A Novel Cross Channel Self-Attention based Approach for Facial Attribute Editing

Meng Xu[1,2], Rize Jin[1,2*], Liangfu Lu[3] and Tae-Sun Chung[4]
[1] School of Computer Science & Technology, Tiangong University
Tianjin, 300387 China
[e-mail: 1931085547@tiangong.edu.cn; jinrize@tiangong.edu.cn]
[2] Tianjin International Joint Research and Development Center of Autonomous Intelligence Technology and Systems, Tianjin, China
[3] School of Medical College, Tianjin University
Tianjin, 300072 China
[e-mail:liangfulv@tju.edu.cn]
[4] Department of Artificial Intelligence Ajou University
Suwon, 16499 South Korea
[e-mail:tschung@ajou.ac.kr]
[*]Corresponding author: Rize Jin

*Received March 8, 2021; revised April 21, 2021; accepted May 11, 2021;
published June 30, 2021*

Abstract

Although significant progress has been made in synthesizing visually realistic face images by Generative Adversarial Networks (GANs), there still lacks effective approaches to provide fine-grained control over the generation process for semantic facial attribute editing. In this work, we propose a novel cross channel self-attention based generative adversarial network (CCA-GAN), which weights the importance of multiple channels of features and archives pixel-level feature alignment and conversion, to reduce the impact on irrelevant attributes while editing the target attributes. Evaluation results show that CCA-GAN outperforms state-of-the-art models on the CelebA dataset, reducing Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) by 15~28% and 25~100%, respectively. Furthermore, visualization of generated samples confirms the effect of disentanglement of the proposed model.

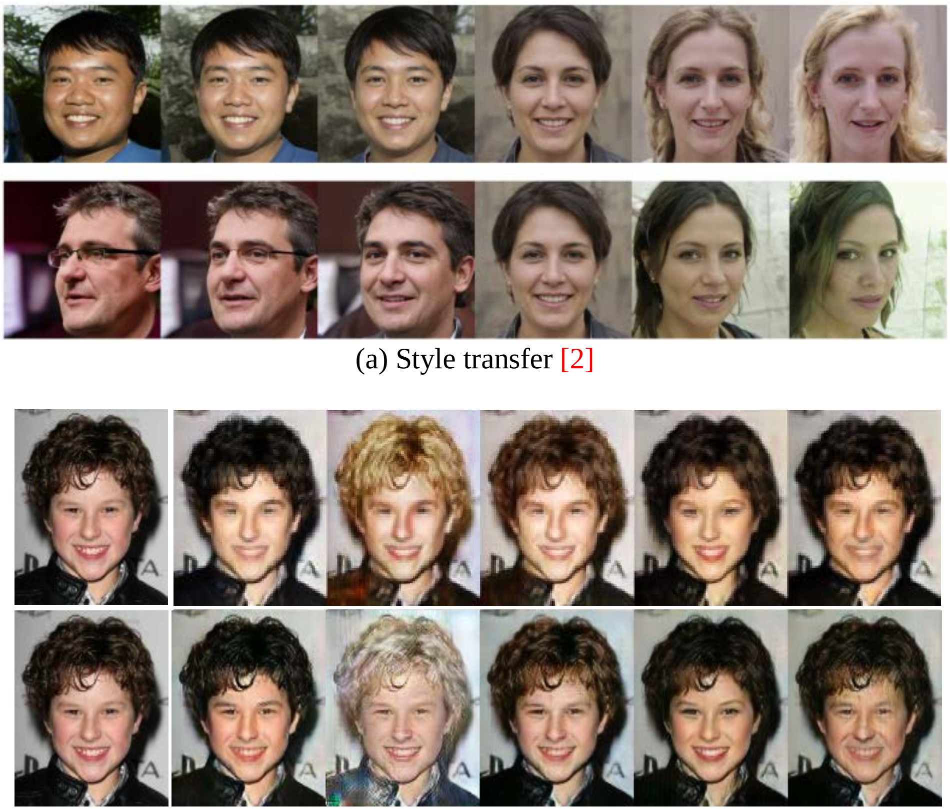
Keywords: Generative Adversarial Network, Cross Channel Self-Attention, Image Translation, Style Transfer, Facial Attribute Editing

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61806142, and in part by the Tianjin Science and Technology Program under Grants 18JCYBJC44000 and 19PTZWHZ00020.

http://doi.org/10.3837/tiis.2021.06.010 **ISSN : 1976-7277**

1. Introduction

Significant progress has been made over the past few years in Generative Adversarial Networks (GANs), the quality of images produced by GANs has improved rapidly. State-of-the-art GANs [1, 2, 3] can produce high-fidelity face images. However, it remains challenging to reduce the impact on irrelevant attributes while editing the target attributes. For example, how to reduce the impact on the background or other details of the human face when changing the hair color attribute? Some approaches [2, 3] embed the input image as latent code in an intermediate vector space, and their generators allow more linear and less entangled representation of different factors of variation. However, due to the entanglement of different semantics in the latent space, editing along one attribute can still lead to unexpected changes in other semantics. Specifically, face style transfer heavily affects image background and it fails to complete the separation of front and back scenes, as shown in **Fig. 1**.



(a) Style transfer [2]

(b) Facial attribute editing [1, 4]

Fig. 1. The entanglement phenomenon found in fine-grained image synthesis.

With the above limitation of existing methods in mind, we argue that there is no need to restrict the editing along the predefined linear path, as we just focus on the target attribute itself and reduces the impact on irrelevant attributes. The self-attention model has the advantage of modeling long range, multi-level dependencies of image regions. Self-Attention Generative Adversarial Network (SAGAN) [4] makes the fine details of each image position carefully coordinated with the fine details of the distant parts of the image. Dual Attention

Network for Scene Segmentation (DANet) [5] captures feature dependencies by channel-wise feature alignment in both channel and spatial dimensions, simultaneously, effectively leveraging the relationships between object classes or stuff in the global view. However, the original self-attention score is calculated with vectors as basic units, and the direct adoption of the self-attention module to image processing causes problems with coarse-grained alignment, while the image is calculated with pixels/feature points as basic semantic units, not vectors. Furthermore, features in the image can be decomposed/disentangled and be represented across multiple layers, however the previous methods are performing self-attention only on the same feature map, failing to realize that each map is just a part of a whole.

To better entangle related attributes located in different feature maps, and separate them from each other, especially, from background information, we propose a new model design, namely Cross Channel Self-Attention Generative Adversarial Network (CCA-GAN), for facial attribute editing. Specifically, to achieve pixel/feature level transformation and fine-grained image generation, we introduce a cross channel self-attention module prior to convolution, to accelerate the training, the cross channel self-attention module in the discriminator was used only before the first convolution layer. Furthermore, fusing multi-channel features will facilitate connectivity of feature maps.

The main contributions of our work are threefold as follows.

- We propose the Cross Channel Self-Attention to enhance the quality of facial attribute editing.
- A fusion method with different channel pixels, feature points is proposed to enhance the connectivity between feature maps and better satisfy the geometry and semantic structure constraints.
- We have conducted extensive experiments on the widely-used benchmark dataset for facial attribute editing, namely CelebA dataset, to evaluate the effectiveness of Cross Channel Self-Attention, and of the models based on the novel self-attention.

The remainder of the paper is organized as follows. Section 2 reviews related work. The proposed Cross Channel Self-Attention for Facial Attribute Editing is presented in Section 3. In Section 4 we experiment with the CelebA dataset. We conclude this paper in Section 5.

2. Related Works

Due to their great potential in generating high-quality, photo-realistic images, Generative Adversarial Networks (GANs) [6, 7] have been widely applied to image editing [8, 9], super-resolution [10, 11], image inpainting [12, 13], video synthesis [14, 15], etc. Despite this tremendous success, GANs still lack control over the generation process. Recently, many attempts have been made to improve the generation quality and fine-grained editing [16, 17, 18].

Semantic or structure coherent editing expects the model to modify the specified attributes yet maintain other information of the input image. In order to achieve this, existing models rely on an auxiliary classifier [1, 19], but the edited faces cannot better entangle the related attributes located on different feature maps and separate the faces from the background. Another successful line of research on manipulating the latent space of a pretrained GAN generator. The generator of StyleGAN [2, 3] allows for a linear, rather entangled, representation of different variation factors. However, adopting linear editing latent space can also lead to unexpected changes in other semantics. Unlike previous learning-based methods, our method is to introduce a cross-channel self-attention module before convolution, which can not only achieve pixel/feature level transformation and fine-grained image generation, but also integrate multi-channel characteristics.

3. Cross Channel Self-Attention for Facial Attribute Editing

In this section, we present the general architecture of the proposed model, and then discuss the Cross Channel Self-Attention mechanism in details.

3.1 Overview

As depicted in **Fig. 2**, the proposed Cross Channel Self-Attention Generative Adversarial Network (CCA-GAN) is composed of two neural networks, a generator and a discriminator, which are adversaries of each other. The generator (G) is trained to study the mapping between the source domain and multiple target domains, i.e., it converts the input image s to the output image $G(s,t)$ given the target domain label t, and G is also reconstructing the generated image $G(s,t)$ provided with the original domain label o back to the original image $G(G(s,t),o)$. Besides the adversarial functionality that judges the true and false of the input image, the discriminator (D) also utilizes an auxiliary classifier [20] to assess the generation quality in terms of specified target labels. Specifically, in the provided example, D outputs two predictions, one with 1 dimension and the other 5 dimensions, indicating the probability of the input image being real and the probability distribution over target domains, respectively.

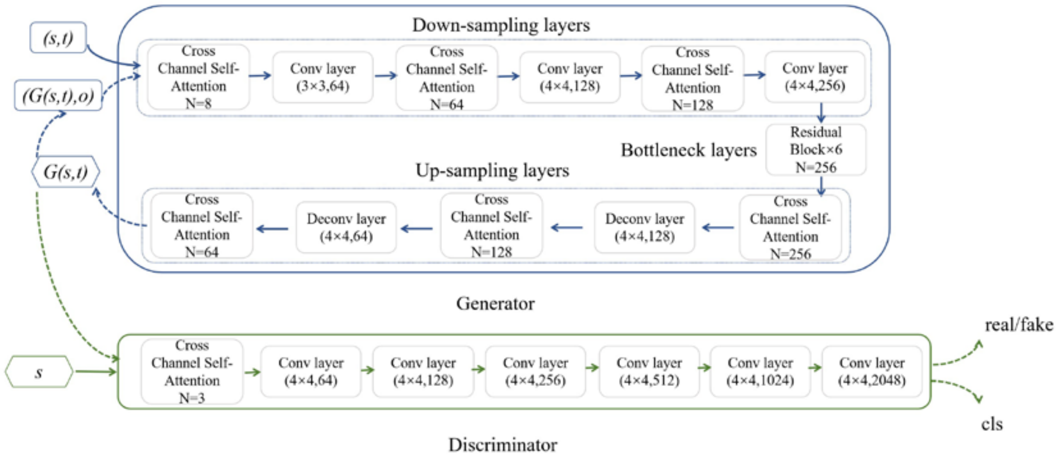


Fig. 2. The overall architecture of CCA-GAN. Cross Channel Self-Attention Module is added before each layer of convolution operation in G, and before the first layer of convolution operation in D.

The generator mainly consists of three main components, i.e., down sampling layers, bottleneck layers, and up sampling layers. 6 residual blocks were adopted for bottleneck layer [21] and cross channel self-attention is applied before each convolution layer to realize inter-channel pixel/feature level attention. To reduce the computational complexity, we suggest to insert the proposed self-attention layer only before the first convolution in the discriminator model.

3.2 Cross Channel Self-Attention

Due to the limitation in the size of receptive field, the convolution operation in traditional

GANs cannot reasonably express the image details. Attention operation is necessary to express global dependencies by establishing association with features. However, the phenomenon of entanglement, especially in facial attribute editing, can be observed, i.e., irrelevant attributes are affected while transforming the specified attribute. Furthermore, note that the calculation of the original self-attention is performed on vectors as a vector is an indivisible unit in word embedding [22, 23]. However, the image has no such constraint, and every pixel feature point has its meaning in the physical world. Based on such observation, we explore a pixel-/feature-wise alignment approach cross channels. Our method could adaptively adjust the feature information of different channels and perform disentangled and controllable edits along various attributes.

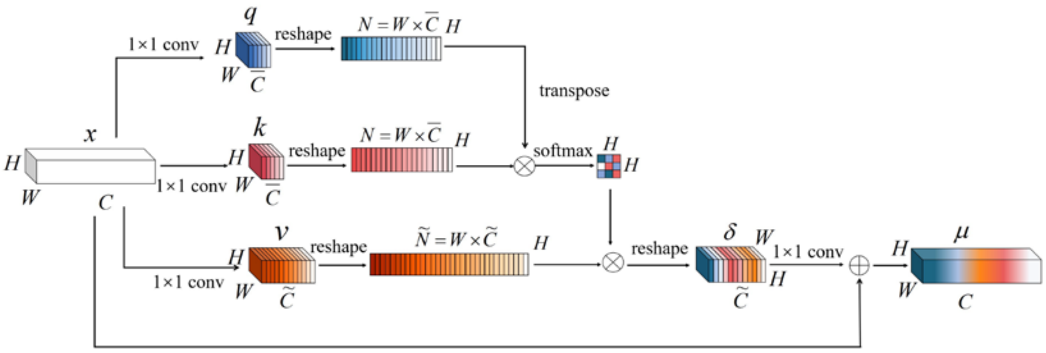


Fig. 3. An illustration of the Cross Channel Self-Attention Operation. x: input the image, H: height, W: width, C: channel ($\overline{C} = C/8$, $\tilde{C} = C/2$).

As shown in **Fig. 3**, the module takes as input the image $x \in R^{H\times W\times C}$, it first goes through convolution layers to obtain two transformations q and k, $\{q, k\} \in R^{H\times W\times \overline{C}}(\overline{C} = C/8)$, where $C \in \{64{,}128{,}256\}$. Then we change the way they fuse into $R^{H\times N}$, where $N = W \times \overline{C}$ is the number of pixels across the channel. We conduct a matrix multiplication between q and the transpose of k, and apply the softmax to obtain the cross channel self-attention map $\alpha \in R^{H\times H}$:

$$\alpha_{j,i} = \frac{exp(k_i \cdot q_j)}{\sum_{i=1}^{H} exp(k_i \cdot q_j)} \tag{1}$$

where $\alpha_{j,i}$ measures the impact of the i^{th} position on j^{th} position.

Meanwhile, we feed x into a convolution layer to a generate a new transformation $v \in R^{H\times W\times \tilde{C}}(\tilde{C} = C/2)$, where $C \in \{64{,}128{,}256\}$, and reshape it to $R^{H\times \tilde{N}}$. Then we perform a matrix multiplication between α and v, and reshape the result to $R^{H\times W\times \tilde{C}}$. δ is the attention layer:

$$\delta_j = \sum_{i=1}^{H} \alpha_{j,i} v_i \tag{2}$$

$$\delta = (\delta_1, \delta_2, \ldots, \delta_j, \ldots, \delta_H) \in R^{H\times \tilde{N}} \tag{3}$$

Finally, we perform an element-wise addition between x and δ to obtain the final output $\mu \in R^{H\times W\times C}$ as follows:

$$\mu_i = \gamma\delta_i + x_i \tag{4}$$

where γ is a learnable scaling parameter, which is initialized with 0.

It is worth noting that we use different channel numbers for the first self-attention layers in the generator and discriminator. In order to encode the source domain image as well as the target domain information, we increase the value of C to (3+t) for the generator, where t is the number of target domain attributes. In contrast, $C = 3$ in the discriminator.

3.3 Loss Functions

The first cost measure we used is an adversarial loss [1].

$$L_{adv} = E_s[logD_{src}(s)] + E_{s,t}\left[log\left(1 - D_{src}\left(G(s,t)\right)\right)\right] \tag{5}$$

where the term $D_{src}(s)$ as a probability distribution over sources given by discriminator D. The generator G is trained to minimize the objective and the discriminator D tries to maximize the it.

In addition, D is optimized by the domain classification loss of real images, and G is optimized by the domain classification loss of fake images, respectively, as follows.

$$L_{cls}^{r} = E_{s,o}[-\,log\,D_{cls}\,(o|s)] \tag{6}$$

$$L_{cls}^{f} = E_{s,t}[-\,log\,D_{cls}\,(t|G(s,t))] \tag{7}$$

where the term $D_{cls}(o|s)$ as a probability distribution over domain labels computed by discriminator D.

In order to constrain the background image pixel similarity, we introduce the L1 distance function to calculate the reconstruction loss between the reconstructed image and the original image, defined as follows.

$$L_{\mathrm{rec}} = E_{s,t,o}[||s - G(G(s,t),o)||_1] \tag{8}$$

The final optimization goal of the model is shown as follows.

$$L_D = -L_{adv} + \lambda_c L_{cls}^{r} \tag{9}$$

$$L_G = L_{adv} + \lambda_c L_{cls}^{f} + \lambda_r L_{rec} \tag{10}$$

where λ_c and λ_r, which are hyper-parameters and tuned on the development set, control the relative importance of domain classification losses (L_{cls}^{r} and L_{cls}^{f}) and reconstruction loss (L_{rec}), respectively. We are planning to make them trainable in the future work, so as to facilitate the practical application of the proposed method.

4. Experiments and Analysis

For evaluating the effectiveness of our model, we compared it with StarGAN [1] and StarGAN+SAGAN [4] on the CelebA dataset [24], and the baseline models all enable multiple domain face property editing tasks using a single network. Evaluation results demonstrate that CCA-GAN achieves new state-of-the-art performance. Next, we first introduce the dataset and implementation details, then we report the experimental results.

4.1 Data and Training Details

The CelebFaces Attributes (CelebA) dataset [24] is a facial attribute dataset with 202,599 color images from 10,177 celebrities, each with 40 attribute annotations, covering a large number of face angles and a cluttered background. We crop and resize the initial 178×218 size images to 128×128. Then, about 200k images are randomly selected as the training and development set, and the remaining is used as the test set. Among them, seven attribute fields are constructed using the following attributes: hair color (black, gold, brown), gender (male/female) and age group (young/old). In addition, the model uses the mask vector (one-hot encoding), which allows our model to pay attention to the specified tag provided by CelebA, ignoring the unspecified tag. The corresponding position of the specified tag is set to 1, and the rest positions are set to 0.

We implement CCA-GAN using TensorFlow. We use $\lambda_c = 1$ and $\lambda_r = 1$ in all of our experiments. The batch size is 32, and Adam optimizer [25] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ are used throughout the experiments. We horizontal flipped the images for data augmentation with a probability of 0.5. After training the discriminator five time, the generator is trained once. We trained the models for 20 epochs, where the learning rate of the first 10 epochs is fixed as 0.0001, and then it decreases linearly to 0 for the following epochs.

4.2 Evaluation Metrics

To assess the quality of generated samples, we employ two standard metrics: Fréchet Inception Distance (FID) [26] and Kernel Inception Distance (KID) [27]. Unlike the simple Inception Score (IS), which evaluates only the distribution of generated data, FID and KID compare the distributions of generated images and real images. A lower the FID or KID metric indicates higher visual similarity between the two. We use the same amount of real and fake samples for evaluation, so that we can compare the scores fairly. This is necessary because metrics such as FID can produce highly biased estimates [27], using a larger sample size can result in significantly lower scores.

4.3 Results

As seen in **Table 1**, CCA-GAN has achieved a significant improvement on the FID and KID metrics on the challenging face dataset CelebA, compared to StarGAN and StarGAN+SAGAN. Specifically, our model has lowered the FID score by 15% to 28% and by 17% to 27%, compared to respective models. In this evaluation, StarGAN+SAGAN could not outperform the baseline model StarGAN on attributes "Black Hair" and "Golden Hair". In contrast, our model has lowered the store of the respective attributes by 15% and 20% over StarGAN, and 17% and 24% over StarGAN + SAGAN. This suggests that the proposed attention approach is much more effective in generating diverse samples than the conventional self-attention mechanism. Even greater improvements can be observed in the KID evaluation. Our model has lowered the score by 23% to 100% in the most target attributes.

Table 1. FID and KID evaluation results

Metric	**Models**	**Black hair**	**Golden hair**	**Brown hair**	**Gender**	**Aged**
FID	StarGAN	0.80	1.01	0.81	0.93	0.89
	StarGAN+ SAGAN	0.82	1.06	0.78	0.92	0.87
	CCA-GAN	**0.68**	**0.81**	**0.61**	**0.67**	**0.65**
	CCA-GAN (λ_r=0)	0.78	0.93	0.72	0.86	0.78
KID	StarGAN	2.00±1.05	1.97±0.47	1.42±0.58	1.88±0.67	2.10± 0.97
	StarGAN+ SAGAN	1.15±0.51	2.66±0.94	1.23±0.62	1.12±0.72	1.45± 0.39
	CCA-GAN	1.37±0.74	1.52±0.41	1.75±0.87	**0.00±0.71**	1.21± 1.67
	CCA-GAN (λ_r=0)	**0.48±0.47**	**1.47±0.50**	**0.52±0.45**	0.84±0.50	**0.24±0.24**

It has been shown that the reconstruction loss plays an important role in the image retention of the original geometric shape and semantic structure constraints [1, 28]. However, in the FID evaluation, a variant model (CCA-GAN with $\lambda_r = 0$) improves the score by 3% to 12% over StarGAN, and by 5% to 12% over StarGAN + SAGAN, in all target attributes. This suggests that the proposed cross channel self-attention mechanism plays a much more important role in preserving the original geometry and semantic consistency of the image. Similarly, in the KID evaluation, the variant model improves by 25% to 89% over StarGAN, and 45% to 89% over StarGAN+SAGAN. At the same time, it slightly outperforms the vanilla CCA-GAN by 3% to 80% except for the "Gender" attribute. We conjecture that KID is an unbiased estimator, compared to FID. It is further demonstrated that the proposed self-attention approach plays a key role in preserving the original features of the face, and improves the quality and diversity of the generated images.

4.4 Case Study and Visualization

Fig. 4, **Fig. 5**, and **Fig. 6** show generated samples of facial attribute editing. We can see that StarGAN and StarGAN+SAGAN generate images with poor quality, and they cannot effectively disentangle the specified ones from the other attributes and the background information. For example, in the third column of **Fig. 4**, the image background turned golden as the hair color changes. In the fifth column of **Fig. 4**, the existing models tend to change the texture of the collar when transforming the gender, and **Fig. 5** has the same visual problem. In **Fig. 6**, StarGAN tends to yellow the skin when generating black hair, and there is a noticeable change in the details of the hair when generating golden hair. StarGAN+SAGAN also makes the pixels worse. By contrast, our method yields a more controllable editing and is superior in terms of identity preservation. **Fig. 7** shows a better visualization of face attribute editing from **Fig. 4**, by comparing the color distribution heat map experiment. We can clearly observe that the color distribution of our model is the most similar to the original image after changing the specified attributes.

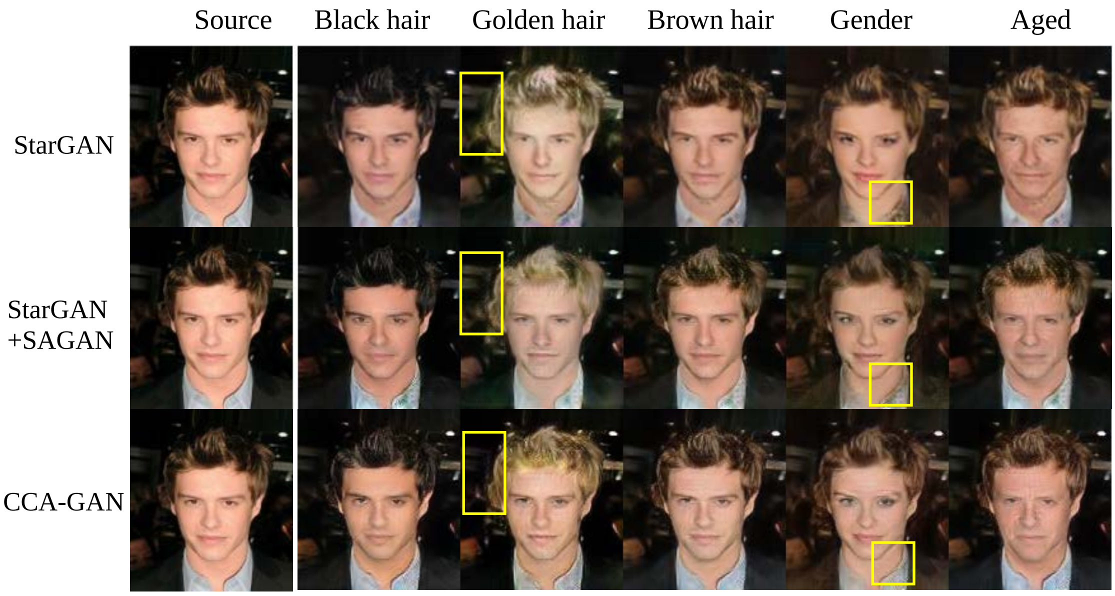


Fig. 4. Samples of male facial attribute editing. The first column displays the source domain image, and the next five columns display the attribute editing results.

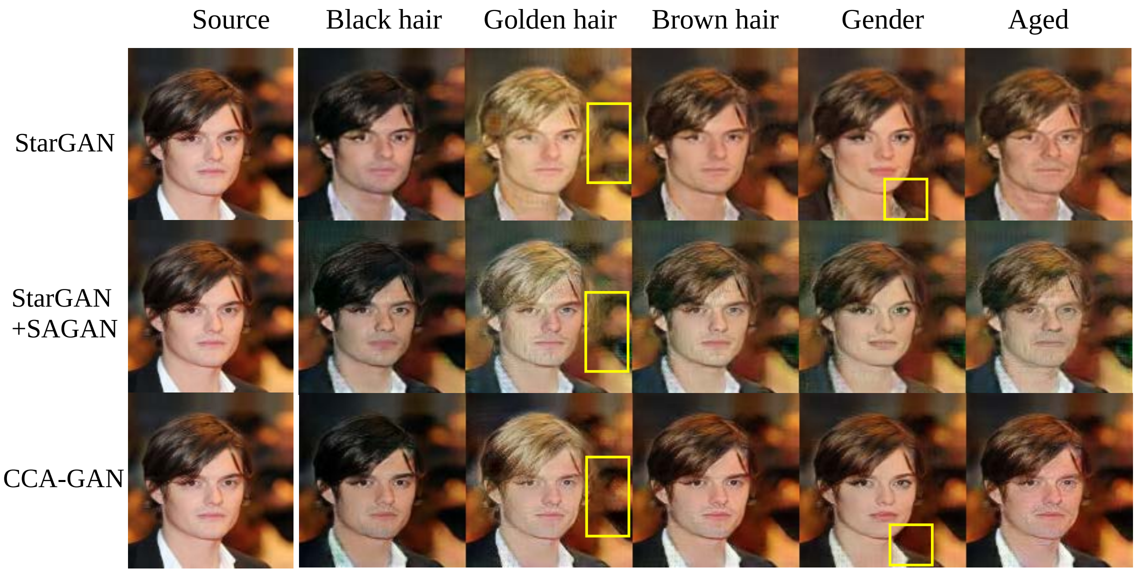


Fig. 5. Samples of male facial attribute editing.

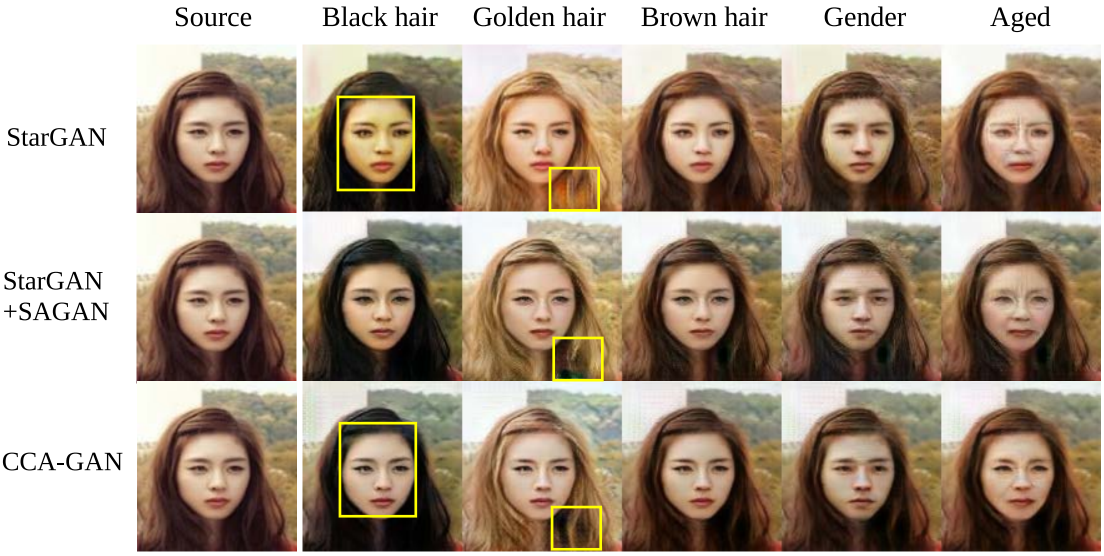


Fig. 6. Samples of female facial attribute editing.

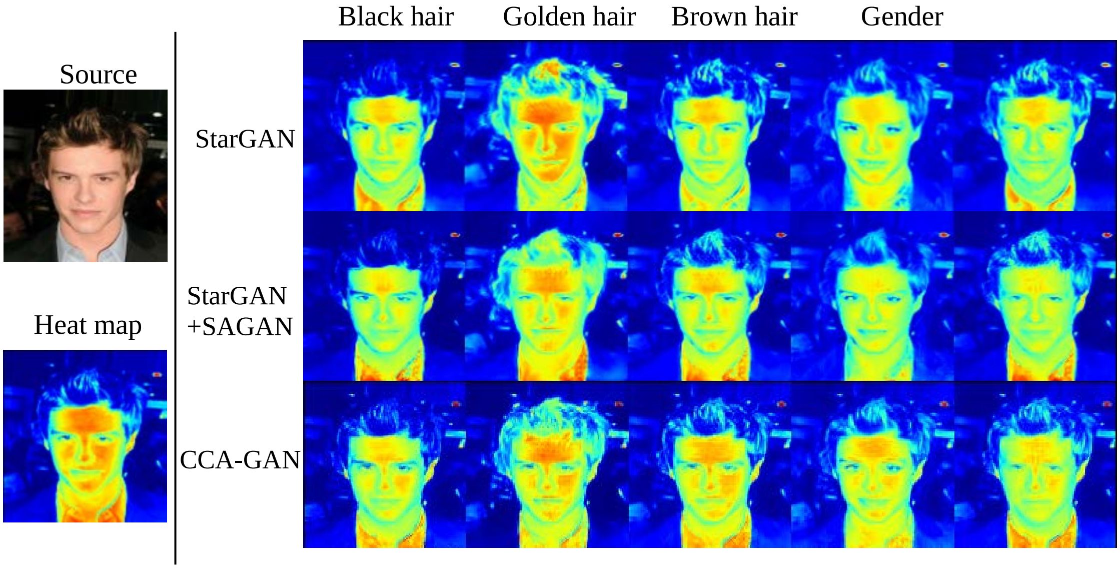


Fig. 7. Heat map visualization of **Fig. 4**.

5. Conclusion

In this work, we introduce CCA-GAN to explore fine-grained control over the generation process for high-level face editing. Our approach features a novel learning framework that calculates the importance of multiple feature channels and realizes pixel-level feature alignment and conversion to guide the image generation process of StarGAN. As a result, we are able to disentangle a variety of semantics and achieve precise edits along different facial attributes. Extensive evaluations demonstrate the superior performance of the proposed approach both in image quality and semantic coherence compared to state-of-the-art works. In order to demonstrate the generalization ability of the proposed model, further experiments on more types of real and artistic face datasets are needed, which we leave as future work.

References

[1] Choi. Y, Choi. M, Kim. M, Ha. J. W, Kim. S, and Choo. J, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. of IEEE Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8789-8797, June, 2018. Article (CrossRef Link)

[2] Karras. T, Laine. S, and Aila. T, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2020. Article (CrossRef Link)

[3] Karras. T, Laine. S, Aittala. M, Hellsten. J, Lehtinen. J, and Aila. T, "Analyzing and improving the image quality of stylegan," in *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*, United States, pp. 8110-8119, June, 2020. Article (CrossRef Link)

[4] Zhang. H, Goodfellow. I. J, Metaxas. D. N, and Odena. A, "Self-attention generative adversarial networks," in *Proc. of ICML*, pp. 7354–7363, July, 2018.

[5] Fu. J, Liu. J, Tian. H, Li. Y, Bao. Y, Fang. Z, and Lu. H, "Dual attention network for scene segmentation," in *Proc. of IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 3146–3154, June, 2019. Article (CrossRef Link)

[6] Goodfellow. I, Pouget-Abadie. J, Mirza. M, Xu. B, Warde-Farley. D, Ozair. S, Couville. A, and Bengio. Y, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, 2020. Article (CrossRef Link)

[7] Ma. X, Jin. R, Sohn. K. A, Paik. J. Y, and Chung. T. S, "An Adaptive Control Algorithm for Stable Training of Generative Adversarial Networks," *IEEE Access*, vol. 7, pp. 184103-184114, 2019. Article (CrossRef Link)

[8] Lample. G, Zeghidour. N, Usunier. N, Bordes. A, Denoyer. L, and Ranzato. M. A, "Fader networks: Manipulating images by sliding attributes," in *Proc. of Neural Information Processing Systems (NIPS 2017)*, vol. 30, pp. 5969-5978, 2017. Article (CrossRef Link)

[9] Bau. D, Strobelt. H, Peebles. W, Wulff. J, Zhou. B, Zhu. J.-Y. and Torralba. A, "Semantic photo manipulation with a generative image prior," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 59, July, 2019. Article (CrossRef Link)

[10] Ledig. C, Theis. L, Huszar. F, Caballero. J, Cunningham. A, Acosta. A, Aitken. A, Tejani. A, Totz. J, and Wang. Z, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 105–114, July, 2017. Article (CrossRef Link)

[11] Wang. X, Yu. K, Wu. S, Gu. J, Liu. Y, Dong. C, Qiao. Y, and Loy. C. C, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. of ECCV Workshops*, Germany, pp. 63–79, September, 2018. Article (CrossRef Link)

[12] Yeh. R. A, Chen. C, Lim. T. Y, Schwing. A. G, Hasegawa-Johnson. M, and Do. M. N, "Semantic image inpainting with deep generative models," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6882–6890, July, 2017. Article (CrossRef Link)

[13] Yu. J, Lin. Z, Yang. J, Shen. X, Lu. X, and Huang. T, "Free-form image inpainting with gated convolution," in *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul Korea, pp. 4470-4479, Nov, 2019. Article (CrossRef Link)

[14] Wang. T.-C, Liu. M.-Y, Zhu. J.-Y, Liu. G, Tao. A, Kautz. J, and Catanzaro. B, "Video-to-video synthesis," in *Proc. of the 32nd International Conference on Neural Information Processing Systems*, pp. 1152–1164, 2018. Article (CrossRef Link)

[15] Wang. T.-C, Liu. M.-Y, Tao. A, Liu. G, Catanzaro. B, and Kautz. J, "Few-shot video-to-video synthesis," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5013–5024, December, 2019.

[16] Arjovsky. M, Chintala. S, Bottou. L, “Wasserstein generative adversarial networks,” in *Proc. of International Conference on Machine Learning*, Sydney, Australia, pp. 214-223, August, 2017. Article (CrossRef Link)

[17] Gulrajani. I, Ahmed. F, Arjovsky. M, Dumoulin. V, and Courville. A, “Improved training of wasserstein GANs,” in *Proc. of International Conference on Neural Information Processing Systems*, vol. 30, pp. 5769–5779, 2017. Article (CrossRef Link)

[18] Miyato. T, Kataoka. T, Koyama. M, and Yoshida. Y, “Spectral normalization for generative adversarial networks,” in *Proc. of International Conference on Learning Representations*, Vancouver, Canada, 2018.

[19] Choi. Y, Uh. Y, Yoo. J, and Ha. J.-W, “StarGAN v2: Diverse image synthesis for multiple domains,” in *Proc. of IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 8188–8197, June, 2020. Article (CrossRef Link)

[20] Odena. A, Olah. C, Shlens. J, “Conditional image synthesis with auxiliary classifier gans,” in *Proc. of International conference on machine learning*, Sydney, Australia, pp. 2642-2651, Augus, 2017. Article (CrossRef Link)

[21] He. K, Zhang. X, Ren. S, and Sun. J, “Deep residual learning for image recognition,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, June, 2016. Article (CrossRef Link)

[22] Bahdanau. D, Cho. K, and Bengio. Y, “Neural machine translation by jointly learning to align and translate,” in *Proc. of International Conference on Learning Representation*s, San Diego, CA, USA, 2014. Article (CrossRef Link)

[23] Vaswani. A, Shazeer. N, Parmar. N, Uszkoreit. J, Jones. L, Gomez. A. N, Kaiser. L. G, and Polosukhin. I, “Attention is all you need,” in *Proc. of International Conference on Neural Information Processing Systems*, vol. 30, pp. 6000-6010, December, 2017. Article (CrossRef Link)

[24] Liu. Z, Luo. P, Wang. X, and Tang. X, “Deep learning face attributes in the wild,” in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 3730–3738, December, 2015. Article (CrossRef Link)

[25] Kingma. D. P, and Ba. J. L, “Adam: A method for stochastic optimization,” in *Proc. of International Conference on Learning Representations*, San Diego, CA, USA, 2015. Article (CrossRef Link)

[26] Heusel. M, Ramsauer. H, Unterthiner. T, Nessler. B, and Hochreiter. S, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017. Article (CrossRef Link)

[27] Binkowski. M, Sutherland. D. J, Arbel. M, and Gretton. A, “Demystifying mmd gans,” in *Proc. of International Conference on Learning Representations*, Vancouver, Canada, 2018.

[28] Zhu. J.-Y, Park. T, Isola. P, and Efros. A. A, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, Honolulu, HI, USA, pp. 2242–2251, 2017. Article (CrossRef Link)

Meng Xu received the Bachelor's degree in software engineering from Tiangong University, Tianjin, China, in 2019. She is currently a graduate student at Tiangong University. Her current research interests include computer vision, image generation, deep learning.

Rize Jin received the M.S. and Ph.D. degrees in computer engineering from Ajou University (AU), South Korea, in February 2011 and February 2015, respectively. He was an Assistant Professor with the Software Department, AU. Before joining AU, he was a Postdoctoral Researcher with the Department of Computer Engineering, Korea Advanced Institute of Science and Technology (KAIST), South Korea. He is currently a professor with the School of Computer Science and Technology, Tiangong University, China. His research interests include natural language processing, and deep learning.

Liangfu Lu received the B.S. and M.S. degrees in computational mathematics from Ludong University and Nanjing University of Aeronautics and Astronautics, China in 2001and 2004, respectively, and his Ph.D. in computer science from Tianjin University, China in 2008. He worked as a visiting scholar in the University of Technology, Sydney, Australia, from Sept. 2011 to Sept.2012. He is currently an associate professor in Medical College of Tianjin University. His research interests include deep learning, image processing and compressive sensing etc. He has published over 40 papers in top journals and conference proceedings and principally investigated some research projects NaturalScience Foundation of China.

Tae-Sun Chung received the B.S. degree from KAIST, Daejeon, South Korea, in 1995, and the M.S. and Ph.D. degrees from Seoul National University, Seoul, South Korea, in 1997 and 2002, respectively, all in computer science. He is currently a Professor with the Department of Artificial Intelligence, Ajou University, Suwon, South Korea. His current research interests include flash memory storage, database systems, and machine learning.