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

Nan Yang, *Member, IEEE,* MengChu Zhou, *Fellow, IEEE,* Zhi Han, *Member, IEEE,* Liang Qi, *Member, IEEE,* Yandong Tang, *Member, IEEE,* and Tianran Wang

A. Predefined Table

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. AWM realization

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 $\mathbb{R}^{k\times512}$, 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.

$$g(\mu,\sigma)=\begin{cases}\mu_i^n+\sigma_i^n>\mu_i^{n-1}, & \text{if } i=1,2\\ \mu_i^n+\sigma_i^n<\mu_i^{n-1}, & \text{if } i=3,4,5,6,7\end{cases} \tag{1}$$

We analyze the image generated from full-space and sub-space using diversified image evaluation metrics [4]. Mean Square Error (MSE) [5], Root Mean Square Error (RMSE) [6], and Peak Signal to Noise Ratio (PSNR) [7], all of which can be used to evaluate the differences between the full-space and sub-space images from a global perspective. They treat all pixels in an image equally while not reflecting the full range of human visual characteristics. The Universal Quality Index (UQI) [8] measures image quality by image relevance, image brightness, and contrast differences. The higher the value, the better the image quality. However, it cannot correlate all subjective evaluations, which leads to its instability issue. As a full-reference image quality evaluation metric, Structural Similarity Measure (SSIM) [9] measures the similarity of images in brightness, contrast, and structure. It outperforms PSNR in terms of image de-noising and similarity evaluation. In practice, it calculates different blocks and finally averages the results to ensure stability and satisfy a human visual system (HVS) [10] for local information. Multi-Scale Structural Similarity Index (MS-SSIM) [11] calculates the structural similarity index by combining multiple images at different scales. It can be more robust than SSIM when the observation condition changes. Visual Information Fidelity (VIF) [12] is an image quality evaluation metric based on statistical image models, image distortion models, and human visual system models. Compared with PSNR and SSIM, it

has higher consistency in subjective vision. The higher the VIF value, the better the image quality.

We use (1) as AWM's convergent criterion, where i is an index of a metric in Table II, e.g., μ_1 has same meaning as that of $\mu_{\mathbf{MSE}}$. μ_i^n and σ_i^n denote the mean and variance of the i'th metric with the number of iterations n. $\downarrow$ indicates that the smaller the value is, the higher the image quality, while $\uparrow$ means the opposite in Table II. AWM converge if it meets (1). By fully considering the mean and variance of multiple metrics, we eliminate the effects of oscillation with different batch input images and the proposed algorithm becomes more robust than those using an individual metric only.

TABLE II
The convergent criterion is based on the following metrics. $\uparrow$ denotes that the higher, the better. $\downarrow$ denotes that the lower, the better.

Configuration	**MSE**	**RMSE**	**PSNR**	**UQI**	**SSIM**	**MS-SSIM**	**VIF**
index	1	2	3	4	5	6	7
trend	$\downarrow$	$\downarrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$

C. 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 [13], [14], [15] and introduce the Pearson correlation coefficient [16] as another metric. The results are shown in Tables III and IV. 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 [17], [18] is defined as:

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

TABLE III
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 IV
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

We can derive the sample size N as:

$$N = \frac{\left(z_{\boldsymbol{\alpha}/2}\right)^2 \sigma^2}{E^2} \tag{3}$$

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 (3) 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 [19]. 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 (3). We use 2000 samples that the real estimated error is less than 5%, and the experimental results become more authentic than 1537.

TABLE V
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

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

TABLE VI
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

TABLE VIII
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

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;

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

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, no. 4, pp. 1–14, 2021. [Online]. Available: http://arxiv.org/abs/2102.02766

[2] 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.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8107–8116, 2020.

[4] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image Quality Assessment: Unifying Structure and Texture Similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[5] D. Guo, S. Shamai, and S. Verdú, "Mutual information and MMSE in Gaussian channels," *IEEE Int. Symp. Inf. Theory - Proc.*, vol. 51, no. 4, p. 347, 2004.

[6] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[7] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *2010 20th Int. Conf. Pattern Recognit.* IEEE, 2010, pp. 2366–2369.

[8] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, 2002.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[10] K. A. Panetta, E. J. Wharton, and S. S. Agaian, "Human visual system-based image enhancement and logarithmic contrast measure," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 38, no. 1, pp. 174–188, 2008.

[11] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *2017 IEEE Int. Conf. Image Process.* IEEE, 2017, pp. 4277–4281.

[12] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[13] 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, no. 4, pp. 2004–2018, 2022.

[14] Y. Shen and B. Zhou, "Closed-Form Factorization of Latent Semantics in GaNs," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1532–1540. [Online]. Available: http://arxiv.org/abs/2007.06600

[15] J. Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9909 LNCS. Springer, 2016, pp. 597–613.

[16] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Springer Top. Signal Process.* Springer, 2009, vol. 2, pp. 1–4.

[17] C. Fohlin, "Financial systems," *Handb. Cliometrics*, pp. 393–430, 2016.

[18] T. Augustin, G. Walter, and F. P. Coolen, *Statistical inference*. Cengage Learning, 2014.

[19] D. M. Wolcott, A. Duarte, and F. W. Weckerly, "Statistical inference," in *Encycl. Ecol.*, 2018, pp. 199–205.