Attribute Group Editing for Reliable Few-shot Image Generation Supplementary Material

Guanqi Ding, Xinzhe Han, Shuhui Wang, Shuzhe Wu, Xin Jin, Dandan Tu, Qingming Huang

This supplementary document is organized as follows:

- Appendix A provides the ablation study for the number of test categories in VGGFaces [1] (Section 4.4).
- Appendix B provides the demonstration of the assumption of Gaussian distribution of $\mathcal{W}^+$ space.
- Appendix C provides visualizations of the interpretable semantics discovered by unsupervised image manipulation method SeFa [5]. It is not able to handle multi-class image generation nor distinguish category-relevant and category-irrelevant attributes like AGE.
- Appendix D provides visualizations of the ablation "Sample Train" (Section 4.3).
- Appendix E provides additional visualizations of one-shot image generalization from AGE (Section 4.5).
- Appendix F provides additional visualizations and analysis of failure cases from AGE (Section 4.7).
- Appendix G provides additional visualization of disentangled attribute editing directions after SVD (Section 4.6).

A. Quantitative Results on VGG Faces Test Split

In the experiment, another interesting phenomenon is that FID and LPIPS is highly correlated with the number of categories in the test set. For fair comparison, we test our AGE model on VGGFaces [1] with different numbers of categories in the test split. The quantitative test results are shown in Table 1. We can find that the more categories in the test split, the lower FID score and higher LPIPS score the model will get. This is conducive to a more comprehensive evaluation of the model. Our AGE model achieves a better quantitative result with FID 34.86 and LPIPS 0.3294 when there are 572 categories in the test split.

Table 1. Ablations of different numbers of categories in the test split on VGG Faces.

# Categories	VGG Faces	
	FID(↓)	LPIPS(↑)
2	78.83	0.2974
50	41.07	0.3189
200	36.09	0.3212
572	**34.86**	**0.3294**

B. Demonstration of the Assumption of Gaussian Distribution.

We make an assumption that the distribution of the samples in $\mathcal{W}^+$ space obeys Gaussian distribution in Eq. 17.

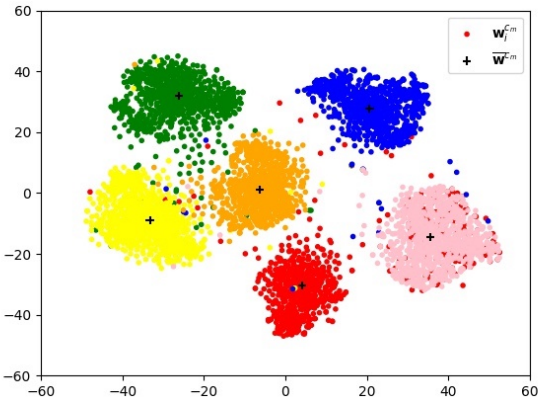


Figure 1. Visualization of $\mathcal{W}^+$ space after TSNE.

This assumption is from StyleGAN that different images can be generated from a center image with linearly interpolation along different directions in the embedding space. In Figure 1, we further illustrate the latent embeddings of samples from 6 different categories after TSNE. The distribution of different categories does roughly follow Gaussian distribution.

C. Comparison with Unsupervised Image Manipulation Methods

The core of the editing-based few-shot image generation is to identify the category-relevant and category-irrelevant attributes in the latent space without explicit supervision. Similar to AGE, unsupervised image manipulation methods [4–6] also study the semantic factorization of a pre-trained GAN. However, they only focus on single-category

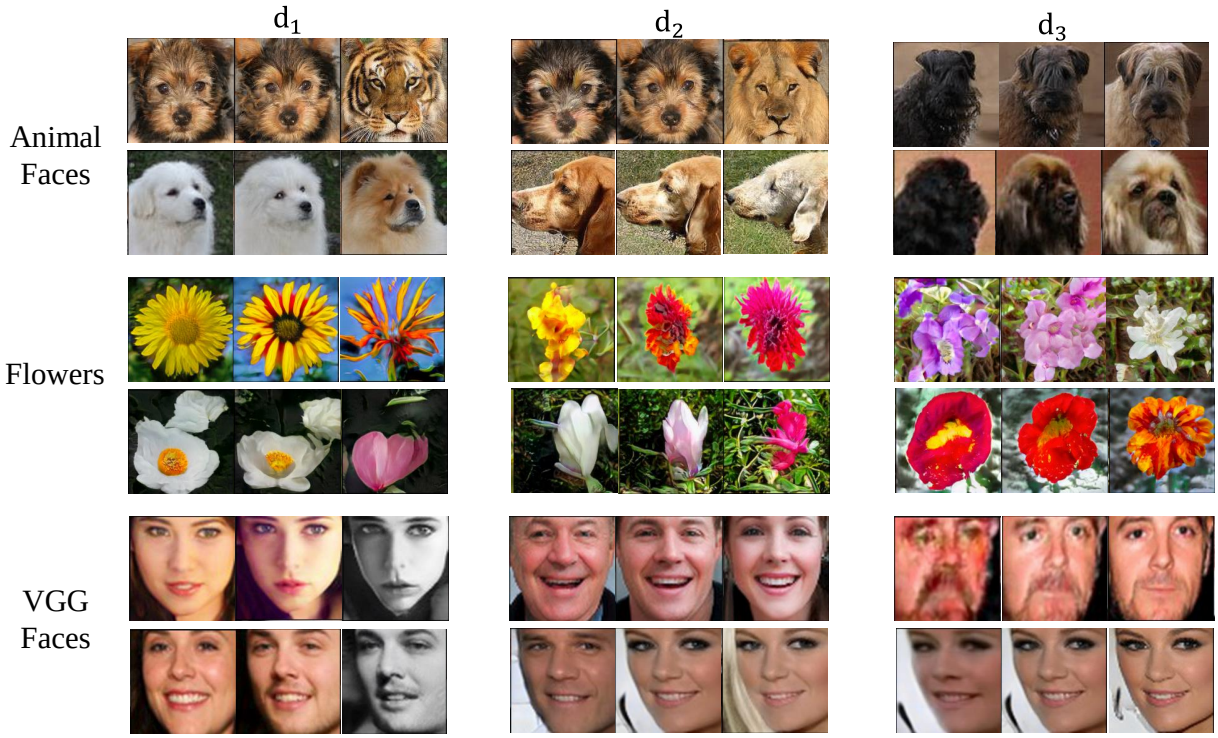


Figure 2. Visualizations of interpretable directions discovered by SeFa. The left and the right images are edited from the middle one. Moving the latent vectors along the discovered directions apparently changes the categories of the images.

image generation that does not care about the categorical information. In order to verify if they can distinguish the category-irrelevant directions for few-shot image generation, we conduct the recent proposed method SeFa [5] on three multi-class image generation datasets. SeFa performs a closed-form factorization on the latent semantics according to the weights of the generator, which is one of the best unsupervised attribute factorization and manipulation methods.

Figure 2 shows the first three directions discovered by SeFa. In complicating datasets Animal Faces [2] and Flowers [3], the category-irrelevant attributes and category-relevant attributes are all entangled. The interpretable semantics are hard to distinguish. Editing along a single direction changes multiple attributes and results in an image of a completely different category (*e.g.* from a dog to tiger, the shape of the petals, etc.). In VGGFaces [1], despite achieving better disentanglement, the top important semantics discovered by SeFa are almost category-relevant including the sex and the shape of the face. In contrast, AGE can factorize the category-irrelevant attributes from the category-relevant attributes, which is the most important for few-shot image generation.

D. Images Generated from "Sample Train"

In Section 4.3, we provide the ablation "Sample Train" that randomly samples $\Delta \mathbf{w}$ of seen categories from the train set and directly use it to edit the unseen categories. As shown in Table 2 in the main paper, it degrades a lot on FID compared with AGE.

In this section, we provide samples generated from "Sample Train" in Figure 3. As shown in the generated samples, directly using the sampled $\Delta \mathbf{w}$ to edit the input images is very unstable. Although some high-quality images can be generated, most images are crashed or change category. This also further proves the necessity of the attribute factorization of AGE.

E. Additional Visualizations for AGE

We provide more samples generated by AGE in Figure 4 and Figure 5.

F. Failure Case Analysis for AGE

We provide more failure cases generated by AGE in Figure 6. Failure cases can be divided into three classes: inversion failure, category change, and editing failure. Most

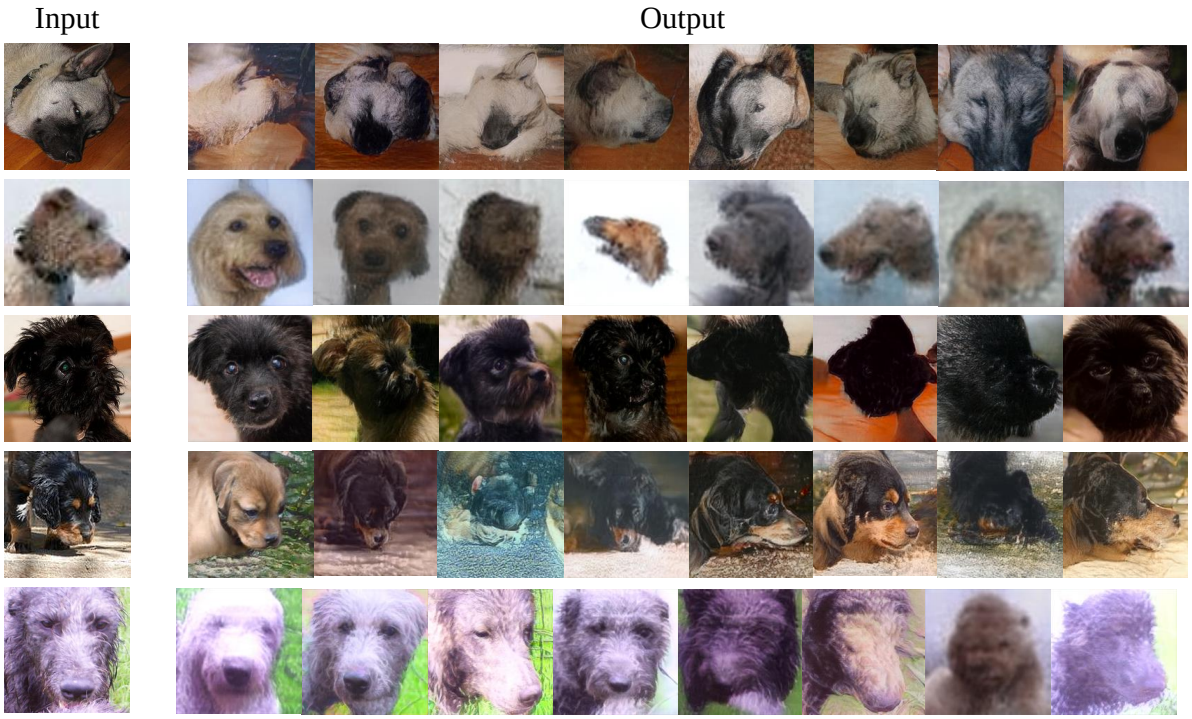


Figure 3. Images edited by random $\Delta\mathbf{w}$ sampled from seen categories.

crashed cases of AGE are caused by the failure of GAN inversion as shown in the Figure 6a. Our editing starts from the latent representation of GAN inversion. Therefore, if the inversion representations cannot reconstruct the input images, both of the attribute factorization and manipulation will fail. GAN inversion is not stable when there aren't enough samples. In Flowers [3], many important category-relevant attributes are lost after inversion. In VGGFaces [1], all glasses are missing after inversion, therefore, this attribute is completely ignored during training.

Second, some editing from AGE may cause category change. This is because some category-irrelevant attributes learned by AGE are not shared among all categories (*e.g.* the number of petals and the shape of cats' face). This situation is more common in the Flowers [3] dataset since the intra-category variations of different kinds of flowers are very distinct.

Third, since the sampling process of sparse representation is based on the statistics of the whole training set, the editing generated from AGE may lead to crashes in the images when encountering extreme cases. We hope our new editing perspective can inspire further researches towards better attribute disentanglement free from pre-trained GAN inversion methods.

G. Additional Disentangled Attribute Editing Directions

We provide more visualizations of disentangled attribute editing directions in different layers/groups learned by AGE in Figure 7. These images are edited along the directions factorized with SVD. The details have been provided in Section 4.6 in the main paper.

As shown in Figure 7, different directions in different layers control different attributes. Editing along certain directions can roughly change one specific attribute continuously. The lower layers like $\mathbf{w}_0$, $\mathbf{w}_1$, $\mathbf{w}_2$ mainly control structure attributes like posture, ear, eye, and head. The higher layers like $\mathbf{w}_3$, $\mathbf{w}_4$, $\mathbf{w}_5$ mainly control surface attributes like hair and color. This is in line with the rules of most GANs' latent spaces.

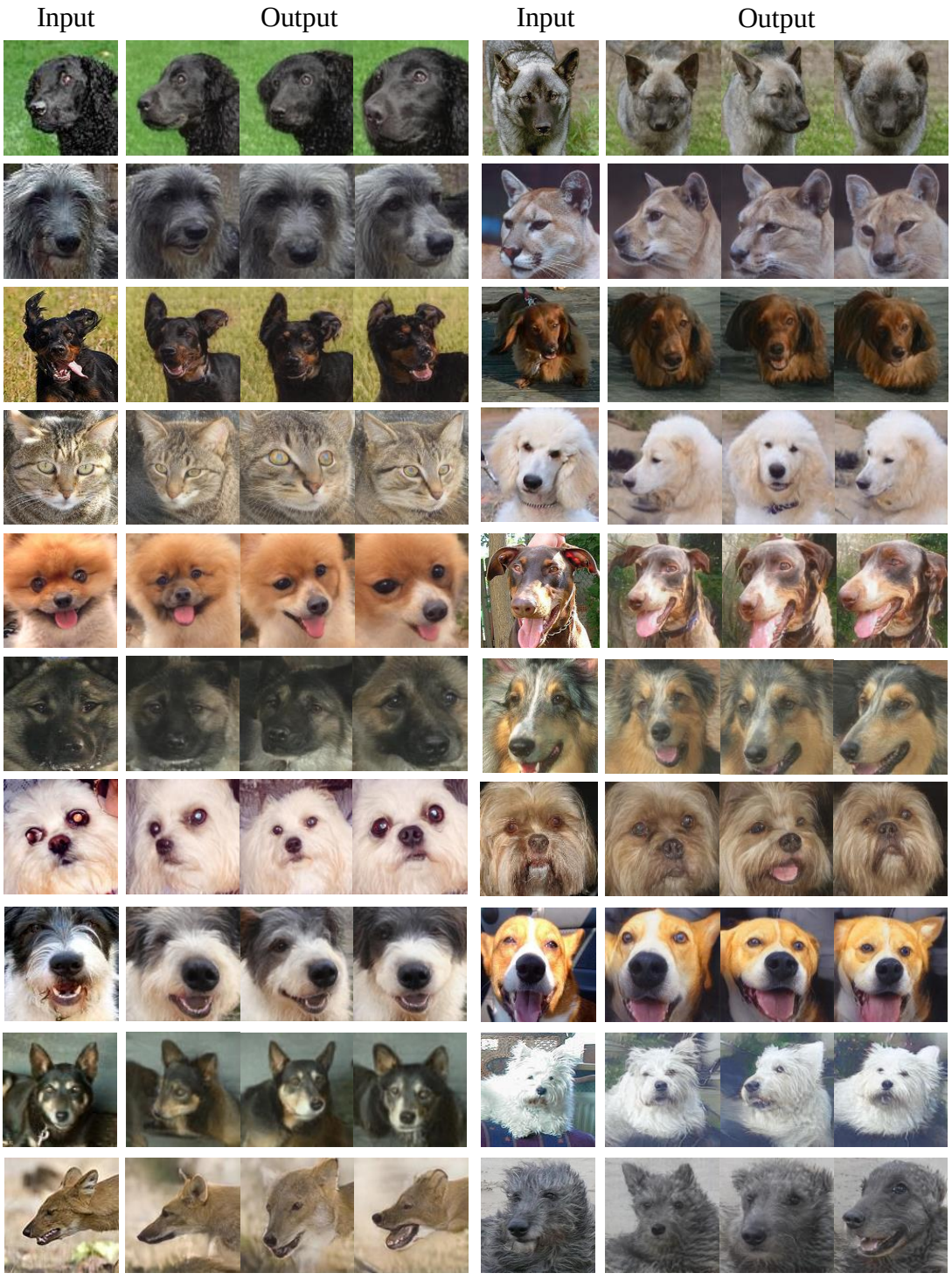


Figure 4. One-shot image generation by AGE on Animal Faces.

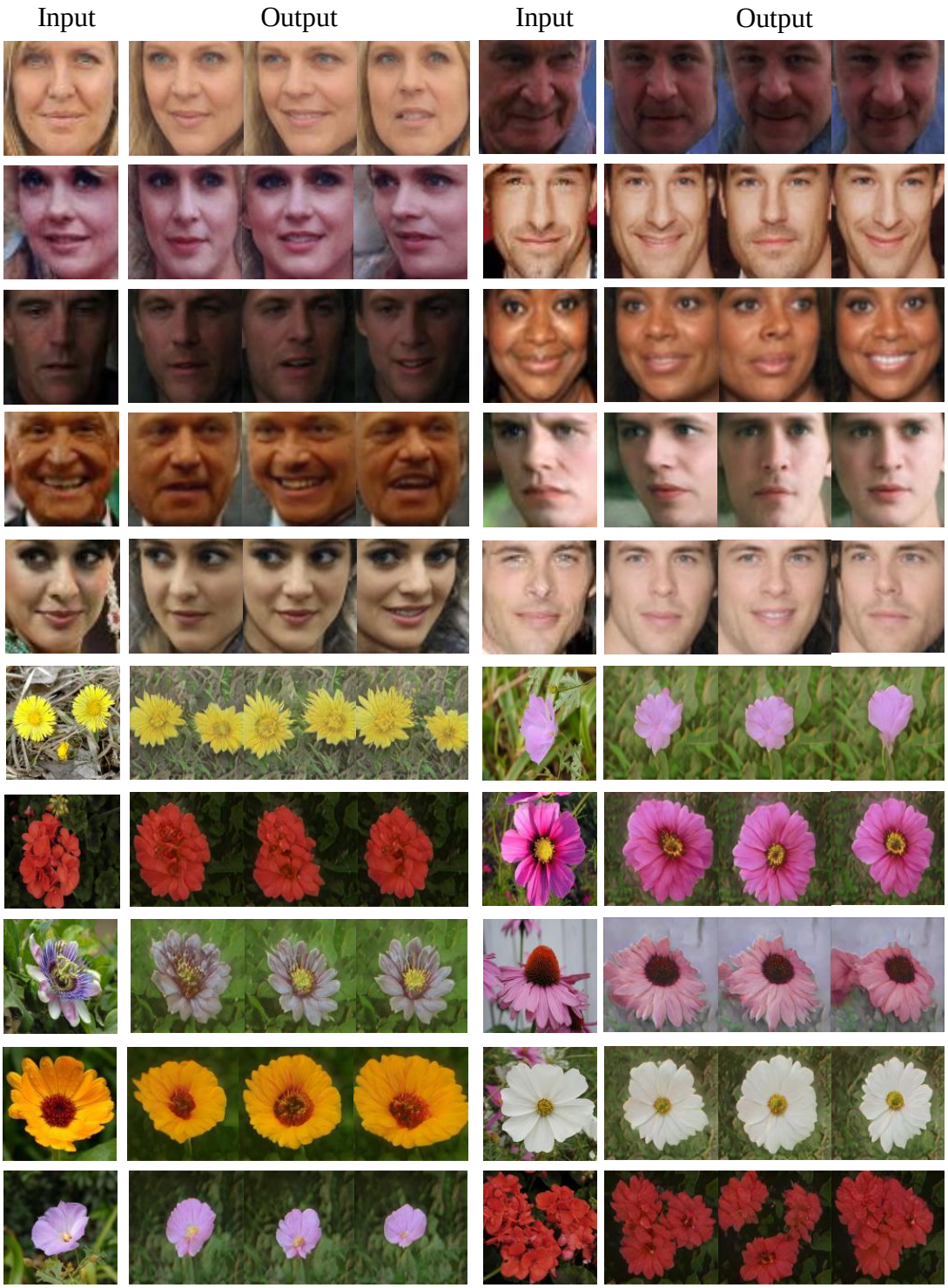
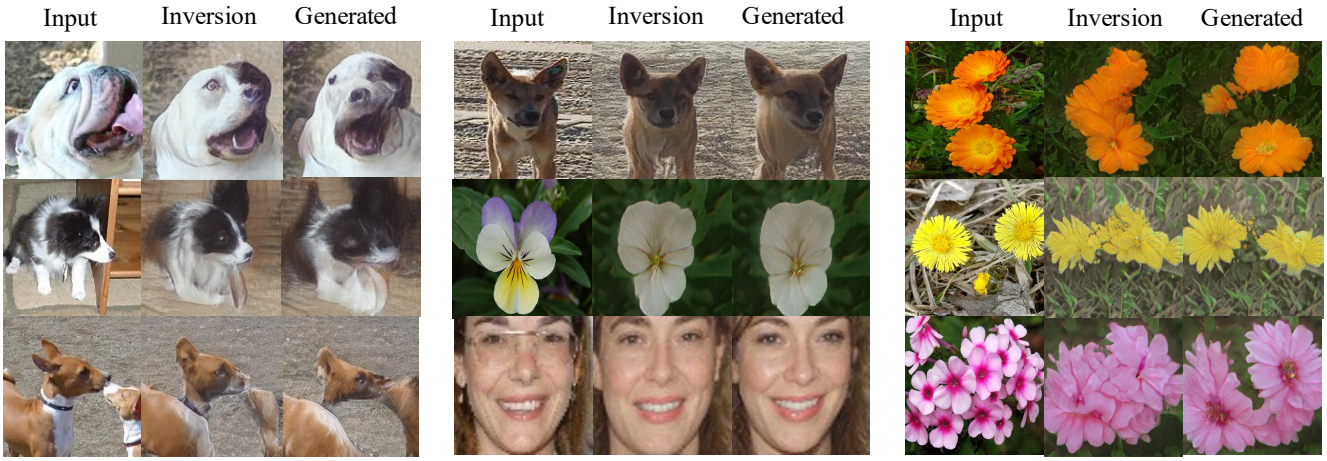


Figure 5. One-shot image generation by AGE on VGGFaces and Flowers.



(a) Inversion Failure

(b) Category Change

(c) Editing Failure

Figure 6. Failure Cases from AGE.

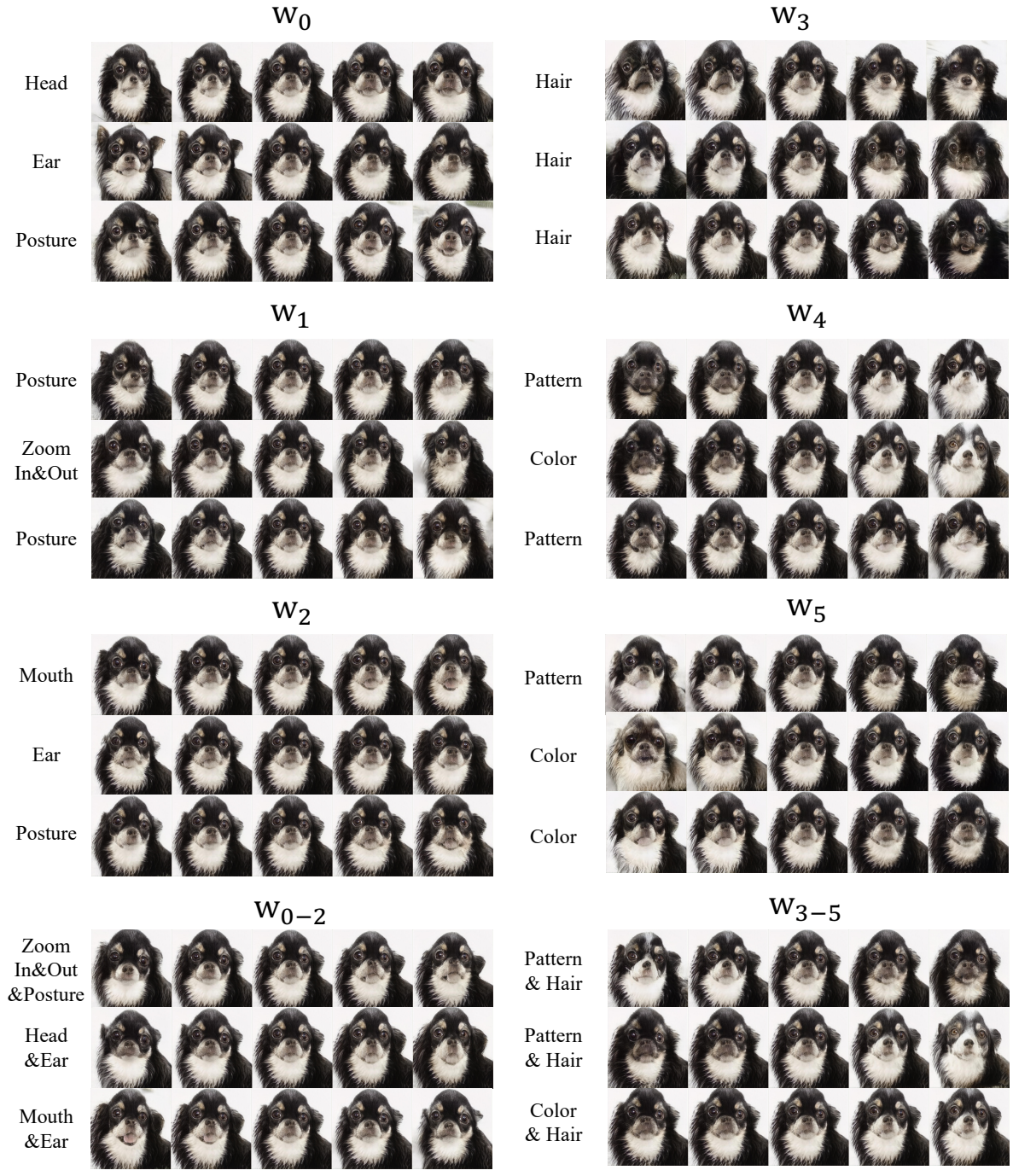


Figure 7. Visualizations of disentangled attribute editing directions in different layers/groups learned by AGE.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018. 1, 2, 3

[2] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsueprvised image-to-image translation. In *ICCV*, 2019. 2

[3] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 2, 3

[4] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*, 2020. 1

[5] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 1, 2

[6] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *ICCV*, 2021. 1