Controllable Facial Attribute Editing via Gaussian Mixture Model Disentanglement

Bo Li[a], Shu-Hai Deng[a], Bin Liu[*a], Yike Li[a], Zhi-Fen He[a], Yu-Kun Lai[b], Congxuan Zhang[c], Zhen Chen[c]

[a]*School of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China*
[b]*School of Computer Sciences and Informatics, Cardiff University, Cardiff, UK*
[c]*Key Laboratory of Nondestructive Testing Ministry of Education, Nanchang Hangkong University, Nanchang, China*

Abstract

Generative adversarial networks (GANs) have made much progress in the field of high-quality and realistic facial image synthesis in recent years. However, compared with their powerful generation ability, it is difficult for users to modify the desired attributes of the resulting image while keeping the others. How to disentangle the latent space of pre-trained GANs is essential and critical for controllable image synthesis. In this paper, a novel controllable facial attribute editing algorithm based on the Gaussian mixture model (GMM) representation is proposed. First, we assume that the latent variables with respect to each facial attribute lie in a subspace of the whole latent manifold composed of a fixed number of learned features, and each attribute subspace can be modeled by a GMM. Then, to avoid unintended changes during attribute editing, a coordinate accumulation strategy with orthogonal regularization is introduced to enhance the independence of distinct attribute subspaces which helps improving the controllability of attribute editing. In addition, a resampling strategy is utilized to improve the stability of the model. Through qualitative and quantitative experimental results, the proposed method achieves the state-of-the-art performance on facial attribute editing, and improves the controllability of desired attribute editing.

Keywords: Gaussian mixture model, latent space decoupling, orthogonal constraints on manifold spaces, semantic editor

1. Introduction

The basic principle of generative adversarial neural networks (GANs) is to learn the mapping function from latent space to real data space through adversarial training [1]. Once the mapping function is learned, GANs can generate realistic images through random sampling in the latent space. Although GANs have made significant progress in the field of high quality facial image synthesis, how to effectively control the attributes of the generated results is still a challenge. For example, when using GANs to generate a face image, various facial attributes (expression, gender, etc.) are entangled together and difficult to conduct desired attribute editing while keeping the other attributes.

In order to enhance the controllability, some researchers [2–5] strive to explore the latent space of GANs by linear projection. They argue that the subspace after projection can represent some semantic information with respect to facial attributes, and the semantic information learned by GANs can be reused to reasonably control the image generation process. GANSpace [4] performed principal component analysis (PCA) on the latent variables sampled

* Corresponding author.
E-mail addresses: nyliubin@nchu.edu.cn

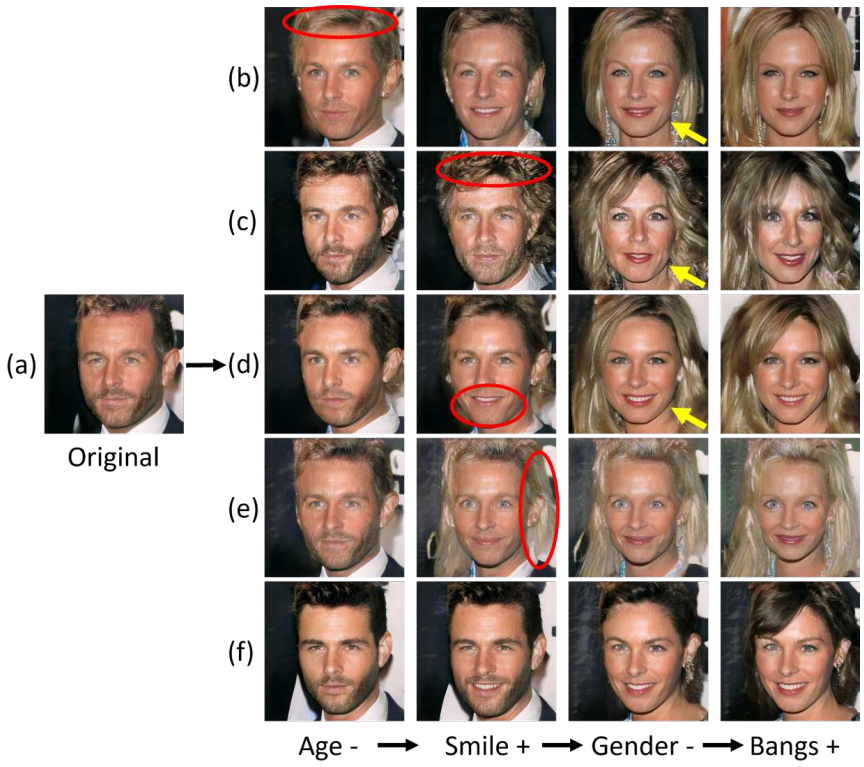

Figure 1: The results of facial attributes editing by various methods. (a) The original image, (b) - (f) are respectively the editing results by GANSpace [4], SeFa [5], InterFaceGAN [3], EYE [6] and the proposed method. Each column represents the results of manipulating a specific attribute by different methods. "+" or "-" means the editing direction of desired attribute, which represents "strengthening" or "weakening", and "→" shows the effect before and after editing. Yellow arrow represents the pose variation and red circle means that this character has changed. Our method improves the controllability of desired attribute editing while keeping the others.

from the latent space of a pre-trained GAN and found some editable directions corresponding to semantic attributes through manipulating eigenvalues and eigenvectors. In SeFa [5], a general unsupervised closed-form factorization method was proposed for latent semantic discovery. It aims at finding the main directions which cause the most significant change of the projected latent code. InterFaceGAN [3] utilized support vector machines to learn the boundary hyperplanes of different semantic attributes with the supervision of pre-defined attribute tags, then the facial attributes can be edited by moving along the normal vectors of the hyperplanes. Enjoy Your Editing (EYE) [6] can edit the facial attributes by manipulating the column of a matrix, which integrates the directions of all attributes. However, due to the high-dimensional and nonlinear entangling property, the above methods based on linear projection are difficult to disentangle complex semantic attributes, as shown in Fig. 1. The existing methods tend to change other facial attributes of the original image when editing one attribute[1], such as the pose (GANSpace, SeFa, InterFaceGAN) or identity information (SeFa). Another example is that these methods may change people's hair or beard when editing "smile" attribute. In contrast to the above models, our method can not only improve the controllability of desired attribute editing, but also maintain other attributes well as shown in Fig. 1(f).

In this paper, a novel facial attribute editing algorithm based on the Gaussian mixture model (GMM) representation is proposed to decouple the latent space of pre-trained GANs. We assume that the latent codes with respect to each facial attribute lie in a subspace of the whole latent manifold of a pre-trained GAN model, such as PGGAN [7] or StyleGAN [8], and each attribute subspace can be modeled by a GMM. In order to improve the disentanglement performance, a coordinate accumulation strategy with orthogonal regularization is proposed to enhance the independence

[1]It is noted that, for fairness, the editing results of different methods are selected when the edit amplitude of the specific attribute reaches about 80%.

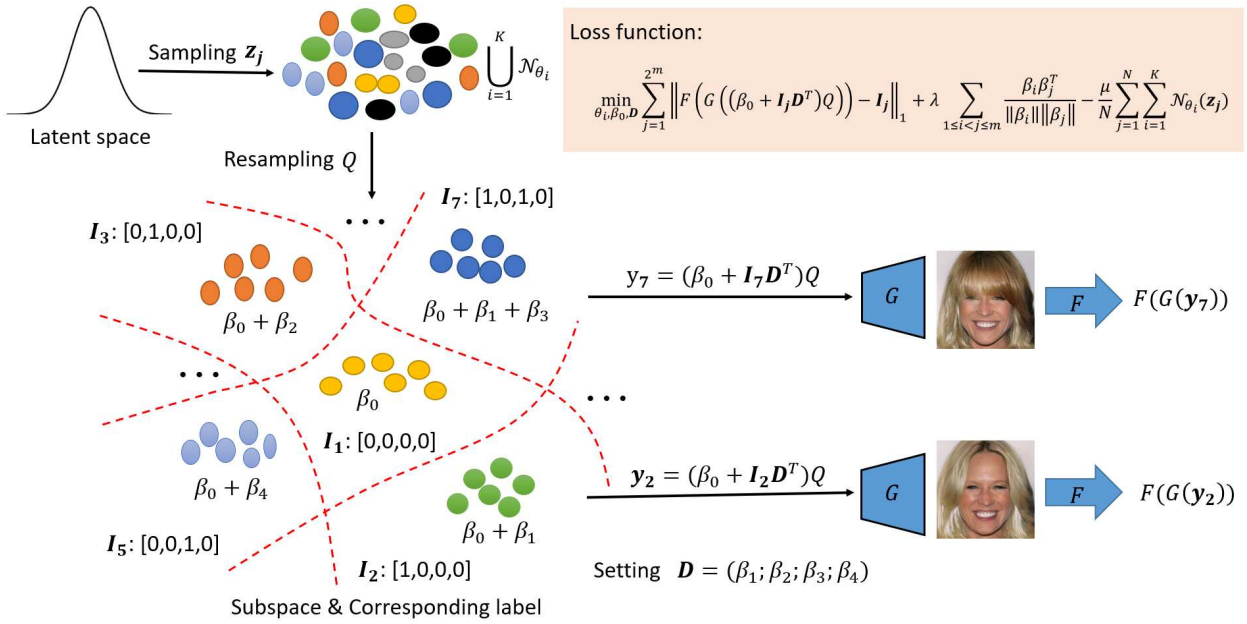

Figure 2: Latent space disentanglement based on GMMs. Given the latent space of a pre-trained GAN, we first use maximum likelihood method to estimate the basis feature distribution functions, and then semantic subspace for each attribute can be built by a GMM, and meaningful semantic directions are explored by a coordinate accumulation strategy.

between different subspaces.

Our contributions are summarized below:

- A novel controllable facial attribute editing algorithm based on Gaussian mixture model (GMM) representation is proposed.
- A coordinate accumulation strategy with orthogonal regularization is proposed to improve the disentangling performance.
- Compared with the state-of-the-art supervised and unsupervised approaches, our method gains better performance on the controllability of specific facial attribute editing.

2. Related Work

In recent years, GANs have attracted widespread attention due to their great potential in generating high-quality realistic images [7–11]. It usually takes a hidden vector (a.k.a. latent code) sampled from a Gaussian distribution as input and outputs a realistic synthesized image. The generation performance of GANs has made great progress in recent years, e.g., PGGAN [7] and StyleGAN [8] can produce high quality facial images with 1024×1024 resolution. However, compared with the powerful generation ability, how to control or edit the generation results without repeated training is still an open problem, i.e., for a generated facial image, users may attempt to modify one or more attributes, such as expression, gender or age as shown in Fig. 1, while retaining other information of the input face.

An intuitive way for controllable attribute editing is training GANs with attribute value as a condition [12–16]. For example, SaGAN [13] improved the controllability of GANs by introducing the spatial attention mechanism into GAN framework, which focuses on altering the attribute specific regions while keeping the rest unchanged. MagGAN [17] introduced a fine-grained facial attribute manipulation strategy by a mask-guided reconstruction loss. Cascade EF-GAN [18] used a progressive facial expression editing strategy with local focuses to solve the artifacts and blurs during attribution manipulation. Wu et al. [19] presented an unsupervised facial expression editing techniques by disentangling the identity and expression of a facial image. To speed up the editing process of facial image, Lin et

al. [20] proposed Anycost GAN based on quick preview features in modern rendering software. In [16], a novel latent space factorization model L2M-GAN was proposed. It is trained end-to-end and improves the edit controllability by disentangling attribute-relevant and irrelevant features. However, the GANs are required to be trained repeatedly to accomplish the controllable editing in these models.

Some recent research has indicated that GANs have actually learned some meaningful semantics in the latent space during the training process. Therefore, the generation results can be potentially controlled and edited by exploring the latent space of a pre-trained GAN without retraining the model. Bau et al. [21, 22] explored the role of each activation in the generation process of GANs with the supervision of semantic segmentation, which is able to control certain objects by modifying some activations. Yang et al. [2] explored the causality between the activations and semantics occurring in the output image quantitatively by evaluating the layer-wise semantic representations at different abstraction levels. For real image editing, GANs inversion methods [23–29] are studied to find the corresponding latent variables and activations for follow-up semantic attribute editing. GANSpace [4], SeFa [5], InterfaceGan [3] and EYE [6] are the most related work to this paper. GANSpace and SeFa are unsupervised analysis methods, while InterFaceGAN is trained with semantic tags as supervision. In GANSpace, the principal axes were found by applying PCA transformation in the latent space, while the interpretable directions can be directly computed in a closed-form solution in SeFa. The support vector hyperplanes are learned with the supervision of semantic attributes in InterfaceGAN, and users can edit the facial attributes by moving along the normal of hyperplanes. EYE [6] can edit the facial attributes by manipulating the column of a directional matrix. However, confined by the linear property, these four methods based on linear projection are difficult to disentangle complex semantic attributes.

In order to explore the physical mechanism and logic relationship contained in big data, the disentangled representation learning aims to mine the multi-level and multi-scale latent generative factor and promotes the development of deep learning. Bengio et al. [30] first introduced the concept of disentanglement representation in the survey of representation learning. Locatello et al. [31] explained the disentanglement representation learning from data generation. Bouchacourt et al. [32] presented a Multi-Level Variational Autoencoder (ML-VAE) for learning a disentangled representation of grouped data. Sanchez et al. [33] proposed to use mutual information to disentangle the common information and specific information for images sharing some attributions. In this paper, a coordinate accumulation strategy with orthogonal regularization is utilized to enhance the independence of disentanglement.

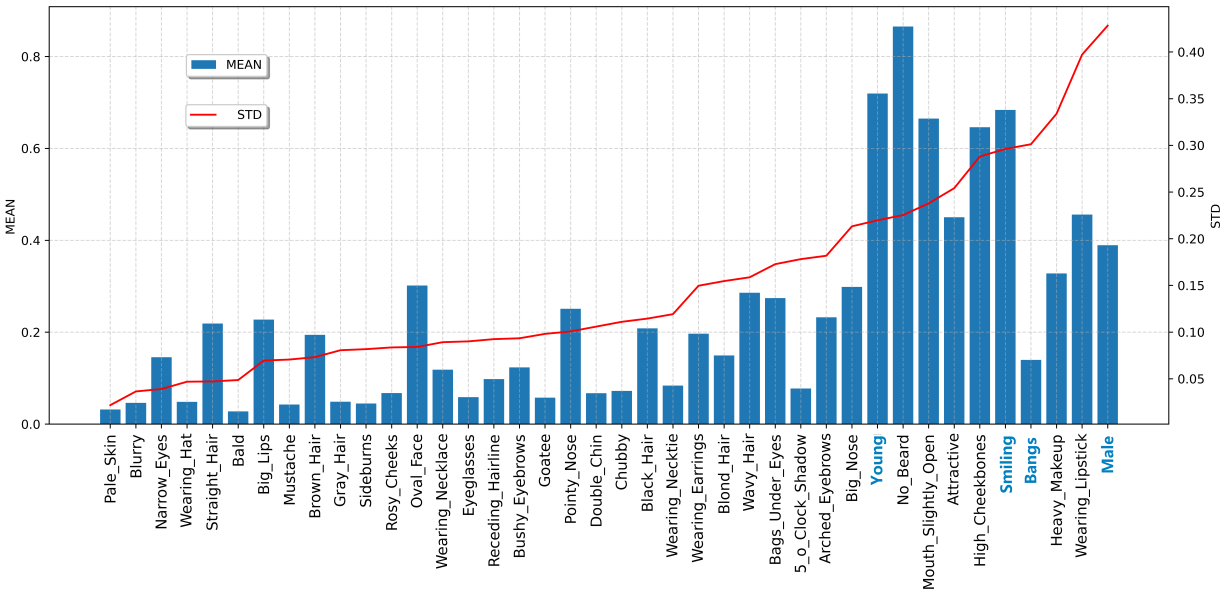

Figure 3: The mean value and standard deviation of each attribute score in sampled 500k images.

3. Approach

In this section, a novel controllable facial attribute editing algorithm based on GMM representation is proposed. The whole framework is shown in Fig. 2. We assume that the whole facial attribute manifold space can be represented by a group of basis features, i.e., eyes, mouth, cheek, color and shape of hair, etc. In order to enhance the diversity of semantic editing, each basis feature is represented by a Gaussian distribution, and the parameters of each Gaussian function can be learned through maximum likelihood estimation (MLE), as shown in Fig. 2. Based on the above assumption, each latent code $\mathbf{z}$ sampled from the latent space of a pre-trained GAN will lie in the union distribution space of the basis features. Furthermore, the latent codes corresponding to each attribute will compose a subspace, and each attribute subspace can be represented by a linear combination of basis features, which can be formulated as a Gaussian mixture model. In this paper, a coordinate accumulation strategy with orthogonal regularization is utilized to improve the disentanglement of distinct attributes. Note that, all analytical and experimental results in this paper are conducted on the two popular pre-trained GAN models, including PGGAN [7] and StyleGAN [8].

3.1. Training Data Collection

The training data is collected similar to InterfaceGAN [3]. For the given pre-trained PGGAN model [7] or StyleGAN model [8], a high-quality facial image $\mathbf{X}$ can be generated by

$$\mathbf{X} = G(\mathbf{z}) \tag{1}$$

where $G(\cdot)$ is the generator function of PGGAN or StyleGAN, $\mathbf{z} \in \mathbb{R}^d$ is the latent code sampled from a normal distribution, and $\mathbf{X}$ is the generated facial image with resolution 1024×1024 in PGGAN or 256×256 in StyleGAN. Then, a pre-trained facial attribute classification network F [34] is utilized to predict the semantic attribute score of a generated face image,

$$\mathbf{s} = F(\mathbf{X}) = F(G(\mathbf{z})) \tag{2}$$

where $\mathbf{s} \in \mathbb{R}^m$ represents the scores of face image $\mathbf{X}$ on m facial attributes. Each element of vector $\mathbf{s}$ in the range of $[0, 1]$ indicates the probability to the specific attribute. For example, the gender attribute can be defined from female (0) to male (1). If the two opposite situations of each attribute are taken into account, the whole facial attribute manifold can be split into 2^m subspaces. The main purpose of this paper is to learn the disentangled attribute subspace representation.

In order to make sure that the distribution of the sampled codes can be expected to cover all of the attribute subspaces, we randomly sample 500K latent codes and predict the attribute scores according to Eq. 2. Then, we calculate the mean value and standard deviation for each attribute, see Fig. 3. Based on the attribute distribution (big standard deviation) and visual effect (easy to show), 4 primary attributes including smile, gender, bangs, age are chosen in this paper, which implies that the whole facial attribute manifold will be split into 2^4 attribute subspaces, as shown in Fig. 2.

In this paper, each attribute subspace can be represented by a 4-d indicator vector $\mathbf{I}$, whose element is either 0 or 1 indicating the negative or positive of this attribute. The meaning of each dimension is respectively smile, gender, bangs and age. For example, $[0, 0, 0, 0]$ represents "the smile is negative (calm); gender is female; bangs do not exist; age is negative (old)", while $[1, 1, 1, 1]$ indicates "the smile is positive (laugh); gender is male; bangs exist; age is positive (young)".

To reduce ambiguous samples, e.g., middle-aged person for age attribute, we pick 2K samples with the highest score and 2K with the lowest score for each attribute, and a total of 4K samples with significant attributes are selected.

3.2. Basis Feature Distribution Learning based on MLE

Assuming that the whole facial attribute space can be represented by a group of learnable basis features, which is very popular in sub-manifold learning. In order to enhance the diversity of semantic editing, each basis feature is represented by a multivariate Gaussian distribution $\mathcal{N}_{\theta_i}$, where $\theta_i = (\mu_i, \Sigma_i)$ is the mean and standard deviation parameters. Therefore, each latent code $\mathbf{z}$ should lie in the union space spanned by the whole basis features,

$$\mathbf{z} \sim \bigcup_{i=1}^{K} \mathcal{N}_{\theta_i}, \tag{3}$$

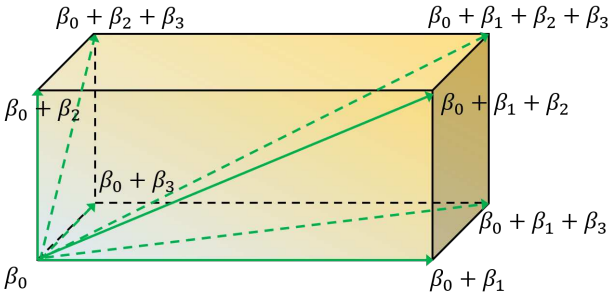

Figure 4: Illustration of coordinate accumulation in the case of $m = 3$.

where, K is the number of Gaussian distributions.

To ensure that the basis features can cover the latent space of GANs, maximum likelihood estimation (MLE) is used to learn the parameters of each distribution,

$$\begin{aligned} \max_{\theta_i} \frac{1}{N} \sum_{j=1}^{N} p(\mathbf{z}_j \sim \bigcup_{i=1}^{K} \mathcal{N}_{\theta_i}) \\ = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{K} \mathcal{N}_{\theta_i}(\mathbf{z}_j) \end{aligned}, \tag{4}$$

where, N represents the number of latent code samples and it is set as 16K in our experiments, $\mathbf{z}_j$ is the latent variable with significant facial attributes and p is the probability function.

3.3. Attributes Disentanglement based on GMM

In this section, we assume that the latent codes corresponding to each attribute lie in a subspace of the whole basis feature manifold. Therefore, each attribute subspace $\mathcal{A}_j$ can be represented by a linear combination of basis features, which can be formulated as a GMM,

$$p(\mathbf{y} \in \mathcal{A}_j) = \sum_{i=1}^{K} \alpha_j^{(i)} \mathcal{N}_{\theta_i}(\mathbf{y}) , \tag{5}$$

where α_j is the learnable coefficients of the GMM for the j-th attribute subspace, and $\alpha_j^{(i)}$ is its i-th component. In order to ensure the latent codes in subspace $\mathcal{A}_j$ comply with the desired attribute $\mathbf{I}_j$, an objective function is designed as follows

$$\min_{\alpha_j} \|F(G(\mathbf{y})) - \mathbf{I}_j\|_1, \ \forall \mathbf{y} \in \mathcal{A}_j . \tag{6}$$

For solving the optimization problem of a continuous normal distribution involved in Eq. 6, a discrete resampling strategy is introduced. A random vector from each basis feature distribution $\mathcal{N}_{\theta_i}$ is sampled, then a basis matrix $\mathbf{Q}$ is constructed with each sampled vector as a row. Combined with Eq. 5, the objective function on the whole 2^m subspaces can be reformulated as

$$\min_{\alpha_j} \sum_{j=1}^{2^m} \|F(G(\alpha_j \mathbf{Q})) - \mathbf{I}_j\|_1 . \tag{7}$$

In order to decouple the semantic attribute subspaces more effectively, a coordinate accumulation strategy with orthogonal regularization is proposed as shown in Fig. 4. Let β_0 represent the basis attribute, i.e., the dumb attribute with indicator function $[0, 0, 0, 0]$ as shown in the last row of Fig. 5. Then the specific attribute subspace $\mathcal{A}_j$ can be found by $\alpha_j = \beta_0 + \mathbf{I}_j(\beta_1; \beta_2; ...; \beta_m)^T$, where $\beta_1, \beta_2, ..., \beta_m$ are the semantic coordinates of m attributes derived from the basis attribute β_0. Once the model is trained well, these directions can be utilized to edit the corresponding semantic attributes of an image.

To further improve the disentanglement performance, an orthogonal regularization between semantic coordinates $\beta_1, \beta_2, ..., \beta_m$ is introduced by minimizing the discrete cosine distance. Finally, the objective function in Eq. (7) can be reformulated as

$$\min_{\beta_0, \mathbf{D}} \sum_{j=1}^{2^m} \|F(G((\beta_0 + \mathbf{I}_j \mathbf{D}^T)\mathbf{Q})) - \mathbf{I}_j\|_1 + \lambda \sum_{1 \le i < j \le m} \frac{\beta_i \beta_j^T}{\|\beta_i\| \|\beta_j\|}, \tag{8}$$

where λ is a regularization parameter, $\mathbf{D} = (\beta_1; \beta_2; ...; \beta_m)$.

Finally, the whole loss function can be formulated as follows

$$\min_{\theta_i, \beta_0, \mathbf{D}} \sum_{j=1}^{2^m} \|F(G((\beta_0 + \mathbf{I}_j \mathbf{D}^T)\mathbf{Q})) - \mathbf{I}_j\|_1 + \lambda \sum_{1 \le i < j \le m} \frac{\beta_i \beta_j^T}{\|\beta_i\| \|\beta_j\|} - \frac{\eta}{N} \sum_{j=1}^{N} \sum_{i=1}^{K} \mathcal{N}_{\theta_i}(\mathbf{z}_j), \tag{9}$$

where hyperparameters λ and η are set to 1 in our paper, and satisfactory results can be obtained after training on 40K resampled points.

Figure 5: Illustration of the central sampling of each attribute subspace. For example, all the central images with the first element is 1 in the indicator vector are laughing and all the central images with the third element is 1 have bangs.

4. Experiments

We conducted experiments with PyTorch and Python3.8 on a PC equipped with a GTX 2080Ti, 32GB RAM and 3.6GHz Intel Core i9-9900K CPU. Adam optimizer with learning rate 5×10^{-3} is used to train our model and the optimization process takes about 8 hours for 4 facial attribute disentanglement in the CelebAhq1024-PGGAN model [7], and about 6 hours for ffhq256-StyleGAN model [8]. Due to the layout, we only show the experimental results on PGGAN in the manuscript, and the experimental results on StyleGAN are shown in the supplementary material.

4.1. Effectiveness of Attribute Decoupling

Tag consistency. As we select 4 facial attributes, smile, gender, bangs and age, a total of 2^4 attribute subspace can be obtained. In this section, an experiment is designed to evaluate the decoupling performance of the proposed

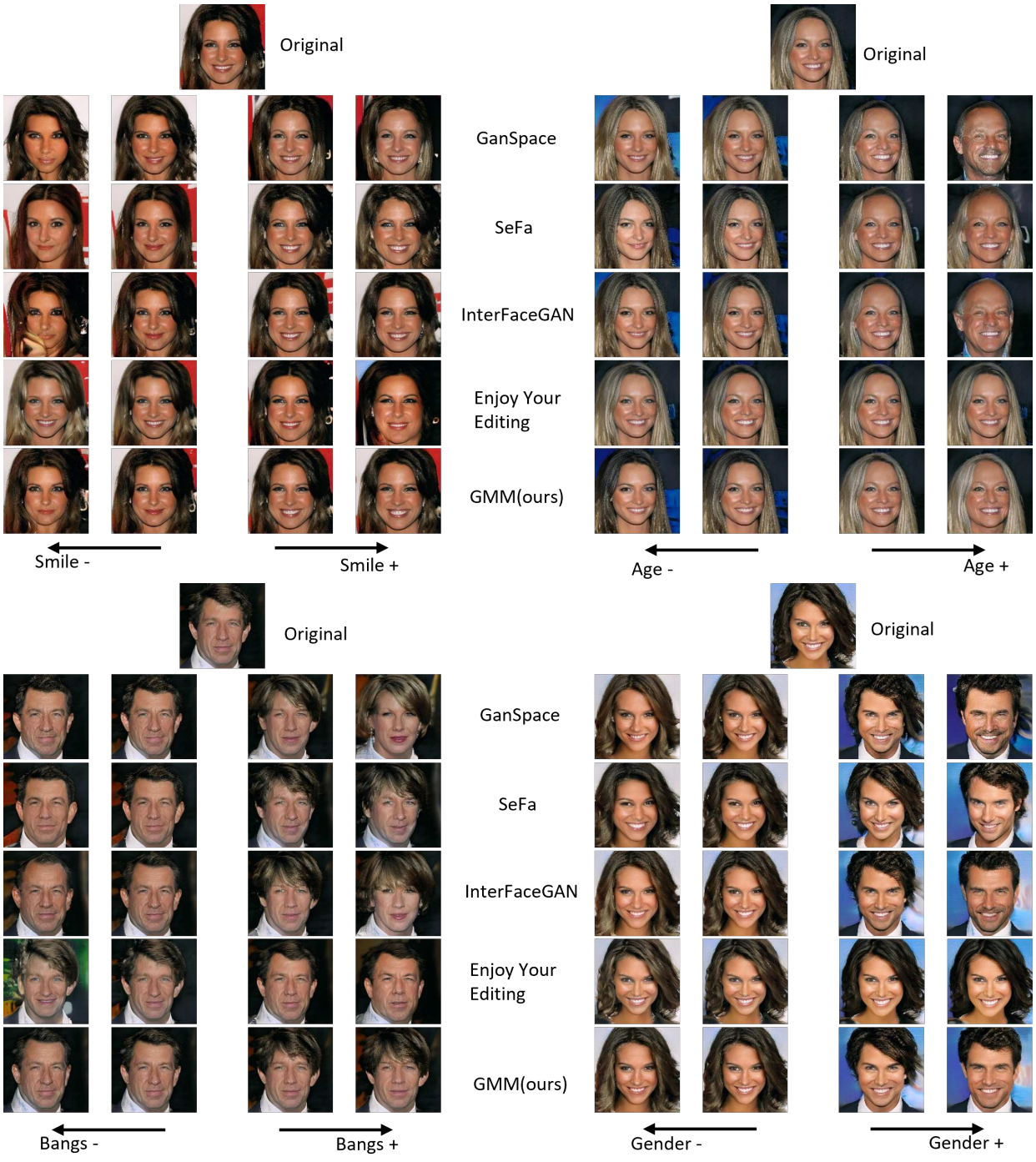

Figure 6: Single attribute editing results for PGGAN model. Each row represents the effect of editing in the opposite directions by different methods.

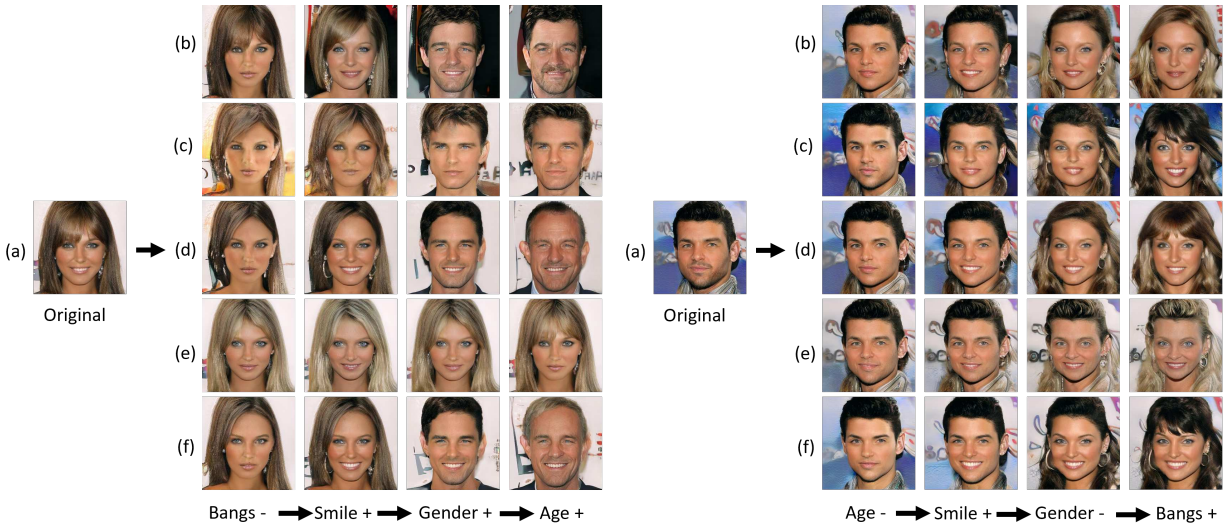

Figure 7: Multiple attribute editing results for PGGAN. (a) The original images, (b) - (f) are respectively the editing results of GANSpace, SeFa, InterFaceGAN and Our results. "+" or "-" means the editing direction of desired attribute and "→" shows the effect before and after editing.

method by checking the central samples of each attribute subspace. The mean vector μ_i of each feature distribution $\mathcal{N}_{\theta_i}$ is sampled and concatenated in a row to construct the sample matrix $\hat{\mathbf{Q}}$. Then the central image of attribute subspace $\mathcal{A}_j$ can be generated by $G((\beta_0 + \mathbf{I}_j\mathbf{D}^T)\hat{\mathbf{Q}})$ as introduced in Eq. (8). The central images are illustrated in Fig. 5. The indicator vector $\mathbf{I}_j$ below each image shows the attribute information of the space, which are (smile, gender, bangs, age). We can find that the central image generated by each subspace does match the indicator tag, and has the most distinctive corresponding attribute, which validates the effectiveness of attribute decoupling.

Identify consistency. In order to illustrate that the proposed method could maintain the identify information well during editing, we utilize a popular face verification algorithm [35], which is used to obtain the identify similarity between two images. The algorithm states that a threshold of 1.1 would classify each pair correctly. Table. 1 shows the mean identity discrepancy on the 1000 images during editing with different step length. We can find that almost all indicators are below 1.1 except for some extreme cases with big step length. For example, when the "gender" attribute is edited with step length 3, the identity discrepancy is over 1.1. We analyze that it is reasonable because big change of "gender" easily affects the variation of other attributes, such as hair, skin color, etc. The experimental results validate the effectiveness of the proposed method on identity preserving during editing process.

	Step length	-3	-2.4	-1.8	-1.2	-0.6	0.6	1.2	1.8	2.4	3
1	Bangs	0.668	0.605	0.528	0.433	0.303	0.338	0.531	0.685	0.806	0.898
2	Age	1.101	0.992	0.846	0.651	0.403	0.389	0.609	0.771	0.892	0.982
3	Gender	1.024	0.968	0.887	0.755	0.526	0.559	0.826	0.977	1.062	1.113
4	Smile	0.928	0.834	0.715	0.562	0.364	0.353	0.526	0.650	0.744	0.817

Table 1: ID discrepancy during the attribute manipulation for PGGAN model. Larger number means lower similarity.

4.2. Latent Space Attribute Editing

As illustrated in Fig. 4, β_i implies the editing direction of the i-th facial attribute, while the combination of different β_i refers to multiple attributes editing. Hence, we can conduct the attribute editing through the following formula

$$\mathbf{z}^{'} = \mathbf{z} + \sum_{i=1}^{m} \omega_i \frac{\beta_i \hat{\mathbf{Q}}}{\|\beta_i \hat{\mathbf{Q}}\|}, \tag{10}$$

where $\mathbf{z}$ is the initial latent code, $\omega_i \in [-3, 3]$ represents the step length of edit along the i-th attribute, $\hat{\mathbf{Q}}$ is constructed in the same way as the previous section. Positive values indicate that the attribute is "strengthened" and negative means "weakened".

In the following, the editing performance of the proposed method is compared with the state-of-the-art methods: GANSpace [4], SeFa [5], InterFaceGAN [3] and EYE [6], qualitatively and quantitatively.

4.2.1. Single Attribute Editing Comparisons

By setting the coefficients of other attributes to 0 in Eq. 10, we can conduct single attribute editing. Fig. 6 shows the editing results on each attribute for PGGAN model. We can find that the "gender" attribute is tangled with "bangs" and "age" in both GANSpace and InterfaceGAN, while SeFa introduces numerous unrelated changes, such as the hairstyle when editing "age", or face shape while editing "bangs". Similar editing results appear on StyleGAN model (in supplementary material). A similar thing happens with the EYE, for example, editing bangs in PGGAN causes age changes, and it is difficult to effectively edit other attributes. Compared with the state-of-the-art methods, the proposed method can decouple the attributes better and preserve the other attributes well when editing one attribute.

4.2.2. Multiple Attributes Editing Comparisons

Similarly, we can conduct multiple attributes editing by the combination of different editing directions as shown in Eq. 10. In this experiment, the attributes are edited cascadedly. Some cases of desired editing are shown in Figs. 1 and 7. It is obvious to find that the proposed method outperforms the other state-of-the-art methods in most samples, and can maintain other attributes well when editing specified attribute.

4.2.3. Quantitative Comparison

In this section, we conduct quantitative experiments to evaluate the disentanglement performance with two popular metrics. In addition, a subjective user study is also designed for further evaluation.

	Method	Smile					Gender				
		0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1
1	GANSpace	0	0.179	0.258	0.321	0.380	0	0.422	0.525	0.599	0.667
2	SeFa	0	0.214	0.304	0.380	0.453	0	**0.409**	**0.503**	**0.571**	**0.632**
3	InterFaceGAN	0	0.169	0.252	0.320	0.393	0	0.440	0.546	0.631	0.716
4	EYE	0	0.267	0.461	0.600	0.758	0	0.562	0.647	0.700	0.747
5	ours(abliation)	0	0.181	0.259	0.325	0.383	0	0.461	0.562	0.636	0.701
6	ours	0	**0.142**	**0.219**	**0.284**	**0.350**	0	0.426	0.526	0.601	0.672
		Bangs					Age				
		0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1
1	GANSpace	0	0.381	0.478	0.541	0.620	0	0.339	0.572	0.763	0.936
2	SeFa	0	0.407	0.451	0.478	0.503	0	0.211	0.350	0.473	0.598
3	InterFaceGAN	0	0.410	0.455	0.486	0.515	0	0.245	0.411	0.568	0.724
4	EYE	0	0.320	0.407	0.462	0.512	0	0.329	0.520	0.675	0.816
5	ours(abliation)	0	0.306	0.337	0.358	0.376	0	0.192	0.319	0.433	0.546
6	ours	0	**0.261**	**0.299**	**0.328**	**0.350**	0	**0.163**	**0.282**	**0.397**	**0.524**

Table 2: Mean-AD vs. edit amplitude of the corresponding facial attribute demonstrated for PGGAN model. Lower mean-AD indicates better disentanglement.ours(ablation) means ours without coordinate accumulation strategy .

Disentanglement Accuracy. Two metrics are utilized in this paper, Mean-AD and re-scoring analysis. Mean-AD referred to as mean attribute dependency (AD) proposed in [36]. It measures the influence of an editing along a certain direction to other irrelevant attributes, which is calculated by the score difference of attribute classifier F for those attributes. Smaller AD implies better disentangled performance. The Mean-AD is computed and shown in Tab. 2. It can be seen that, compared with other methods, the average AD of the proposed method is smaller for most

GANSpace	smile	gender	bangs	age
smile	0.42	0.14	0.04	0.2
gender	0.06	0.41	0.13	0.61
bangs	0.07	0.14	0.15	0.21
age	0.19	0.31	0.08	0.46

SeFa	smile	gender	bangs	age
smile	0.5	0.21	0.08	0.36
gender	0.15	0.53	0.14	0.3
bangs	0.15	0.12	0.36	0.26
age	0.17	0.24	0.17	0.6

InterFaceGAN	smile	gender	bangs	age
smile	0.63	0.15	0.1	0.26
gender	0.12	0.68	0.18	0.55
bangs	0.1	0.14	0.6	0.22
age	0.17	0.33	0.19	0.7

EYE	smile	gender	bangs	age
smile	0.3	0.19	0.09	0.43
gender	0.24	0.28	0.05	0.85
bangs	0.14	0.06	0.15	0.33
age	0.25	0.34	0.03	0.77

Ours(ablation)	smile	gender	bangs	age
smile	0.55	0.25	0.2	0.21
gender	0.1	0.6	0.11	0.15
bangs	0.08	0.25	0.73	0.31
age	0.33	0.64	0.42	0.52

Ours	smile	gender	bangs	age
smile	0.63	0.14	0.07	0.22
gender	0.06	0.63	0.06	0.22
bangs	0.05	0.14	0.62	0.13
age	0.21	0.3	0.13	0.68

Figure 8: Re-scoring analysis on the semantic manipulation achieved by different methods on PGGAN model.

facial attributes, which means the proposed method has better performance on attribute disentanglement and subspace segmentation.

In order to evaluate the disentangling performance between the current editing attribute and each of the rest attributes, another popular metric referred to re-scoring analysis used in [3, 5] is adopted. Re-scoring analysis is designed to quantitatively evaluate whether the editing along the identified directions can properly represent the corresponding attributes. Each image is edited distinguishably along a certain attribute direction, and we record the score difference along each attribute. Finally, the average scoring is computed and shown in Fig. 8. It can be obviously found that the proposed method preserves irrelevant attributes the best while editing along an identified attribute direction in most cases.

User Study. To evaluate attribute editing results under human perception, we randomly choose 20 edited images for each attribute from 1000 images, and 100 users are asked to select the best edited image among all methods. To avoid bias, we randomise the order of image pairs shown to the participants and their left/right positions. User evaluation criteria are defined as: 1) whether the attribute is edited correctly; 2) whether the area irrelevant to the attribute is reserved; 3) whether the edited image is realistic. Fig. 9 shows the results over all users, it shows that our method outperforms the compared methods on user study.

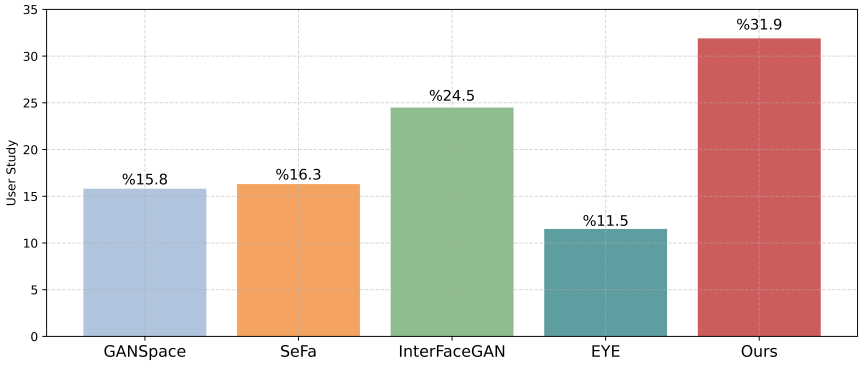

Figure 9: User Study results of facial attribute editing on four attributes: smile, bangs, age, gender.

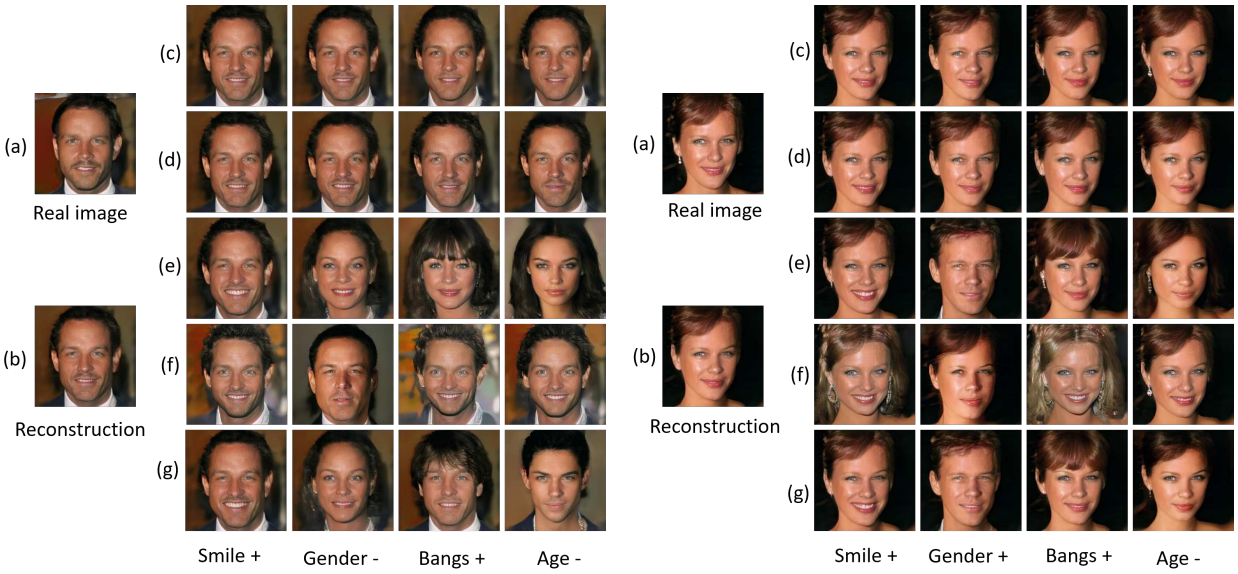

Figure 10: Editing results on real images for the pre-trained PGGAN model. (a) The real image, (b) The reconstruction image, (c) - (g) are respectively the editing results of GANSpace, SeFa, InterFaceGAN, EYE and the proposed method. Each column represents the results of manipulating a specific attribute with different methods. "+" or "-" means the editing direction of desired attribute, which represents "strengthening" or "weakening".

4.3. Real Image Attribute Editing

In this section, we illustrate the performance on editing real images. As the edit operation is performed in the latent space of GANs, a GAN inversion preprocessing is required to find the corresponding latent code of a real face image. Similar to [3, 4], an additional encoding network is trained to learn the inverse mapping of PGGAN and StyleGAN.

Fig. 10 shows the editing performance of four attributes on two real examples for different pre-trained PGGAN models (The results of StyleGAN are included in the supplementary material). It can be seen that the controllability of both GANSpace and SeFa is weak and the changes of attributes are not obvious. InterfaceGAN can get more meaningful results, however, the controllability of desired attribute editing is not satisfactory. For example, when editing "bangs" and "age", the gender of the person is incorrectly changed. When editing the "smile" or "age" attribute, other attributes will also be changed obviously for the method of EYE.

In contrast, the proposed method can edit the specified attribute properly while maintain the others, benefiting from the proposed GMM subspace representation.

Figure 11: The ablation study of coordinate accumulation strategy in PGGAN. (a) The original image, (b) and (c) are respectively the editing results of the proposed method w/o and with coordinate accumulation method. "+" or "-" means the editing direction of desired attribute.

4.4. Ablation Study

In this section, the effectiveness of the coordinate accumulation strategy with orthogonal regularization is demonstrated in an ablation study. Figure 11 shows the results of without or with the coordinate accumulation strategy. We can find that the results equipped with this strategy achieve better performance on keeping the pose and identity property while editing a specified attribute. In addition, quantitative metric on 1000 test images also illustrates the performance of coordinate accumulation strategy, as shown in Table. 2.

Next, an experiment is conducted to show that coordinate accumulation with orthogonal constraint does promote the disentangling performance dramatically. For each unique attribute (smile, gender, bangs, age), the coefficients of the corresponding Gaussian mixture model (GMM) is a 100-dimensional vector, which measures the weight on each basis feature representation. The visualizations of the coefficients estimated by the model equipped with/without orthogonal regularization are shown in Fig. 12 (PGGAN [7]). It can be seen that the model equipped with orthogonal regularization produces sparse dominant representations for both GAN models, which assist with the interpretation of improvements on disentangling performance.

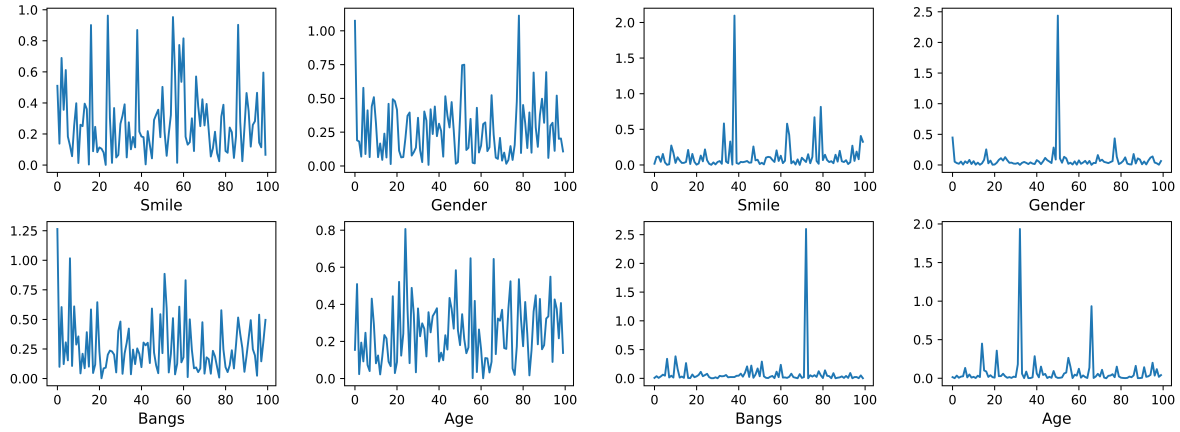

Figure 12: Visualization of GMM coefficients estimated by the model equipped without/with orthogonal regularization for PGGAN model [7].

5. Conclusions

This paper proposes to use GMMs to decouple the semantic space of pre-trained GAN (e.g., PGGAN, StyleGAN) face synthesis models for controllable attribute editing of facial image. A coordinate accumulation strategy with orthogonal regularization is proposed to enhance the independence of distinct attribute subspaces. Qualitative and quantitative experimental results show that the proposed method gains the state-of-the-art performance compared with most related work.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China under Grant (62172198, 61762064, 62041604), Jiangxi Science Fund for Distinguished Young Scholars (20192BCBL23001), Natural Science Foundation of Jiangxi Province (20202BABL212005,20224ACB202008), Science and Technology Research Project of Jiangxi Provincial Education Department (GJJ191152), Innovation Fund Designated for Graduate Students of Jiangxi Province (TC2020-S522),PhD Start-up Fund of Nanchang Hangkong University (EA202107269), Science and technology project of Jiangxi Education Department (DA202207128) and the Opening Project of Nanchang Innovation Institute, Peking University. The authors would like to thank the anonymous referees and the editors for their helpful comments and suggestions.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[2] C. Yang, Y. Shen, B. Zhou, Semantic hierarchy emerges in deep generative representations for scene synthesis, International Journal of Computer Vision 129 (5) (2021) 1451–1466.

[3] Y. Shen, C. Yang, X. Tang, B. Zhou, Interfacegan: Interpreting the disentangled face representation learned by gans, IEEE transactions on pattern analysis and machine intelligence (2020).

[4] E. Härkönen, A. Hertzmann, J. Lehtinen, S. Paris, Ganspace: Discovering interpretable gan controls, Advances in Neural Information Processing Systems 33 (2020) 9841–9850.

[5] Y. Shen, B. Zhou, Closed-form factorization of latent semantics in gans, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1532–1540.

[6] P. Zhuang, O. Koyejo, A. G. Schwing, Enjoy your editing: Controllable gans for image editing via latent space navigation, arXiv preprint arXiv:2102.01187 (2021).

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

[8] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[9] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International conference on machine learning, PMLR, 2019, pp. 7354–7363.

[10] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, arXiv preprint arXiv:1802.05957 (2018).

[11] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, arXiv preprint arXiv:1809.11096 (2018).

[12] W. Shen, R. Liu, Learning residual images for face attribute manipulation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4030–4038.

[13] G. Zhang, M. Kan, S. Shan, X. Chen, Generative adversarial network with spatial attention for face attribute editing, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 417–432.

[14] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, S. Wen, Stgan: A unified selective transfer network for arbitrary image attribute editing, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3673–3682.

[15] Y.-C. Chen, X. Shen, Z. Lin, X. Lu, I. Pao, J. Jia, et al., Semantic component decomposition for face attribute manipulation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9859–9867.

[16] G. Yang, N. Fei, M. Ding, G. Liu, Z. Lu, T. Xiang, L2m-gan: Learning to manipulate latent space semantics for facial attribute editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2951–2960.

[17] Y. Wei, Z. Gan, W. Li, S. Lyu, M.-C. Chang, L. Zhang, J. Gao, P. Zhang, Maggan: High-resolution face attribute editing with mask-guided generative adversarial network, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[18] R. Wu, G. Zhang, S. Lu, T. Chen, Cascade ef-gan: Progressive facial expression editing with local focuses, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5021–5030.

[19] R. Wu, S. Lu, Leed: Label-free expression editing via disentanglement, in: European Conference on Computer Vision, Springer, 2020, pp. 781–798.

[20] J. Lin, R. Zhang, F. Ganz, S. Han, J.-Y. Zhu, Anycost gans for interactive image synthesis and editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14986–14996.

[21] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, A. Torralba, Seeing what a gan cannot generate, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4502–4511.

[22] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, A. Torralba, Semantic photo manipulation with a generative image prior, arXiv preprint arXiv:2005.07727 (2020).

[23] G. Perarnau, J. Van De Weijer, B. Raducanu, J. M. Álvarez, Invertible conditional gans for image editing, arXiv preprint arXiv:1611.06355 (2016).

[24] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, A. A. Efros, Generative visual manipulation on the natural image manifold, in: European conference on computer vision, Springer, 2016, pp. 597–613.

[25] F. Ma, U. Ayaz, S. Karaman, Invertibility of convolutional generative networks from partial measurements, Advances in Neural Information Processing Systems 31 (2018).

[26] J. Gu, Y. Shen, B. Zhou, Image processing using multi-code gan prior, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3012–3021.

[27] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, A. Courville, Adversarially learned inference, arXiv preprint arXiv:1606.00704 (2016).

[28] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, arXiv preprint arXiv:1605.09782 (2016).

[29] J. Zhu, D. Zhao, B. Zhou, B. Zhang, Lia: Latently invertible autoencoder with adversarial learning (2019).

[30] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1798–1828.

[31] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: international conference on machine learning, PMLR, 2019, pp. 4114–4124.

[32] D. Bouchacourt, R. Tomioka, S. Nowozin, Multi-level variational autoencoder: Learning disentangled representations from grouped observations, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[33] E. H. Sanchez, M. Serrurier, M. Ortner, Learning disentangled representations via mutual information estimation, in: European Conference on Computer Vision, Springer, 2020, pp. 205–221.
[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[35] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
[36] Z. Wu, D. Lischinski, E. Shechtman, Stylespace analysis: Disentangled controls for stylegan image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12863–12872.