Supplementary file for the main article A Subspace-based Method for Facial Image Editing

Nan Yang, *Member, IEEE,* MengChu Zhou, *Fellow, IEEE,* Liang Qi, *Member, IEEE,* Xin Luan, Yandong Tang, *Member, IEEE,* Xiaofeng Li, *Fellow, IEEE,* and Zhi Han, *Member, IEEE,*

A. *Predefined Table*

Tov et al. [1] gives a specific definition to clarify latent space, i.e., $\mathbb{Z}$, $\mathbb{W}$, $\mathbb{W}^k$, $\mathbb{W}_*$, and $\mathbb{W}_*^k$. The latent space $\mathbb{Z}$ is a predefined distributed space, such as a Gaussian distribution. $\mathbb{W}$ is obtained from $\mathbb{Z}$ by passing through an affine transformation, which is a style code latent space in $\mathbb{R}^{512}$. The extended space $\mathbb{W}^k$ contains k different inputs with dimension of 512, where k is the number of style inputs of the generator. For example, a generator capable of synthesizing images at a resolution of 1024×1024 operates in the extended $\mathbb{W}^{18}$ space corresponding to the 18 different style inputs. $\mathbb{W}_*$ denotes that individual style codes are not limited to $\mathbb{W}$, i.e., outside the range of StyleGAN's mapping function, while they have the same style codes in all layers. $\mathbb{W}_*^k$ indicates that individual style codes are not limited to $\mathbb{W}$, and it has k different style codes in all layers. $\mathbb{W}_*$ and $\mathbb{W}_*^k$ are in $\mathbb{R}^{512}$ and $\mathbb{R}^{k\times 512}$, respectively. To summarize, $*$ denotes the space is not limited to $\mathbb{W}$. If all the k style codes are equal, the space is still $\mathbb{W}$. If all the k style codes are different, the space is $\mathbb{W}_*^k$. We follow the standard StyleGAN training mode [2], [3] and select $\mathbb{W}$ space to perform modulation. This is because we need to limit the space to $\mathbb{W}$. We do not use space $\mathbb{W}_*$ since there is no style-mixing operation in modulation. The symbols, terms, and definitions are listed in Table. I.

TABLE I
Symbols, terms, and definitions.

α_L	The left lower bound of editing interval
α_R	The right upper bound of editing interval
$\otimes$	Elementwise multiplication
$\uparrow$	The higher, the better
$\downarrow$	The lower, the better
$\leftarrow$	An updating operation
E	An estimated error
M	An iterative mask
N	The minimum sample size
W	A real matrix that contains full space style weights
W_s	A real matrix that contains subspace style weights
SW	A sliding window with size 5
N_P	The number of parameters
G_S	Generator size
N_L	The total style layers
N_F	The number of floating point operations per second
N_{MA}	The number of multiply adds
M_{GC}	Memory usage of GPU and Central Processing
M_S	Memory to store style weight vectors
M_M	Memory to store modulated style weight vectors
M_R	Memory ratio to save a style weight vector
x	An image generated from full space
x'	An image generated from subspace
x_{edi}	An edited image in subspace
$\boldsymbol{z}$	A vector denotes a latent code
$\boldsymbol{d}$	A vector denotes an attribute semantic
α	A scalar coordinate relative to the raw editing point
U	A real matrix in $\mathbb{R}^{m\times m}$
V^T	A real matrix in $\mathbb{R}^{n\times n}$
Σ	A singular value vector that is subject to $n < m$
μ_i^n	The mean of the i'th metric with the number of iterations n
σ_i^n	The variance of the i'th metric with the number of iterations n
σ_{ci}	An index of Σ when AWM converges
σ_{ei}	An end index of Σ
θ_G	The full space parameters of a generator
θ_{G_s}	The subspace parameters of a generator
F_e	An operator to extract well-trained parameters
F_e'	An operator to transfer modulated parameters to style weight vector
F_d	A weight matrix decomposition function, SVD in this work
$\mathbf{C}$	A operator to concatenate different style weight vectors to get a matrix
$\mathbf{C}'$	A operator to split a matrix to get different style weight vectors
$\mathbb{Z}$	A predefined distributed space, e.g., Gaussian distribution
$\mathbb{W}$	A translated space from $\mathbb{Z}$ through an affine transformation
$\mathbb{W}_s$	A subspace of $\mathbb{W}$
$\mathbb{W}^k$	An extended space that contains k different inputs
$\mathbb{W}_*$	A space that is not limited to $\mathbb{W}$
$\mathbb{W}_*^k$	A space that is not limited to $\mathbb{W}$ and contains k different inputs

B. *Analysis Compared to Diffusion-based Models*

We carefully study the unsupervised attribute editing methods based on the conditional diffusion model and StyleGAN [3], including approaches such as StyleCLIP [4], CLIP2latent [5], CLIP2StyleGAN [6], and others. Our analysis focuses on three key aspects: i) how these methods achieve attribute editing and their limitations, ii) how the AWM algorithm works, and iii) what are the strengths of AWM when compared to these diffusion-based models.

i) The above-mentioned models use various techniques to extract editing directions from the latent space of a pre-trained StyleGAN based on textual descriptions using CLIP [7]. It enables text-driven attribute editing. One of these methods, CLIP2latent [5], employs a diffusion model conditioned on CLIP embeddings to sample latent vectors from StyleGAN. At inference time, it generates the CLIP embedding for text input and uses this as the condition for the clip2latent network. By performing denoising diffusion, the clip2latent network generates a StyleGAN latent code which can then be used for image editing. However, diffusion-based methods have the following limitations:

Computational complexity: A diffusion model requires significant computational resources to process large images and may not be feasible for real-time applications. Additionally, the method involves numerous parameters that require tuning, which can be time-consuming.

Limited applicability: Such a model relies on a pre-trained StyleGAN generator and CLIP model for joint language-vision embedding. As StyleCILP [4] points out, these methods may not be possible to manipulate images to a point where they lie outside the domain of the pre-trained generator or remain inside the domain but in regions less well-covered by the

generator. Similarly, text prompts that map into areas of CLIP [7] space that are not well-populated by images may not yield visual manipulations that faithfully reflect the semantics of the prompt. It may also be challenging to achieve drastic manipulations in visually diverse datasets.

Lack of fine-grained control: Such a model operates by gradually modifying the image over time, resulting in a smooth and natural-looking transition between the original and modified versions. However, this also means that users have limited control over the exact changes made to the image, making it difficult to achieve specific modifications.

ii) Our proposed Adaptive Weight Modulation (AWM) is to find a subspace with an orthogonal decomposition that allows the independent modification of individual attributes in a generative model without affecting other attributes. The method involves several modules and operations to achieve this goal, including supervised prior knowledge discovery and a convergent criterion establishment.

One of the critical features of AWM is the use of a sliding window and interactive mask, which control the number of diagonal values and corresponding parameters that participate in at each iteration. This approach ensures that only the parameters corresponding to the selected diagonal values are used for face generation, enabling independent attribute modification.

iii) The advantage of AWM over a diffusion model lies in its ability to independently modify individual attributes in a generative model without affecting others. It is achieved by finding a subspace with an orthogonal decomposition that yields the maximum effective parameters in full space to generate high-quality faces. As a result, the subspace possesses three key advantages: 1) it contains the maximum number of effective parameters to generate delicate faces while sharply reducing subspace parameters (**Low complexity**); 2) it can generate high-fidelity faces just as if the full space is used (**High applicability**); and 3) it offers consistency at the pixel, perceptual, and visual levels for a specific person (**Fine-grained control**), which is hard to realize with a diffusion model.

AWM also employs a sliding window and interactive mask to control the number of diagonal values and corresponding parameters used for face generation, ensuring independent attribute modification. Furthermore, AWM does not require a large amount of training data to learn the underlying distribution of images or train an extra encoder/transformer to realize face editing, which is a limitation of diffusion models.

C. *Motivation for AWM*

Why should we obtain an orthometric subspace? 1) An orthometric subspace enables manipulation of one attribute while leaving the others unchanged, resulting in more natural-looking and realistic image editing results. 2) It can improve model interpretability by identifying specific features responsible for particular aspects of the data. Therefore, an orthometric subspace enables disentangled representation learning, which improves the ability to manipulate and interpret data in various applications.

Aiming to overcome the limitations: AWM addresses these limitations by finding a subspace with an orthogonal decomposition in StyleGAN latent space while not training an extra encoder or transformer mapping network. As a result, the subspace possesses three key advantages: 1) it contains the maximum number of effective parameters to generate delicate faces while sharply reducing subspace parameters (Low complexity); 2) it can generate high-fidelity faces just as if the full space is used (High applicability); and 3) it offers consistency at the pixel, perceptual, and visual levels for a specific person (Fine-grained control), which is not possible with the existing diffusion models.

Hence, we have developed AWM, which allows for independent attribute modification while maintaining high quality and consistency in face generation and editing.

TABLE II
The cosine distance of five different attribute semantics.

Cosine	Pose	Glasses	Age	Smile	Surprise
Pose	1.0000	-0.0025	0.0177	0.0142	0.0054
Glasses		1.0000	0.0817	0.0393	-0.0448
Age			1.0000	0.0682	-0.0703
Smile				1.0000	-0.1034
Surprise					1.0000

TABLE III
The Pearson correlation coefficient of five different attribute semantics.

Pearson	Pose	Glasses	Age	Smile	Surprise
Pose	1.0000				
Glasses	-0.0032	1.0000			
Age	-0.0344	0.1626	1.0000		
Smile	0.0307	0.0771	0.1355	1.0000	
Surprise	0.0119	-0.0906	-0.0142	-0.2081	1.0000

D. *Motivation for MST*

Why should we find an editing interval? The process of attribute editing can be conceptualized as a motion along a curve in the latent space, which can be referred to as the "editing curve." The editing curve can have steepest-change points where unexpected performances could occur, such as jumping from one attribute to another or changing facial identity.

However, the existing methods do not provide a robust and principled way to prevent unexpected performances when editing at the steepest-change points, which may result in totally undesired effects on facial identity and image quality.

Hence, we propose Maximum Slope Truncation (MST) to address the challenges posed by the editing process in the latent space. It offers a rigorous theoretical derivation and computational formula that provides an editing interval and balances the trade-off between image editing and facial identity preservation. An editing interval can preserve facial identity while allowing for effective image editing. As a result,

it enables more precise and controlled editing in the latent space.

TABLE IV
The cosine distance of edited codes on $\mathbb{Z}$.

Cosine	Pose	Glasses	Age	Smile	Surprise
Pose	1.0000	-0.0796	0.0576	0.0658	0.0653
Glasses		1.0000	0.1227	0.0727	-0.0003
Age			1.0000	0.0959	-0.0285
Smile				1.0000	-0.0705
Surprise					1.0000

TABLE V
The Pearson correlation coefficient of edited codes on $\mathbb{Z}$.

Pearson	Pose	Glasses	Age	Smile	Surprise
Pose	1.0000				
Glasses	-0.1609	1.0000			
Age	0.1160	0.2448	1.0000		
Smile	0.1336	0.1440	0.1910	1.0000	
Surprise	0.1315	-0.0002	-0.0576	-0.1421	1.0000

E. *Key Difference Compared to Existing Design*

*1) **Difference Compared to Diffusion-based Models**:* Our proposed AWM and MST approaches differ from existing designs, such as a diffusion model that is the most relevant to our work. We summarize these differences in terms of their training strategy, design idea, and editing mode.

1) Regarding a training strategy, AWM employs a supervised learning strategy to discover prior knowledge from a pre-trained StyleGAN-based model, which is used to construct a subspace for attribute editing and can save significant amounts of time and computational resources. In contrast, a diffusion model mainly relies on unsupervised learning strategies. It maps the latent spaces of pre-trained CLIP [7] and StyleGAN [3] models to achieve text-driven attribute editing.

2) Regarding a design idea, AWM and MST utilize different mathematical concepts and techniques from a diffusion model. AWM uses an orthogonal decomposition to find a subspace where attribute modification can be done independently. Furthermore, MST uses a slope truncation method and derives a rigorous editing interval to control the variation of the output images. These novel design ideas allow for more precise and efficient editing. In contrast, a diffusion model conditioned on CLIP embeddings is employed to sample latent vectors from StyleGAN. At inference time, it generates the CLIP embedding for text input and uses this as the condition for the clip2latent network.

3) Regarding an editing mode, AWM and MST offer different editing modes in comparison with a diffusion model's. AWM can edit individual attributes independently without affecting others, allowing precise control over image features. MST truncates the output images' slopes in latent space, resulting in a rigorous editing interval and making a trade-off between face editing and facia identity. This editing mode lets users modify the image without drastically changing its features. However, diffusion-based methods rely on a pre-trained StyleGAN generator and CLIP model for joint language-vision embedding. Thus, it is impossible to manipulate images to a point where they lie outside the domain of the pre-trained generator or remain inside the domain but in regions less well-covered by the generator. Similarly, text prompts that map into areas of CLIP space that are not well-populated by images may not yield visual manipulations that faithfully reflect the semantics of the prompt. It may also be challenging to achieve drastic manipulations in visually diverse datasets.

Overall, the AWM and MST approaches offer a more precise and controlled method of attribute editing in the latent space, which is different from existing designs and can thus improve the applicability and efficiency of image editing tasks.

*2) **Difference Compared to InterfaceGAN and SeFa**:* SeFa aims to discover meaningful semantics in an unsupervised manner, and InterfaceGAN [8] tries to disentangle the attributes via sub-space projection. However, they do not consider the problem of discovering a subspace. The goal of AWM is to find a subspace with an orthogonal decomposition, such that can edit an attribute independently but not affect other attributes. Inspired by the reviewer's comment, we have also investigated other unsupervised PCA-based methods [9], [10]. Unfortunately, none of them deals with the same problem as AWM does.

Furthermore, AWM involves several modules and operations to realize the goal, such as discovering suitable supervised prior knowledge and establishing a convergent criterion. The sliding window and interactive mask in AWM are two switches that control the number of diagonal values and corresponding parameters participating in at each iteration. Only the parameters corresponding to the selected diagonal values are used for face generation. SVD [11], [12] is only an adequate tool we use in AWM, which decomposes the weights of face generators and yields singular values. It has the merit of low computational efficiency in comparison with PCA. Our method needs to modulate many parameters. Therefore, computational efficiency is a significant factor that must be considered. That is the reason we adopt SVD in our method rather than PCA.

To summarize, our work aims to discover an orthometric subspace that edits one attribute without affecting others and makes a good trade-off between image editing and facial identity preservation.

F. *Correlation analysis and obtain editing interval*

1) Disentanglement analysis on attributes: We analyze five key facial attributes, i.e., pose, glasses, age, smile, and surprise, which contains subjectively independent (pose vs. glasses) and highly relevant attributes (smile vs. surprise). We follow the local linear hypothesis in image editing [8], [13], [14] and introduce the Pearson correlation coefficient [15] as another metric. The results are shown in Tables II and III. We reveal four major discoveries:

1) Every diagonal element is equal to 1.0, which denotes that all the attributes are autocorrelated;

2) Pose is almost unrelated to other attributes. Smile has a high correlation with surprise on two metrics. This result does make sense since 1) pose seems to be unrelated natively to other attributes, and 2) smile and surprise are an important part of facial emotion. Beyond our expectation, age and glasses, age and smile, age and surprise have a high correlation;

3) Age and smile have a positive correlation, while age and surprise have the negative one. The reason seems to be that smiling and surprise are not simple pixel-to-pixel changes but require muscle movements. Smile and surprise may affect the change in age since they are two global attributes. We find that the facial skin become creased and loosened when editing smiles and surprises, respectively. The above observation explains why age and smile have a positive correlation, while age and surprise have the negative one;

4) Two attributes with a high correlation may cause entanglement, i.e., one attribute transfers to another, such as age, which may change when editing glasses;

2) Disentanglement analysis on space $\mathbb{Z}$ and $\mathbb{W}$: Firstly, we determine the minimum sample size N. For a given $z_{\boldsymbol{\alpha}/2}$ and standard deviation σ, an estimated error is defined as:

$$E = z_{\boldsymbol{\alpha}/2}\frac{\sigma}{\sqrt{N}} \tag{1}$$

We can derive the sample size N as:

$$N = \frac{\left(z_{\boldsymbol{\alpha}/2}\right)^2\sigma^2}{E^2} \tag{2}$$

where $z_{\boldsymbol{\alpha}/2}$ denotes a quantile of normal distribution at $\boldsymbol{\alpha}/2$, and bold $\boldsymbol{\alpha}$ indicates a confidence coefficient. It is sufficient to compute N using (2) since sampling from a standard normal distribution (σ is equal to 1.0) satisfies a condition of infinite population sampling. $\boldsymbol{\alpha}$ is equal to 0.05, which means that we have a confidence coefficient of 95%, and its corresponding $z_{\boldsymbol{\alpha}/2}$ is equal to 1.96 [16], [17]. We set E to 0.05 based on the following considerations: 1) we have an acceptable memory to store these samples. Because the images have a resolution of 1024×1024, we need to have enough memory to save them; and 2) due to the capacity of a face generator, there are approximately 5 failure cases per 100 images. We calculate $N = 1537$ by putting the above parameter values into (2). We use 2000 samples that the real estimated error is less than 5%, and the experimental results become more authentic than 1537.

The quantitative correlation analysis results in $\mathbb{Z}$ space are shown in Tables IV and V. We conclude two conclusions:

1) latent code $\boldsymbol{z}$ inherits the correlation among attributes, which indicates that we are prone to entanglement when editing in $\mathbb{Z}$ space. This change is detrimental to controllable image editing;

2) The correlation between age vs. glasses and age vs. smile is large, while that between age vs. surprise and smile vs. surprise is small. This indicates that the generator is biased when generating one face, i.e., it is inclined to generate aging men/women with glasses or smiles, and to produce a "surprise" face that looks slightly younger than the actual one with no "smile" co-existing;

Fig. 1. An interactive demo of our developed web app.

We further provide the corresponding experimental results in $\mathbb{W}$ space, as shown in Tables VI and VII. We find the following results:

1) The correlation is large and stable after adding an attribute semantic to a latent code in $\mathbb{W}$ space. A generator must have the capacity to maintain facial identity. For a specific person, an edited code should be similar despite different attribute semantics, because of the code should contain enough facial identity information. This means that the identity of one person cannot change with the change in attribute editing;

2) Performing image editing in $\mathbb{W}$ space may not cause entanglement since the latent code does not inherit the correlation among attributes. These attribute semantics are seen as a style code to participate in image generation and editing. It is easier to control the editing process, whereas $\mathbb{W}$ space may "over-disentangle". Therefore, finding a truncation point in $\mathbb{W}$ space, where the slope changes the most, is crucial.

TABLE VI
The cosine distance of edited codes on $\mathbb{W}$.

Cosine	Pose	Glasses	Age	Smile	Surprise
Pose	1.0000	0.4986	0.4982	0.4972	0.4977
Glasses		1.0000	0.4980	0.4968	0.4968
Age			1.0000	0.4967	0.4963
Smile				1.0000	0.4968
Surprise					1.0000

G. Ablation Studies

Unsupervised semantics has an inherent limitation in that the explicit meaning of one semantic is unknown. We propose a guideline, which is a solution for understanding the meaning of a semantic. We choose 5 frequently-used semantics to carry out analysis, i.e., pose, glasses, age, smile, and gender. Correspondingly, 5 eigenvectors with the largest singular values are selected from V^T. We use two metrics, i.e., Euclidean distance [18] and cosine distance [19]. The former denotes a

TABLE VII
The Pearson correlation coefficient of edited codes on $\mathbb{W}$.

Pearson	Pose	Glasses	Age	Smile	Surprise
Pose	1.0000				
Glasses	0.9971	1.0000			
Age	0.9964	0.9960	1.0000		
Smile	0.9942	0.9934	0.9933	1.0000	
Surprise	0.9953	0.9935	0.9925	0.9832	1.0000

relative distance of two attribute semantics in latent space. A smaller Euclidean distance signifies a greater similarity, which means two semantics have a similar expression. The latter measures the angle of two semantics in latent space. A smaller cosine distance signifies a greater similarity. If the two metrics have both minima, we consider the two semantics to have a similar expressive ability. From Table VIII, we can obtain two conclusions:

1) the unsupervised semantic has an expressive ability similar to that of most supervised semantics. They are consistent, e.g., age vs. V_1^T, glasses vs. V_2^T, and pose vs. V_3^T. The unsupervised semantic is a disentangled semantic since it suffers from the restriction of $\mathbf{n}^T\mathbf{n} = 1$;

2) Pose and gender have a small value with V_3^T. This result may contain over-disentangled semantics, i.e., the two semantics are encoded into a new style. 3) We visualize the editing image based on some unsupervised semantics. The results show that these semantics can control their corresponding attribute changes well and the novel guideline is first proposed.

TABLE VIII
The ablation studies for evaluating the meaning of unsupervised semantic. The experiments are conducted on five attribute semantics on two metrics. SS and US denote supervised semantic and unsupervised semantic, respectively.

US \ SS		Pose	Glasses	Age	Smile	Gender
V_0^T	Cosine	0.4899	0.4954	0.5148	0.4917	0.5017
	Euclidean	4.5397	5.8885	5.3287	5.8885	5.0407
V_1^T	Cosine	0.4659	0.4958	**0.4618**	0.5151	0.5039
	Euclidean	4.4625	5.8907	**5.0603**	6.1125	5.0506
V_2^T	Cosine	0.5229	**0.4798**	0.5089	0.4948	0.5112
	Euclidean	4.6437	**5.7951**	5.2995	5.9910	5.0836
V_3^T	Cosine	**0.4593**	0.5171	0.5093	0.4957	**0.4801**
	Euclidean	**4.4412**	6.0160	5.3014	5.9960	**4.9410**
V_4^T	Cosine	0.5109	0.4956	0.5020	**0.4626**	0.5145
	Euclidean	4.6060	5.9717	5.2651	**5.7923**	5.0984

We study the effect of sample size N on attribute analysis. We set the error E as 2% and 3%, which are lower than 5%. The corresponding sample sizes are 4300 and 9600. The quantitative results are shown in Table IX. We can obtain a consistent result based on different larger sample sizes. The values with a small fluctuation is related to random sampling. Hence, 1) this is a correct result based on an error of 5%; 2) $\mathbb{W}$ space is a limited one, which has a strong ability to preserve one's identity while suffering from entanglement easily; and 3) it is difficult for a face generator to have only 2 or 3 bad samples out of 100.

H. Web application

We develop a web app to help users perform facial image editing freely. The app supports unsupervised semantic and supervised semantics (more than 40 attribute semantics). It can also be applied on both official and third-party generators. Users can independently determine the editing effect based on our editing interval, as shown in Fig. 1. Meanwhile, we provide 3 dynamic metrics that change during the editing process. These metrics include identity score [20], cosine distance [19], and Euclidean distance [18]. We fully consider both editing performance and dynamic score for facial editing, which is not considered in existing methods. Our method can realize adaptive editing in comparison with state-of-the-art methods.

I. More Experimental Results

We newly conduct experiments to evaluate the performance of face editing with six state-of-art methods, i.e., TransEditor [21], StyleCLIP [4], CLIP2lantent [6], InterfaceGAN [22], SeFa [13], DNI [23]. The qualitative results are shown in Fig. 2.

The results show that our proposed method can do facial editing well and have the least influence on other attributes when editing a target attribute, e.g., smile and age. We further investigate the inversion performance, which fully reflects the ability of image generation and real image editing. The quantitative results are shown in Table. X.

The quantitative comparison results show that our proposed method can infer the optimal inverted code and generate high-quality faces in the subspace. It further indicates that our proposed method can-not only edit one attribute for a real person well but also ensure image quality.

We also utilize a re-scoring metric to evaluate the attribute entanglement of our editing results on other facial attributes. We choose three attributes: pose, gender, age, and smile. The comparative results of our method with TransEditor [21], StyleCLIP [4], InterfaceGAN [22], DNI [23] are presented in Table XI.

Our experiments show that our method exhibits minimal influence on other facial attributes when editing each specific attribute, indicating that our approach is highly disentangled. For example, this finding suggests that our method effectively disentangles age-related attributes from other facial features, allowing for independent manipulation of age-related attributes while preserving other facial features. Therefore, our approach is well-suited for face editing tasks and outperforms state-of-the-art methods.

We perform experiments on five different face generators. The experimental results are shown in the following Table XII. The results show that config-1 and config-3 tend to get larger singular values while config-2 tends to get smaller singular values in comparison to convergent criterion. The larger singular values can not yield an optimal subspace, while the smaller singular values can not generate fine faces as done

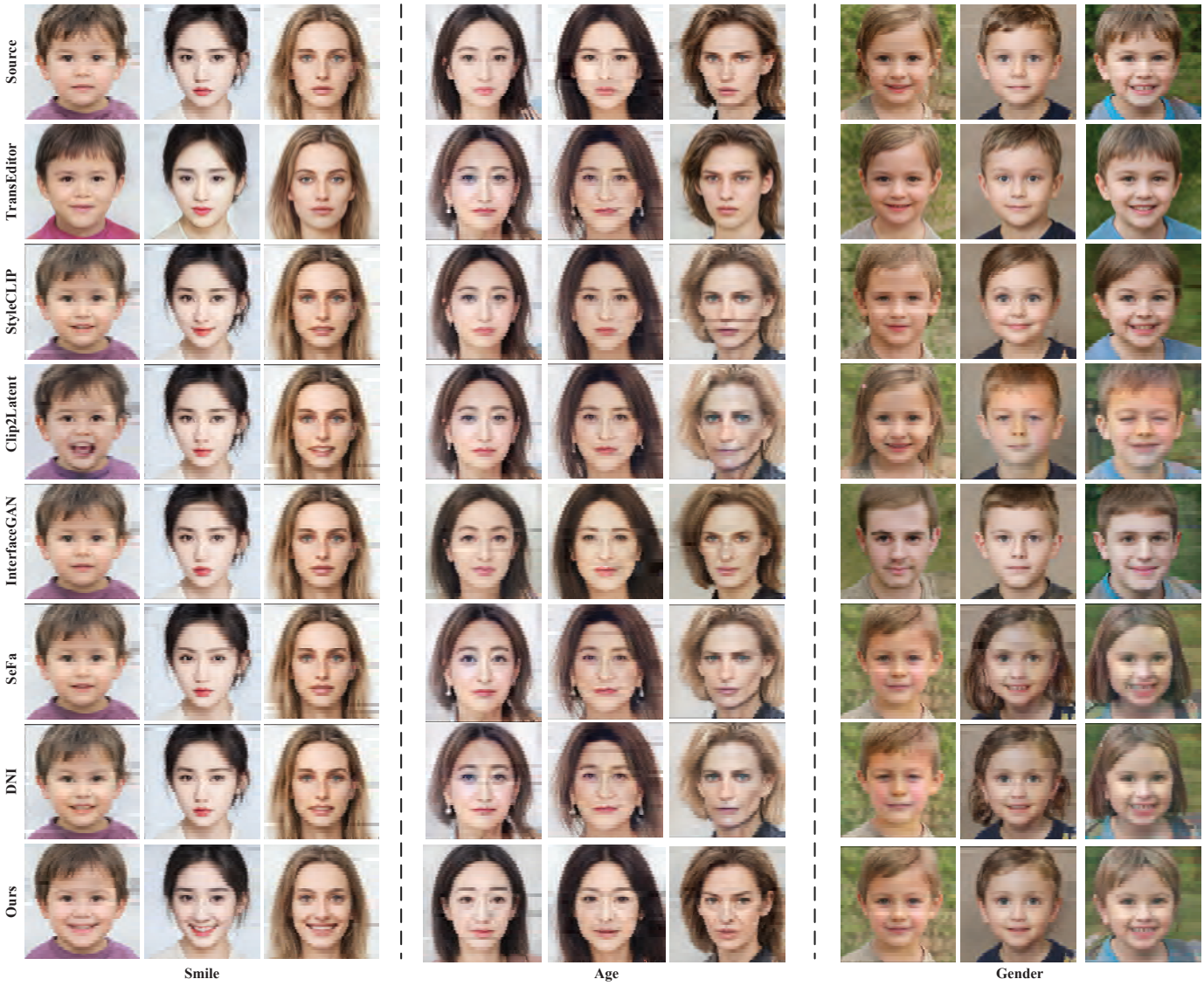

Fig. 2. Qualitative comparison results with state-of arts methods.

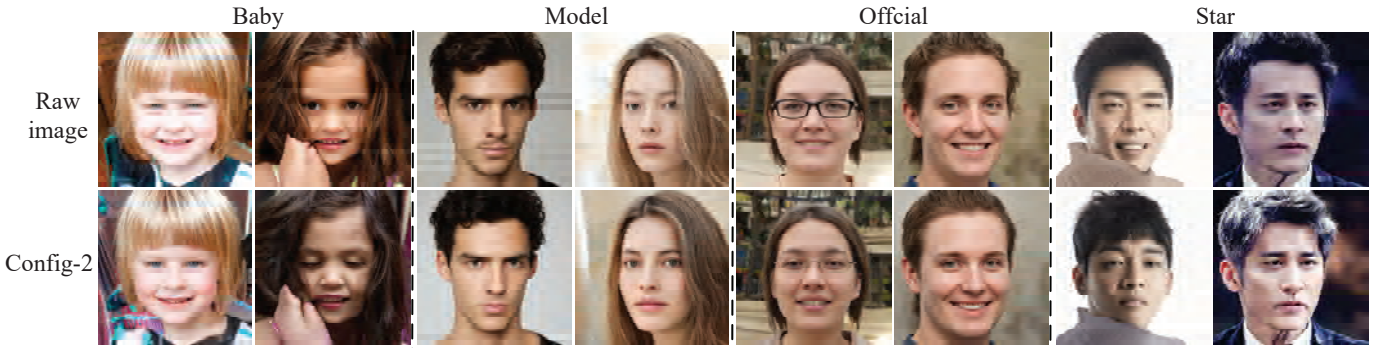

Fig. 3. The comparison results between the raw image and generated image. The subspace is derived using Config-2.

as in full space, e.g., losing facial identity, and producing image artifacts, as shown in Fig. 3.

Fig. 4 shows that all the "bad samples" only appear at or outside the bounds, i.e., left to outside the left bound, right to outside the right bound. The editing state changes most dramatically occur at points α_L and α_R. The gender editing results in Fig. 5 show that our proposed method can well edit the attribute of gender. Furthermore, according to the comparative results, we can find that the attribute of smile and beard are not changed when editing gender. The results

TABLE IX
The ablation studies for evaluating the effect of sample size on attribute analysis. The experimental results are obtained on five attribute semantics on two metrics.

Configuration		Pose		Glasses		Age		Smile		Surprise	
		N_1=4300	N_2=9600	N_1=4300	N_2=9600	N_1=4300	N_2=9600	N_1=4300	N_2=9600	N_1=4300	N_2=9600
Pose	Cosine	1.0000	1.0000	0.4950	0.4950	0.4938	0.4938	0.4979	0.4979	0.4919	0.4919
	Pearson	1.0000	1.0000	0.9889	0.9889	0.9862	0.9862	0.9953	0.9953	0.9820	0.9820
Glasses	Cosine	0.4950	0.4950	1.0000	1.0000	0.4931	0.4931	0.4954	0.4954	0.4889	0.4889
	Pearson	0.9889	0.9889	1.0000	1.0000	0.9847	0.9848	0.9898	0.9898	0.9752	0.9752
Age	Cosine	0.4938	0.4938	0.4931	0.4931	1.0000	1.0000	0.4946	0.4946	0.4871	0.4871
	Pearson	0.9862	0.9862	0.9847	0.9848	1.0000	1.0000	0.9880	0.9880	0.9713	0.9713
Smile	Cosine	0.4979	0.4979	0.4954	0.4954	0.4946	0.4946	1.0000	1.0000	0.4929	0.4929
	Pearson	0.9953	0.9953	0.9898	0.9898	0.9880	0.9880	1.0000	1.0000	0.9843	0.9843
Surprise	Cosine	0.4919	0.4919	0.4889	0.4889	0.4871	0.4871	0.4929	0.4929	1.0000	1.0000
	Pearson	0.9820	0.9820	0.9752	0.9752	0.9713	0.9713	0.9843	0.9843	1.0000	1.0000

TABLE X
Quantitative comparison results with state-of arts methods.

Method	SSIM	PSNR	MS_SSIM	VIF	FSIM	GMSD	LPIPS	DISTS
TransEditor	0.9678	21.9485	0.9608	0.8781	0.7826	0.7974	0.8883	0.8621
StyleCLIP	0.9736	21.7049	0.9759	0.8821	0.7790	0.7906	0.8828	0.8593
InterfaceGAN	0.9703	21.7575	0.9768	0.8850	0.8606	0.8507	0.9365	0.8995
DNI	0.9878	22.8140	0.9797	0.8821	0.8634	0.8523	0.9381	0.8935
Ours	**0.9929**	**22.8665**	**0.9802**	**0.8872**	**0.8699**	**0.8554**	**0.9436**	**0.9016**

TABLE XI
Quantitative editing comparison between TransEditor[1], StyleCLIP[2], InterfaceGAN[4], and DNI[6]. The row-column entry represents the degree of change of the column attribute while editing the row attribute. The comparison shows our method has the least effect on other attributes during editing.

	Pose					Gender					Age					Smile				
Method	InterfaceGAN	DNI	StyleCLIP	TransEditor	Ours	InterfaceGAN	DNI	StyleCLIP	TransEditor	Ours	InterfaceGAN	DNI	StyleCLIP	TransEditor	Ours	InterfaceGAN	DNI	StyleCLIP	TransEditor	Ours
Pose	-	-	-	-	-	0.2308	0.2450	**0.2154**	0.2407	0.2277	0.2538	0.2436	0.2357	0.2154	**0.2119**	0.1231	0.1505	0.1999	**0.1078**	0.1203
Gender	0.0556	0.0321	-	0.0398	**0.0154**	-	-	-	-	-	0.0404	0.0514	0.0507	0.0733	**0.0289**	0.0249	0.0568	0.0384	0.1078	**0.0234**
Age	0.1748	0.1435	-	0.1326	**0.0558**	0.1840	0.1032	0.1042	0.1897	**0.0913**	-	-	-	-	-	0.0909	0.1025	0.1445	0.1642	**0.0515**
Smile	**0.1538**	**0.0825**	**-**	**0.0313**	**0.0727**	**0.2462**	**0.1854**	**0.1846**	0.2047	**0.1818**	0.1077	0.1037	0.1028	0.1077	**0.0909**	-	-	-	-	-

TABLE XII
The studies for verifying convergent criterion.

StyleGAN2	Offcial	Baby	Celebrity	Star	Model
Config-1	σ = 190	σ = 95	σ = 90	σ = 100	σ = 110
Config-2	σ = 70	σ = 60	σ = 50	σ = 55	σ = 50
Config-3	σ = 175	σ = 95	σ = 125	σ = 80	σ = 75
Convergent Criterion	σ = 165	σ = 90	σ = 115	σ = 65	σ = 95

show that our method can edit gender well and have the least influence on other attributes when editing a target attribute. The qualitative comparison results for age editing are shown in Fig. 6.

REFERENCES

[1] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Trans. Graph.*, vol. 40, pp. 1–14, 2021. [Online]. Available: http://arxiv.org/abs/2102.02766

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, 2018.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 4396–4405, 2019.

[4] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," 2021.

[5] J. N. M. Pinkney and C. Li, "clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip," *arXiv preprint arXiv:2210.02347*, 2022.

[6] R. Abdal, P. Zhu, J. Femiani, N. Mitra, and P. Wonka, "Clip2stylegan: Unsupervised extraction of stylegan edit directions," 2022, pp. 1–9.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," 2021, pp. 8748–8763.

[8] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," 2020, pp. 9243–9252.

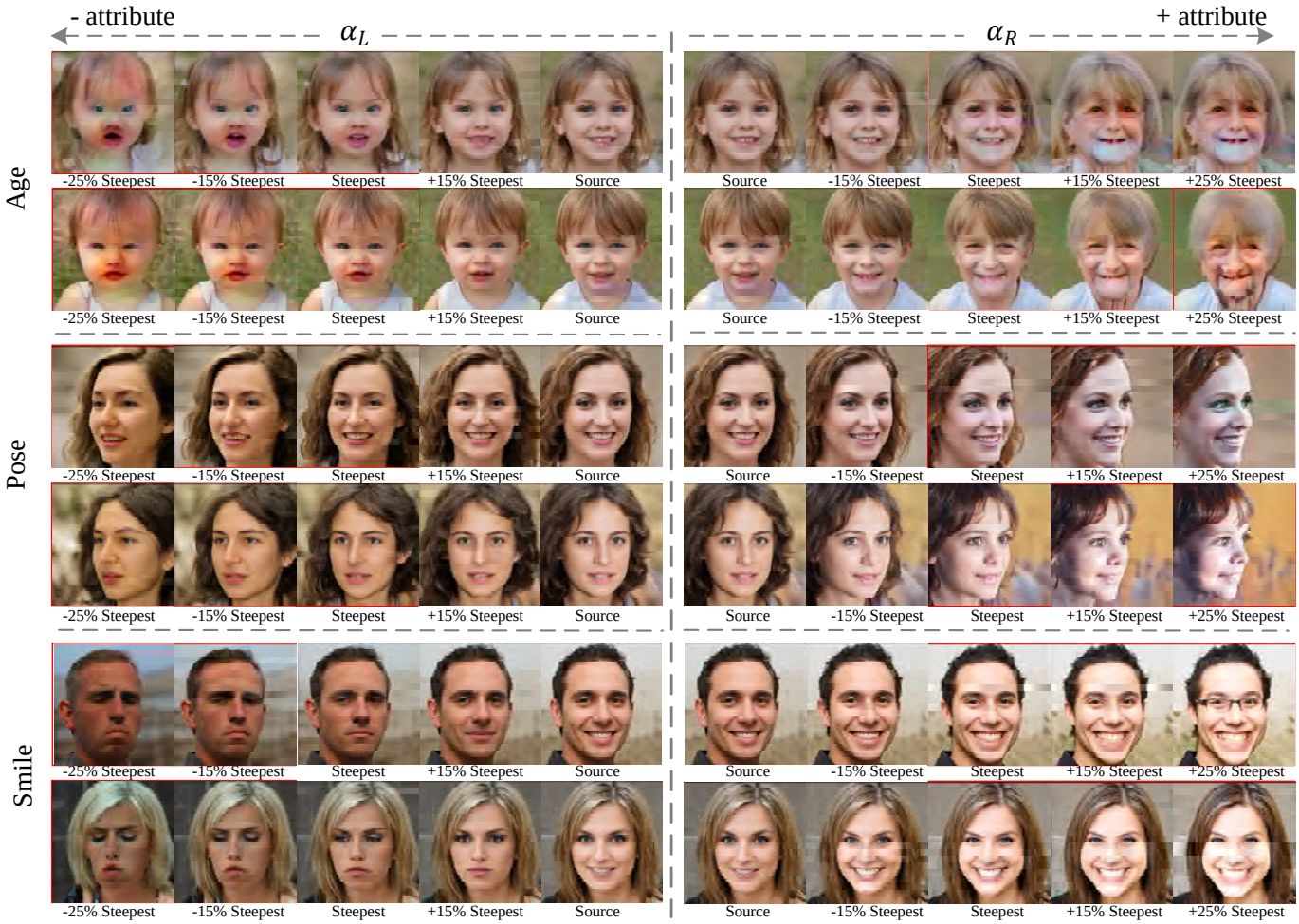

Fig. 4. The state changes in the steepest gradients of $f(X)$, where the steepest-change points occur at points α_L and α_R. Bad editing cases are labeled in the red box.

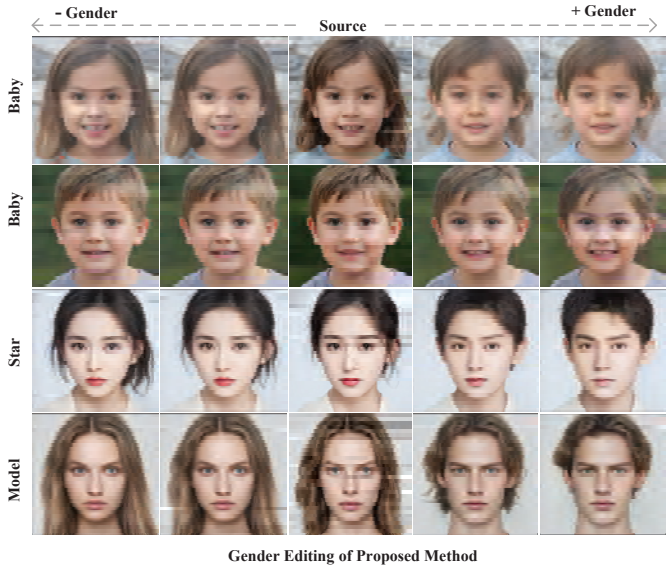

Fig. 5. The comparison of gender editing between the proposed method and peers.

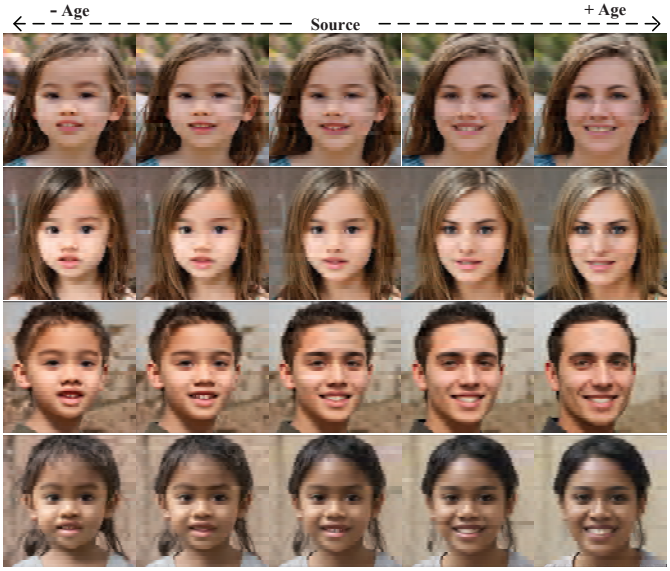

Fig. 6. The editing results for age editing.

[9] A. Daffertshofer, C. J. Lamoth, O. G. Meijer, and P. J. Beek, "Pca in studying coordination and variability: a tutorial," *Clinical biomechanics*, vol. 19, no. 4, pp. 415–428, 2004.

[10] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.
[11] A. Hoecker and V. Kartvelishvili, "Svd approach to data unfolding," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 372, no. 3, pp. 469–481, 1996.
[12] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
[13] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," 2021, pp. 1532–1540. [Online]. Available: http://arxiv.org/abs/2007.06600
[14] J. Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," vol. 9909 LNCS, 2016, pp. 597–613.
[15] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," pp. 1–4, 2009.
[16] D. M. Wolcott, A. Duarte, and F. W. Weckerly, "Statistical inference," pp. 199–205, 2018.
[17] T. Augustin, G. Walter, and F. P. A. Coolen, *Statistical inference*. Cengage Learning, 2014.
[18] L. Wang, Y. Zhang, and J. Feng, "On the euclidean distance of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1334–1339, 2005.
[19] D. Zhang and G. Lu, "Evaluation of similarity measurement for image retrieval," vol. 2, 2003, pp. 928–931.
[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," vol. 2019-June, 2019, pp. 4685–4694.
[21] Y. Xu, Y. Yin, L. Jiang, Q. Wu, C. Zheng, C. C. Loy, B. Dai, and W. Wu, "Transeditor: transformer-based dual-space gan for highly controllable facial editing," 2022, pp. 7683–7692.
[22] Y. Shen, C. Yang, X. Tang, and B. Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 2004–2018, 2022.
[23] N. Yang, Z. Zheng, M. Zhou, X. Guo, L. Qi, and T. Wang, "A domain-guided noise-optimization-based inversion method for facial image manipulation," *IEEE Transactions on Image Processing*, vol. 30, pp. 6198–6211, 2021.