expressions for an example face in the RaFD dataset.

lated and achieve decent styles for target regions.

The accuracy, FID and LPIPS of face images translated by our SART are listed in Table 1, together with that translated by StarGAN, SEAN and FUNIT. One can observe from the table that the accuracy of SART is as high as 69.01%, which is significantly higher than that of StarGAN, SEAN and FUNIT. Also, our method achieves the lowest FID score and LPIPS among these GAN based models.

Results on the RaFD Dataset

We now test the performance of our SART for region level expression translation using RaFD dataset. Fig. 7 shows the translation result of an example face in RaFD, whose eyes or/and mouth are translated from neutral to different expressions like angry, fearful, happy, sad and surprise, etc. In the first and second rows, only the eyes and mouth of the man are respectively translated to different expressions, with other regions fixed. In the third row, both eyes and mouth are translated. One can observe from the figure that our approach can precisely translate the shape and texture of designated facial regions to a target expression, without touching any other regions.

Table 1 shows the accuracy, FID and LPIPS of faces generated by different GAN models. One can observe from the table that the accuracy of SART is as high as 88.32%, which is significantly higher than that of StarGAN and more than 75% higher than that of SEAN and FUNIT. Also, our method achieves the lowest FID of 27.88, which is 13.79 lower than that of FUNIT. Though the FID of SEAN is close to our SART, the expressions translated by SEAN is not accurate, due to the fixed shape defined in the semantic mask.

Fig. 10 in Appendix presents more example faces with different expressions translated by StarGAN, SEAN, FUNIT and our SART, which clearly justify the advantages of our approach, in terms of the visual quality of generated face images.

Results on the CelebAMask-HQ Dataset

We now evaluate the region level gender translation performance of our approach using CelebAMask-HQ dataset. Fig. 11 in Appendix shows the intermediate results for face/hair translation of a man to the styles of a lady. With fixed face, the first row present the intermediate results for hair translation from short/black to long/golden. In contrast, the second row show the intermediate results for face translation from man to lady, with fixed hair style. Fig. 1 further shows the results of a man and lady when their left/right eyes, mouths, hairs and full images are translated to the styles of opposite genders. Again one can observe that our model can precisely translate the style of region controlled by the mask overlaid on the bottom right corner of the generated faces, without touching other regions.

Table 1 lists the accuracy, FID and LPIPS of different approaches. Again, our SART achieves the highest accuracy (97.06%) and lowest FID (31.06) and LPIPS (0.3450).

Fig. 12 in Appendix shows the translation of left/right eyes, nose, mouth and faces to the style of a beautiful lady. When the five regions are translated one by one, one can

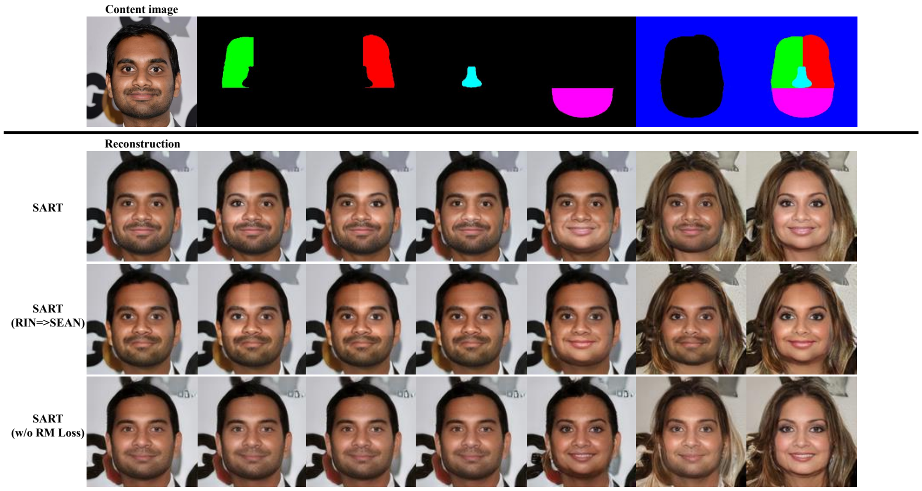


Figure 8: The visual results for ablation studies.

Datasets	Morph			RaFD			CelebAMask-HQ		
Method	Acc(%)↑	FID↓	LPIPS↓	Acc(%)↑	FID↓	LPIPS↓	Acc(%)↑	FID↓	LPIPS↓
StarGAN	60.88	27.89	-	77.28	32.67	-	62.17	47.53	-
SEAN	30.25	48.84	0.2525	13.10	29.61	**0.2610**	72.95	61.06	0.3465
FUNIT	39.02	26.14	0.3152	12.72	41.67	0.2937	93.30	35.17	0.3781
SART	**69.01**	**23.34**	**0.2512**	**88.32**	**27.88**	0.2776	**97.06**	**31.06**	**0.3450**

Table 1: The quantative results of the GAN based models on different datasets.

Method	Acc(%)↑	FID↓	LPIPS↓
SART(RIN⇒SEAN)	96.85	35.01	0.3494
SART(w/o RML)	96.61	33.81	**0.3421**
SART	**97.06**	**31.06**	0.3450

Table 2: The quantative results for ablation studies.

clearly see the make up effects like eye-shadow and whitening of the skin, which beautify the faces to make the ladies look more attractive.

Ablation Studies on CelebAMask-HQ Dataset

To further prove the effectiveness of our proposed RIN block and RM (region matching) loss, we perform an ablation study in this section. We replaced our RIN block with SEAN, removed the RM loss, i.e. set $\lambda_{RM} = 0$ in equation (8), and tested the performance of SART for gender style translation using CelebAMask-HQ dataset. Fig. 8 shows the translation results of different regions for a young man when SART with different settings are applied. Compared with SART using SEAN blocks, the left/right eye and nose (the 2nd, 3rd and 4th columns) translated by the original SART present more lady-like styles, i.e. eye-shadows appear around the eyes and the nose is whitened. When RML is removed, there is no significant difference among the faces presented in the third row when left/right eye and nose are translated, respectively. The long hair in the sixth column actually does not fit the face boundary well.

Table 2 lists the accuracy, FID and LPIPS of different settings. Compared with SEAN, our RIN block significantly reduces FID from 35.01 to 31.06. The accuracy of our SART is also higher than that with SEAN and trained without RML.

Conclusion

This paper proposed a novel self-adaptive region translation framework, named SART, for region level face translation. A region-adaptive instance normalization block and region matching loss are proposed to fuse the per-region style of style and content images, and reduce the influence between different regions, respectively. The proposed SART is evaluated on three datasets and the experiments results demonstrates its effectiveness.

Acknowledgments

The work was supported by the Natural Science Foundation of China under grants no. 61672357, 91959108 and U1713214, and the Science and Technology Project of Guangdong Province under grant no. 2018A050501014.

References

Chen, W.; Shen, L.; and Lai, Z. 2019. Introspective Gan for Meshface Recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, 3472–3476. doi: 10.1109/ICIP.2019.8803594.

Chen, W.; Xie, X.; Jia, X.; and Shen, L. 2019. Texture Deformation Based Generative Adversarial Networks for Multidomain Face Editing. In Nayak, A. C.; and Sharma, A., eds., *PRICAI 2019: Trends in Artificial Intelligence*, 257–269. Cham: Springer International Publishing. ISBN 978-3-030-29908-8.

Chen, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Chu, C.; Shen, L.; and Zheng, Y. 2020. TR-GAN: Topology Ranking GAN with Triplet Loss for Retinal Artery/Vein Classification. In Martel, A. L.; Abolmaesumi, P.; Stoyanov, D.; Mateus, D.; Zuluaga, M. A.; Zhou, S. K.; Racoceanu, D.; and Joskowicz, L., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 616–625. Cham: Springer International Publishing. ISBN 978-3-030-59722-1.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Deng, Q.; Cao, J.; Liu, Y.; Chai, Z.; Li, Q.; and Sun, Z. 2020. Reference Guided Face Component Editing. *arXiv preprint arXiv:2006.02051* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28(11): 5464–5478.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 6626–6637.

Hong, S.; Yang, D.; Choi, J.; and Lee, H. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7986–7994.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jiang, W.; Liu, S.; Gao, C.; Cao, J.; He, R.; Feng, J.; and Yan, S. 2020. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5194–5202.

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

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4401–4410.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D. H.; Hawk, S. T.; and Van Knippenberg, A. 2010. Presentation and validation of the Radboud Faces Database. *Cognition and emotion* 24(8): 1377–1388.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 700–708.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, 10551–10560.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which training methods for GANs do actually converge? *arXiv preprint arXiv:1801.04406* .

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2337–2346.

Ricanek, K.; and Tesafaye, T. 2006. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 341–345. IEEE.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

W. Liu, W. Chen, Y. Z. L. S. 2020. SATGAN: Augmenting Age Biased Dataset for Cross-Age Face Recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, 465–476.

Zhu, P.; Abdal, R.; Qin, Y.; and Wonka, P. 2020. SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5104–5113.