Spatial and Channel-wise Attention in Generative Adversarial Network for Text-to-Face Synthesis and Manipulation

ZHOU YUTONG[1,a)]   SHIMADA NOBUTAKA[1]

Abstract: Several studies have been conducted on text-to-image synthesis techniques that transfer text descriptions into realistic images over recent years. However, due to the lack of dataset, there is almost no previous research focusing on text-to-face(T2F) synthesis, which possesses significant potential in public safety and security, art creation, image editing, image retrieval, *etc.* In this paper, we propose a T2F generative adversarial network with spatial-wise and channel-wise attention mechanisms, which can not only synthesize high-resolution facial images but also manipulate various facial local attributes (*e.g.*, eyes, mouth, skin, hair, age, *etc.*) with the correlated keywords while preserving the identity's facial characteristics. In addition, we collect and annotate a novel Flickr-Faces-HQ with Text descriptions dataset (FFHQ-Text), which consists of high-resolution face images (1024×1024) with text descriptions and bounding boxes. The experimental results on our FFHQ-Text dataset show that the face images generated by the proposed algorithm perform superior both in quality and quantity than the existing text-to-image synthesis approach – AttnGAN.

Keywords: Computer Vision, Generative Adversarial Network, Text-to-image Synthesis

SpatialおよびChannel-wiseなAttention機構と敵対的生成による文章から顔画像生成・編集

周　禹トウ[1,a)]　島田　伸敬[1]

概要: 文章から写実的な画像の生成を行う技術は近年盛んに研究されている挑戦的な研究課題である．文章からの顔画像生成は防犯保安分野・アートの創作・画像編集・画像検索などにおいて大きな可能性を秘めているが，データセットが欠けているため，先行研究はあまりない．本論文では，Spatial-wiseおよびChannel-wiseなAttention機構を導入した敵対的生成ネットワークによる文章から顔画像生成手法を提案する．高解像度顔画像の生成だけでなく，様々な顔局所属性（例えば、目、口、肌、髪、年齢など）を顔全体の人相の同一性を維持しつつ，相関するキーワードを指定して局所的に編集することができる．さらに，新しいFlickr-Faces-HQ高解像度の顔画像（1024×1024）を収集し，テキスト注釈とバウンディングボックスを付けるデータセット（FFHQ-Text）を作る．FFHQ-Textデータセットを用いた実験結果より，提案手法によって生成された顔画像が，既存手法（AttnGAN）より量的質的に優れていることを示す．

キーワード: コンピュータービジョン、敵対的生成ネットワーク、Text-to-image生成モデル

[1] 立命館大学 情報科学研究科，草津市，滋賀県
College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan
[a)] zhou@i.ci.ritsumei.ac.jp

1. Introduction

In the last few decades, the fields of Computer Vision (CV) and Natural Language Processing (NLP) have been made several major technological breakthroughs in deep learning research. Recently, researchers appear interested in combining semantic information and visual information in these traditionally independent fields, such as caption generation [1], visual storytelling [2], visual question answering [3], *etc.* With the advent of Generative Adversarial Networks(GANs) [4] for realistic image synthesis, as a sub-domain of visual generation task, synthesizing images from text description well joint the Computer Vision (CV) and Natural Language Processing (NLP) fields, which is in a period of vigorous development in recent years [5–8]. Text-to-image synthesis contributes to analyzing the relationship between text and realistic images, which has extensive application prospects in art creation, image editing, image retrieval, *etc.*

In recent years, several papers have been explored on the subject of text-to-image synthesis. Most outstanding contributions rely on GAN-based models as their main driving force to generate images from text descriptions. Though they can efficiently produce images with specified shapes and colors, there are still large gaps in human specify requirements, *e.g.* (1) Generate high-resolution images without distorted local details. (2) Selectively change some keywords of a sentence to control different attributes synthesis results.

Nowadays, as one of the earliest practical applications of text-to-image technology in real life, visual reproduction of a suspect's appearance based on eyewitnesses' memory and verbal description is still playing an important role in the criminal investigation field. However, the witness's testimony may differ from the actual physical appearance, caused by fear or other reasons. Therefore, manipulating particular visual features of the synthetic images corresponding to the input description, while retaining the other attributes becomes increasingly essential in text-to-face generation task.

The commonly used datasets for text-to-image synthesis research are CUB [9], Oxford 102 Flowers [10] and COCO [11]. However, because there is no suitable dataset, few studies have focused on text-to-face synthesis techniques. Therefore, due to the ambiguity and complexity of text description, face-to-face synthesis is more challenging than original image-guided face synthesis and has more practical significance and application prospects.

This paper proposed a text-to-face synthesis method for filling in the gap in the text-to-image synthesis research. To summarize, the contributions of our work can be generalized as follows: (1) We propose a spatial-wise and channel-wise visual attention module on text to realistic high-resolution (512×512) face synthesis, which has a wide range of diversity and corresponds to relevant words. In addition, when the facial description has some particular changes, the proposed method can also selectively and locally change the attributes (*e.g.*, color, shape, *etc.*) of individual facial elements (*e.g.*, skin, hair, age, *etc.*) while maintaining the background and the appearance of other unaltered facial characteristics. (2) We propose a novel manual text-to-face dataset (FFHQ-Text) based on the FFHQ dataset [12] with text descriptions for 760 high-resolution female portraits. (3) The experimental results show that our proposed method achieves superior performance on the text-to-face task. Furthermore, we introduce a novel evaluation metric method on the "Age" element conversion task to measure the similarity of generated images after the input text transformation in the four groups.

2. FFHQ-Text Text-to-Face Dataset

In previous years, the primary approach of facial image generation is to draw some sketches by artists, which can help eyewitnesses recall more details about the suspect's appearance step by step [13]. This process is complicated because all suspect's characteristics should be recalled from the witness's memory. Moreover, to accurately match the sketches drawn with the descriptions of witnesses, artists need much training in drawing face sketch. The lack of a text-to-face synthesis dataset will significantly hinder text-to-face generation tasks that have already been proved extremely useful in many tasks, such as image retrieval, public safety, assist with crime investigation, *etc.* Therefore, based on the Flickr-Faces-HQ (FFHQ) dataset [12], we propose a text-to-face dataset that contains high-quality images of human faces in the wild and standard terminology descriptions named "**FFHQ-Text**".

To build a standard text-to-face dataset, we firstly download all facial images from Flickr-Faces-HQ Dataset (FFHQ) web page[*1]. We manually ensure that the downloaded high-quality facial images only contain one complete human face. In this study, we select a total of 760

[*1] https://github.com/NVlabs/ffhq-dataset

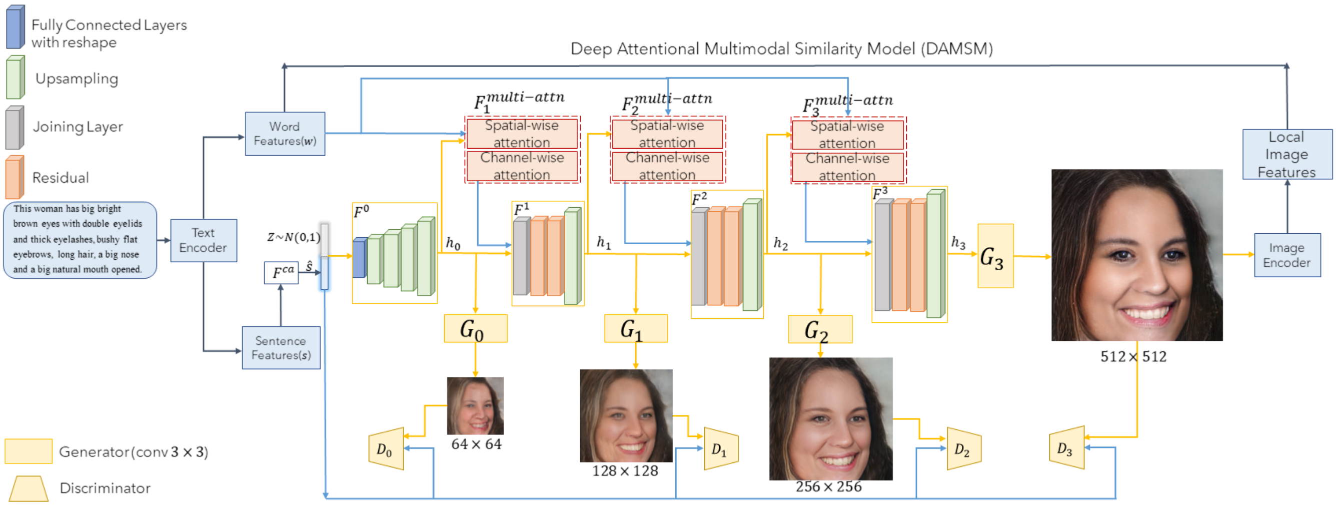


Fig. 1 The architecture of our proposed architecture. The multi-stage network generate 64×64, 128×128, 256×256 and 512×512 face images from coarse to fine.

portraits of female in the wild.

Facial feature selection and description are crucial parts because of two factors: the reliability of facial features and the lack of accuracy in witnesses' description [13]. In this study, we use the human facial text description terminology to setup our dataset FFHQ-Text.

For each face image of different people with different facial attributes, 9 sentences are annotated by 3 different persons to maintain the diversity of text descriptions. To create the standard facial-textual description dataset, we define each ordinary face image's appearance, containing 12 multi-valued facial elements (*e.g.*, eyes, hair, face, age, *etc.*) with 150 attributes (*e.g.*, size, color, shape, range, *etc.*). Especially for the definition of the "Age" element, we set 8 age categories : four for children ages 0-2, 4-6, 8-13, 15-20 (denote as "girl"), and four for adults 25-32, 38-43, 48-53, 60+ (denote as "woman").

3. Spatial and Channel-wise Attention Generative Adversarial Network

We aim to synthesis a high-quality and realistic face image $\hat{I}$ from the given text description T, which is highly correlated with the semantic meaning of the sentence. Furthermore, when the input text description T is modified to T', the synthesized image should well-matched with the modified text description while preserving not mentioned content in $\hat{I}$.

3.1 Architecture

In this paper, we adopt the recent multi-stage architectures of text-to-image synthesis researches [6, 7] as our baseline architecture shown in Fig. 1. Our architecture has four stages: 64×64, 128×128, 256×256 and 512×512, to synthesize high-resolution images from coarse to fine.

Our improved model consists of the spatial-wise and channel-wise mixed models stacked with m generators (G_0,G_1,...,G_{m-1}) and m discriminators (D_0,D_1,..., D_{m-1}) as a tree-like structure. The generated face images ($\hat{i}_0$,$\hat{i}_1$,...,$\hat{i}_{m-1}$) from coarse to fine.

For text encoder, we use the bi-directional Long Short-Term Memory(BiLSTM) [14] to take the input text description T, and convert it into word features $w \in \mathbb{R}^{D \times L}$ with the word number L and dimension D, and a global sentence feature $t \in \mathbb{R}^D$. Due to the fact that each word corresponds to the hidden states (h_0,h_1,...,h_{m-1}) in the BiLSTM, we concatenate the hidden states of each word in both directions of the Long Short-Term Memory (LSTM) to extract the semantic meaning of each word [7]. We also apply the conditioning augmentation F^{ca} [6] to transform the sentence feature t to a standard deviation vector and convert the sentence vector to a global conditional vector $\hat{t}$, which is to improve the diversity of synthesized images. Then, the global sentence feature $\hat{t}$ concatenated with a random noise vector z as an input of the first stage F^0 for reshaping and up-sampling. The first image feature from the first hidden layer h_0 output and the first image synthesis with low-resolution (64 × 64).

Note that AttnGAN [7] is not perfect in capturing coherent global structures, which cause the synthesized image are not likely photorealism. Therefore, different from AttnGAN, after the first stage F^0, we introduce a spatial-wise and channel-wise mixed attention module. The spatial-wise attention takes the word features w and the image feature from the previous hidden layer h as input, and output the word-context vector with spatial-wise attention h_s for image feature. Then, the channel-wise attention takes these corresponding word-context features h_s are further concatenated with the word features w and regarded as input for image synthesizing at the next stage (F^1,F^2,...,F^{m-1}).

For image encoder, the same as [7], we adopt a Convolutional Neural Network (CNN) that maps images to semantic vectors, which is built upon the Inception-v3 model [15] pretrained on ImageNet [16].

3.2 Word-Level Spatial and Channel-wise Attention Modules

We aim to focus on global information to important features and suppressing unnecessary ones [17]. To achieve this, the proposed word-level spatial-wise and channel-wise mixed attention modules utilize the related features from relevant words and different visual portions to enhance more details.

Spatial-wise Attention mainly focuses on 'where' is an informative spatial part [17], which extracts the image feature for recognizing individual words and convert them into a word of context vectors. The spatial attention mechanism attempts to pay more attention to the semantic-related regions [18].

The spatial-wise attention $F^{spatial-attn}(w, h)$ has two inputs as AttnGAN [7]: the word features $w \in \mathbb{R}^{D \times L}$ and the image features from the previous hidden layer h $\in \mathbb{R}^{C \times H \times W}$ (C denotes the channel, H and W denotes the height and width of the feature map respectively.) to generate fine-grained visual details that relevant to the input text descriptions. We first adopt a perceptron layer $P \in \mathbb{R}^{H \times W \times D}$ to convert the word features into the common semantic dimension as image features. We first convert the word features and image features into the common semantic dimension by adopting a perceptron layer P. The transformed word features denote as w' $\in \mathbb{R}^{H \times W \times L}$ (w' $= Pw$). Then, we feed the hidden visual feature $h \in \mathbb{R}^{C \times H \times W}$ into a convolutional layers to acquire three new feature maps h_1,h_2 and h_3, where h_1,h_2,h_3 $\in \mathbb{R}^{C \times H \times W}$. Next, we calculate the i^{th} word-context vector h_s^i by adopt the attention weight of the model attend to the i^{th} word for synthesizing j^{th} sub-region of the image. The spatial-wise attention map can be calculated as: $M^S_{j,i} = \sum_{i=1}^{L} \left\{ \frac{exp(h^i_{s1} h^j_{s2})}{\sum_{i=1}^{L} exp(h^i_{s1} h^j_{s2})} w'_i \right\} \in \mathbb{R}^{L \times L}$. Finally, we set the word-context matrix for image feature by $F^{spatial-attn}(w, h) = (h^0_s, h^1_s, ..., h^i_s, ..., h^L_s) \in \mathbb{R}^{C \times H \times W}$.

Channel-wise Attention mainly focuses on 'what' is important of the image features, which attempts to allocate different significance to the channels of feature maps. The channel-wise attention can be viewed as the process of selecting semantic attributes on the demand of the word-context [18], which associates semantically meaningful parts with meaning corresponding words. Therefore, we apply the channel-wise attention module after the spatial-wise attention module.

Note that spatial-wise attention requires the visual feature to calculate the spatial attention weights, but the visual feature used in spatial attention is not attention-based [18]. Therefore, we introduce a channel-wise attention mechanism after the spatial-wise attention to attend the visual feature.

The channel-wise attention $F^{channel-attn}(w, h_s)$ has two inputs: the word features $w \in \mathbb{R}^{D \times L}$ and the word-context vector for spatial-wise attention weighted image feature $h_s \in \mathbb{R}^{C \times H \times W}$ from the spatial-wise attention module to emphasize interdependent features and improve the feature representation of specific semantic meaning. The same as spatial-wise attention module, we adopt a perceptron layer $P \in \mathbb{R}^{H \times W \times D}$ to convert the word features into the common semantic dimension as image features. Then, we calculate the channel-wise attention map $M^C_{i,j} = \sum_{i=1}^{C} \left\{ \frac{exp(h^i_s w'_j)}{\sum_{i=1}^{L} exp(h^i_s w'_l)} \right\} \in \mathbb{R}^{C \times L}$ to represent the correlation between the i^{th} channel of the image features h_s and the j^{th} word of the input text description. Finally, we set the spatial-wise mixed channel-wise attention matrix for image feature by $F^{channel-attn}(w, h_s) = (h^0_{s\&c}, h^1_{s\&c}, ..., h^i_{s\&c}, ..., h^L_{s\&c}) \in \mathbb{R}^{C \times H \times W}$.

3.3 Objective

For training the proposed method, we introduce two parts of objective loss. The first part is the common min-max function defined as common GANs. The second part is the DAMSM loss, which measures match losses of text descriptions and generated images.

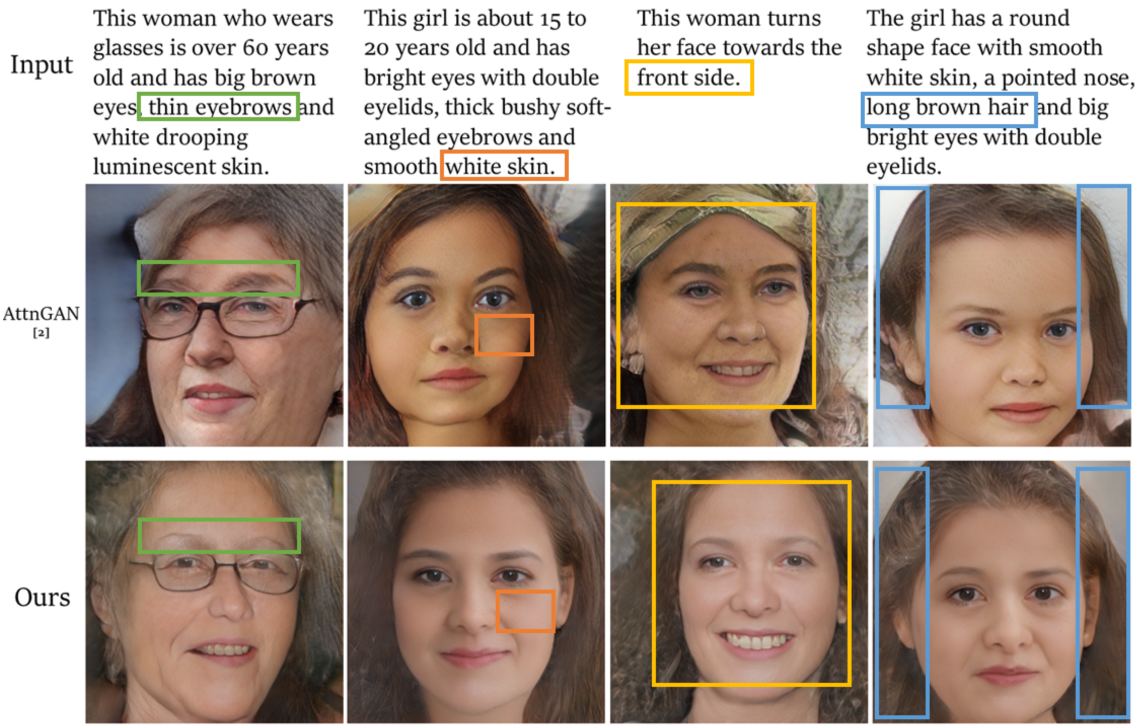
In order to generate high-quality and realistic face images with multiple levels of conditions, we introduce two parts of objective loss: the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$.

3.3.1 Generator Objective Function.

The generator loss $\mathcal{L}_G$ contains an adversarial loss $\mathcal{L}_{G_m}$ and a text-image matching loss $\mathcal{L}_{DAMSM}$ [7] is defined as:

$$\mathcal{L}_G = \sum_{m=1}^{M} \mathcal{L}_{G_m} + \lambda \mathcal{L}_{DAMSM} \qquad (1)$$

where M is the number of stage, λ is a hyper-parameter



This woman who wears glasses is over 60 years old and has big brown eyes, thin eyebrows and white drooping luminescent skin.

This girl is about 15 to 20 years old and has bright eyes with double eyelids, thick bushy soft-angled eyebrows and smooth white skin.

This woman turns her face towards the front side.

The girl has a round shape face with smooth white skin, a pointed nose, long brown hair and big bright eyes with double eyelids.

Fig. 2 Qualitative comparison results of AttnGAN [7] and our method on our FFHQ-Text dataset.

to control the relative importance of the loss. In this paper, we set λ to 5. The $\mathcal{L}_{DAMSM}$ [7] is used to estimate the correlation between the synthesis image and the input text description.

At the m^{th} stage, the adversarial loss $\mathcal{L}_{G_m}$ is defined as:

$$\mathcal{L}_{G_m} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{i}_m \sim P_{G_m}}[log(D_m(\hat{i}_m))]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{i}_m \sim P_{G_m}}[log(D_m(\hat{i}_m, t))]}_{\text{conditional loss}} \tag{2}$$

which contains unconditional loss and conditional loss. The unconditional loss estimates whether the synthetic image is real or fake. The conditional loss estimates how well the synthetic image matches the input text description.

3.3.2 Discriminator Objective Function.

For training the discriminator D to classifies both real images and fake images from the generator, the final discriminator loss $\mathcal{L}_D$ is the adversarial loss function $\mathcal{L}_{D_m}$ is defined as :

$$L_{D_m} = \underbrace{-\frac{1}{2}\mathbb{E}_{i_m \sim P_{Data_m}}[log(D_m(i_m))] - \frac{1}{2}\mathbb{E}_{\hat{i}_m \sim P_{G_m}}[1 - log(D_m(\hat{i}_m))]}_{\text{unconditional loss}}$$
$$\underbrace{-\frac{1}{2}\mathbb{E}_{i_m \sim P_{Data_m}}[log(D_m(i_m, t))] - \frac{1}{2}\mathbb{E}_{\hat{i}_m \sim P_{G_m}}[1 - log(D_m(\hat{i}_m, t))]}_{\text{conditional loss}} \tag{3}$$

where i_m is the true image from the real image distribution P_{Data_m} at the m^{th}, $\hat{i}_m$ is the synthetic image from the model distribution P_{G_m} at the m^{th}.

4. Experiments

We evaluate the proposed model on our FFHQ-Text dataset. FFHQ-Text contains 760 female portraits and each with 9 sentences. We split the FFHQ-Text dataset into 500 training images and 260 testing images. In the training step, the images are randomly chosen, and only 1 sentence is randomly selected from the 9 sentences related to the current image. In the testing step, all sentences in the test set are selected for image synthesis. We optimize our model for 1000 epochs with an Adam optimizer having hyper-parameters β_1=0.5, β_2=0.999 and learning rate 0.0002 for each generators and the discriminators.

4.1 Qualitative Results

4.1.1 Portrait Synthesis

Fig. 2 shows input text descriptions and comparison results of AttnGAN [7] and our proposed method. As the results shown in the second row, some parts are not match the input descriptions, which indicated in green, orange, yellow and blue boxes. On the contrary, our approach produces better facial images in quality and photorealism while well-conditioned to input text descriptions.

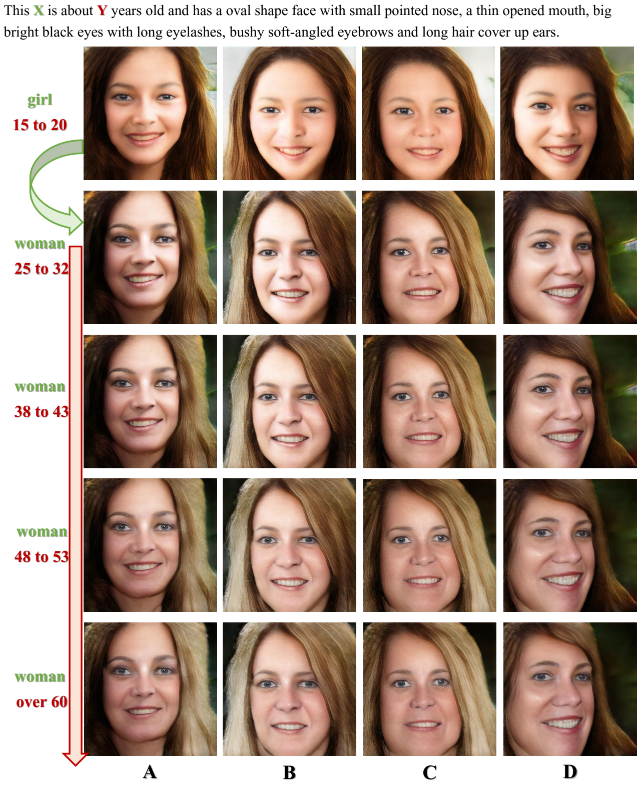

This X is about Y years old and has a oval shape face with small pointed nose, a thin opened mouth, big bright black eyes with long eyelashes, bushy soft-angled eyebrows and long hair cover up ears.
girl
15 to 20
woman
25 to 32
woman
38 to 43
woman
48 to 53
woman
over 60
A
B
C
D

Fig. 3 Qualitative results of **"Age" element conversion** on our FFHQ-Text dataset.

4.1.2 Face Comparing

We introduce a female whose appearance looks between the age classes of 0-2/4-6/8-13/15-20 is denoted as a "girl". Besides, a female whose appearance looks between the age classes of 25-32/38-43/48-53/60+ is denoted as a "woman".

Fig. 3 shows the morphing process of the generated face images. Due to the lack of the girl image dataset at age classes of 0-2, 4-6 and 8-13, we only take the experiment on face image synthesis with the "Age" element manipulate at the range of 15-20, 25-32, 38-43, 48-53, and 60+ respectively. This process also can be seen as a prediction of a girl's growth. Accordingly, we approximate this continuous transformation by converting the "Age" element and not change almost general parts of the generated facial images (*e.g.* facial orientation, face, nose, mouth, *etc.*).

Note that in our FFHQ-Text data set, the number of Asian females (less than 20%) and European females (over 65%) are unequal. Therefore, as the age changes, the experimental results show that the original young girls with brown hair and brown eyes gradually became women with golden hair and silver eyes.

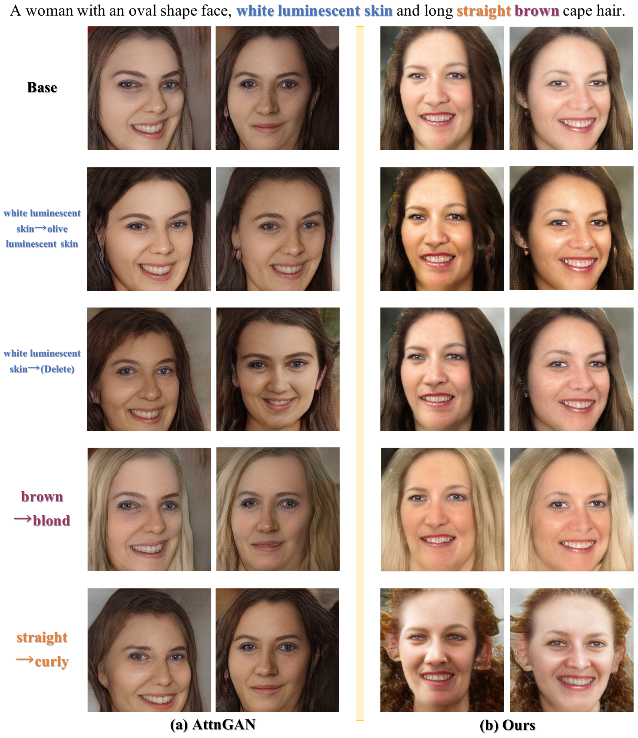

A woman with an oval shape face, white luminescent skin and long straight brown cape hair.
Base
white luminescent skin→olive luminescent skin
white luminescent skin→(Delete)
brown →blond
straight →curly
(a) AttnGAN
(b) Ours

Fig. 4 Qualitative comparison results of AttnGAN [7] and our method about **"Skin"/"Hair" element conversion** on our FFHQ-Text dataset about manipulating part of the word of input text description.

4.1.3 Facial Attribute Manipulation

Furthermore, we also make a comparison with AttnGAN on the facial attribute manipulation task as shown in Fig. 4. The first row shows the generated facial images from the basic input text description:"*A woman with an oval shape face, white luminescent skin and long straight brown cape hair*". Next, we altered four related parts in the input description, respectively. For example, we changed the "white luminescent skin" to "olive luminescent skin" in the basic text description, as shown in the first two columns in the second row, the facial images generated are not match the manipulated text well. In contrast, our proposed method can generate better results while maintaining the same character.

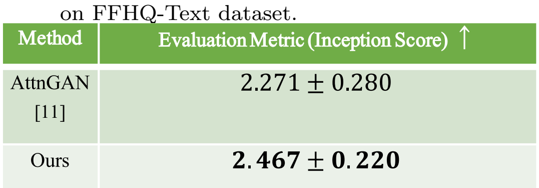
4.2 Quantitative Results

4.2.1 Inception Score (IS)

For quantitative evaluation, we adopt the Inception Score(IS) [19] as one of our evaluation metrics. This metric rewards high quality and varied images and correlates best with human judgments. A higher score means this model can generate more diverse and high-quality images. To do the comparison with AttnGAN, we select 10 partitions of 500 randomly generated samples to compute the

Inception Scores values shown in Table 1 on our FFHQ-Text dataset. Obviously, in terms of image quality, the proposed method is superior to AttnGAN.

Table 1 Inception Scores of AttnGAN and proposed method on FFHQ-Text dataset.

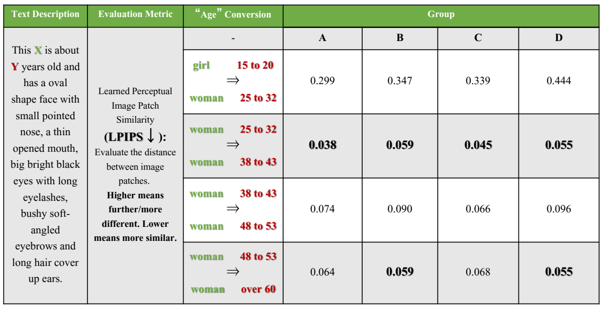


Method	Evaluation Metric (Inception Score) ↑
AttnGAN [11]	2.271 ± 0.280
Ours	$\mathbf{2.467 \pm 0.220}$

4.2.2 Learned Perceptual Image Patch Similarity (LPIPS)

In order to assess the perceptual similarity between two consecutive synthetic face images in each group in Fig. 3, we adopt a novel evaluation metric method, named Learned Perceptual Image Patch Similarity (LPIPS) metric [20]. Table 2 shows the quantitative evaluation of the qualitative results of Fig. 3 of the "Age" element conversion on our FFHQ-Text dataset. A higher value means that they are further different from each other. On the contrary, a lower value means similar to each other. Except for the similarity between the age of 15-20 and 25-32 have significantly declined when input description changing from a "girl" to a "woman," other similarities are incredibly high. The results verify that our method can hardly change other properties while only making changes in age.

Table 2 Face perceptual similarity performance comparison of our proposed method on "Age" element conversion.



Text Description	Evaluation Metric	"Age" Conversion	Group			
		-	A	B	C	D
This X is about Y years old and has a oval shape face with small pointed nose, a thin opened mouth, big bright black eyes with long eyelashes, bushy soft-angled eyebrows and long hair cover up ears.	Learned Perceptual Image Patch Similarity **(LPIPS ↓):** Evaluate the distance between image patches. **Higher means further/more different. Lower means more similar.**	**girl 15 to 20** ⇒ **woman 25 to 32**	0.299	0.347	0.339	0.444
		woman 25 to 32 ⇒ **woman 38 to 43**	**0.038**	**0.059**	**0.045**	**0.055**
		woman 38 to 43 ⇒ **woman 48 to 53**	0.074	0.090	0.066	0.096
		woman 48 to 53 ⇒ **woman over 60**	0.064	**0.059**	0.068	**0.055**

5. Conclusion

This paper proposed a text-to-face synthesis method based on spatial-wise and channel-wise mixed attention architecture that can: 1) generate high-quality and high-resolution face images, 2) manipulate the synthesized images with a partly changed text description. Our proposed method fully utilizes the contextual information between word features and image features in each channel, effectively distinguishing different facial attributes with corresponding semantic meaning. To evaluate the effectiveness of our model, we introduced a novel text-to-face synthesis dataset, which is manually annotated textual description sentences for high-resolution face images. For the text descriptions about a large number of face images described by a few people, we manually annotate from different perspectives to avoid indicating almost the same descriptions for one image and ensure the diversity of generated face images. Compared with AttnGAN [7] on our FFHQ-Text dataset, the proposed method can demonstrate better experimental results qualitatively and quantitatively. We can also effectively preserve the facial attributes that are not mentioned to manipulate when changing specific words in the input text descriptions.

For our future work, we aim at utilizing attention-based sentence-level annotations to facilitate the generation of story-level photo-realistic face images from multi textual descriptions. Moreover, to make the generated images more accurate, we will balance the number of facial images of each race in the future. We also expect that the face image quality performance and control appearance of the proposed model can be further improved. Therefore, it is important to expand the text-to-face synthesis dataset (FFQH-Text) to a sufficiently large scale by more workers annotating text descriptions from different perspectives of other kinds of people.

References

[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

[2] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

[3] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

[4] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014): 2672-2680.

[5] Reed, Scott, et al. "Generative adversarial text to image synthesis." arXiv preprint arXiv:1605.05396 (2016).

[6] Zhang, Han, et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[7] Xu, Tao, et al. "Attngan: Fine-grained text to image generation with attentional generative adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[8] Qiao, Tingting, et al. "Mirrorgan: Learning text-to-image generation by redescription." Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[9] Wah, Catherine, et al. “The caltech-ucsd birds-200-2011 dataset.” (2011).

[10] Nilsback, Maria-Elena, and Andrew Zisserman. “Automated flower classification over a large number of classes.” 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. IEEE, 2008.

[11] Lin, Tsung-Yi, et al. “Microsoft coco: Common objects in context.” European conference on computer vision. Springer, Cham, 2014.

[12] Karras, Tero, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

[13] Kurach, Damian, Danuta Rutkowska, and Elisabeth Rakus-Andersson. ”Face classification based on linguistic description of facial features.” International Conference on Artificial Intelligence and Soft Computing. Springer, Cham, 2014.

[14] Schuster, Mike, and Kuldip K. Paliwal. ”Bidirectional recurrent neural networks.” IEEE transactions on Signal Processing 45.11 (1997): 2673-2681.

[15] Szegedy, Christian, et al. ”Rethinking the inception architecture for computer vision.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[16] Russakovsky, Olga, et al. ”Imagenet large scale visual recognition challenge.” International journal of computer vision 115.3 (2015): 211-252.

[17] Woo, Sanghyun, et al. ”Cbam: Convolutional block attention module.” Proceedings of the European conference on computer vision (ECCV). 2018.

[18] Chen, Long, et al. ”Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[19] Salimans, Tim, et al. ”Improved techniques for training gans.” arXiv preprint arXiv:1606.03498 (2016).

[20] Zhang, Richard, et al. ”The unreasonable effectiveness of deep features as a perceptual metric.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.