

Appendix B for EigenGAN

1. All Interpretable Dimensions of Each Layer

Fig. 1 to Fig. 5 show all interpretable dimensions of each generator layer learned by the proposed EigenGAN on CelebA dataset [3], Fig. 6 to Fig. 9 show all dimensions learned on Anime dataset [1], and Fig. 10 to Fig. 14 show all dimensions learned on FFHQ dataset [2]. We traverse each dimension from -4.5σ to 4.5σ and omit the dimensions with almost no change. “None” in these figures means the corresponding part of the dimension is uninterpretable or difficult to be assigned an attribute name. The smaller the index, the deeper the layer.

2. Effect of the Importance Matrix L_i

In Sec. 3.1 of the main text, we introduce the importance matrix $L_i = \text{diag}(l_{i1}, \dots, l_{iq})$ with l_{ij} deciding the importance or energy of the basis vector \mathbf{u}_{ij} . In Fig. 15, we compare dimensions with different importance values. As can be seen, dimensions with large importance value control large variations while dimensions with small importance value control slight variations. It cannot be judged whether the larger one of two large enough importance values controls larger variation, because it is hard to quantify the semantic variations. However, it can be sure that dimensions with small importance value can only control slight changes or even no change. Therefore, to some extent, the importance matrix can be used to select the learned semantic dimensions by discarding dimensions with small importance value. Note that the values of L_i across different layers ($i = 1, 2, \dots, 6$) cannot be directly compared since they belong to different subspaces.

3. Network Architectures

The architectures of the generator and the discriminator of EigenGAN are shown in Fig. 16.

References

- [1] Gwern Branwen, Anonymous, and Danbooru Community. Danbooru2019 portraits: A large-scale anime head illustration dataset, 2019. 1, 7, 8, 9, 10
- [2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In

IEEE Conf. Comput. Vis. Pattern Recog., 2019. 1, 11, 12, 13, 14, 15

- [3] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, 2015. 1, 2, 3, 4, 5, 6
- [4] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Int. Conf. Mach. Learn.*, 2013. 17

L6 (Layer 6, shallowest)

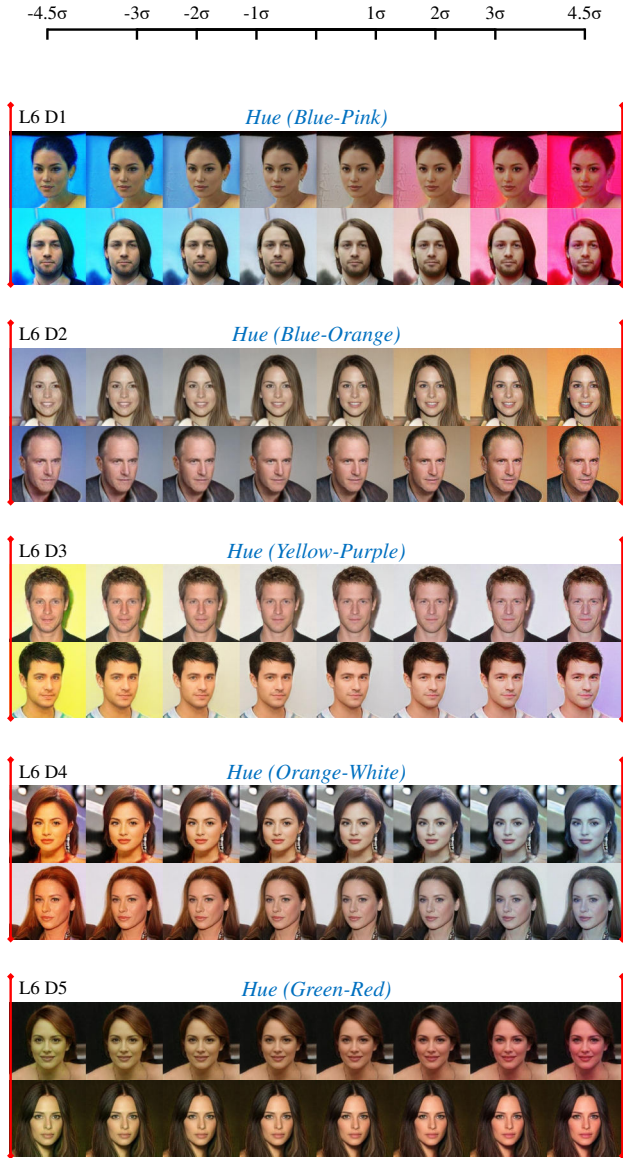
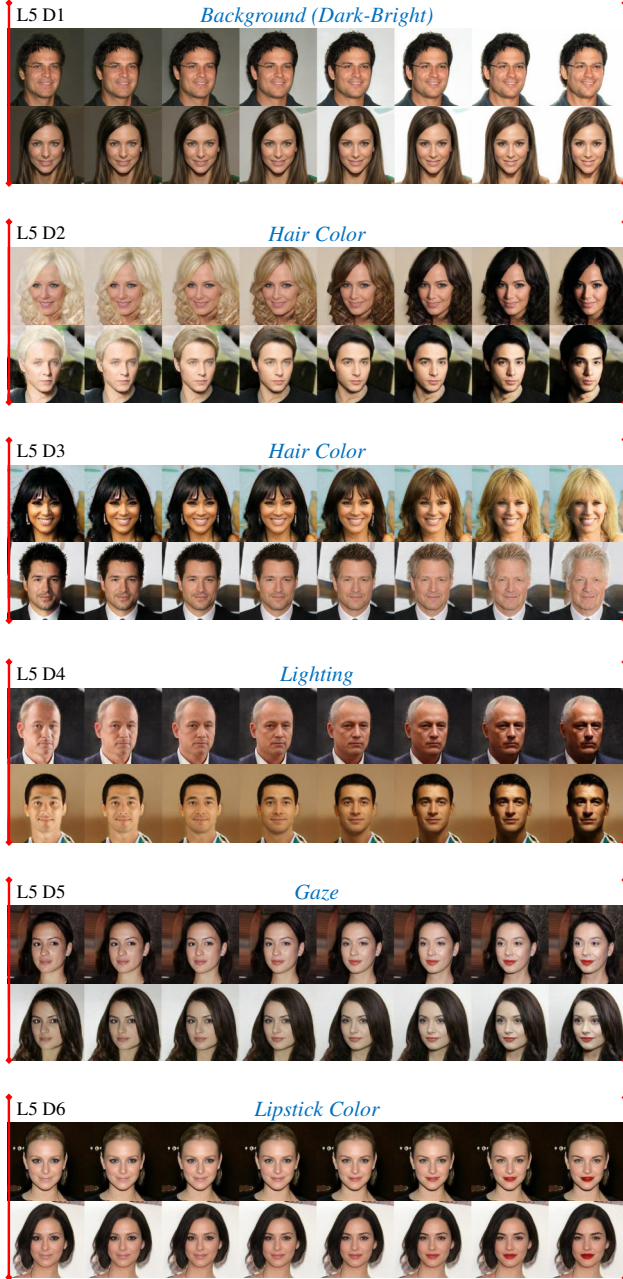
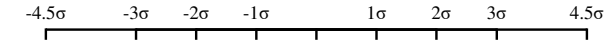


Figure 1. Interpretable dimensions of layer 6 (the shallowest) for CelebA dataset [3].

L5 (Layer 5)



L4 (Layer 4)

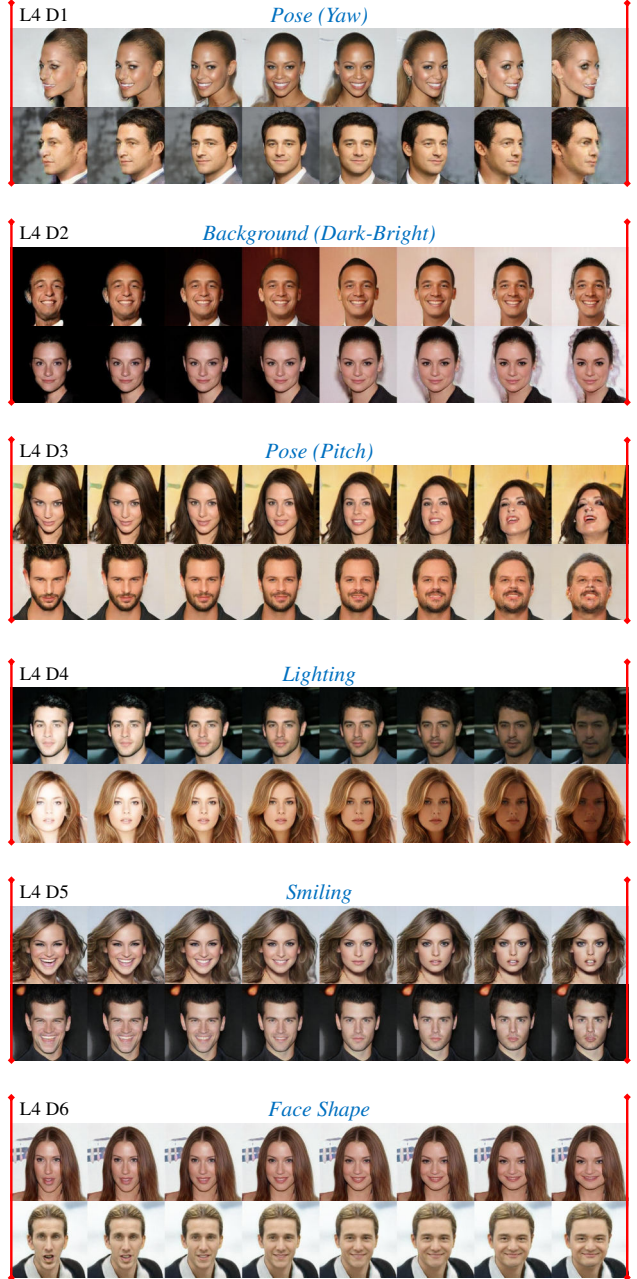
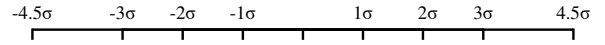


Figure 2. Interpretable dimensions of layer 5 (left) and layer 4 (right) for CelebA dataset [3].

L3 (Layer 3)



Figure 3. Interpretable dimensions of layer 3 for CelebA dataset [3].

L2 (Layer 2)



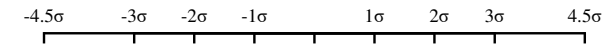
Figure 4. Interpretable dimensions of layer 2 for CelebA dataset [3]. “None” means uninterpretable.

L1 (Layer 1, deepest)



Figure 5. Interpretable dimensions of layer 1 (the deepest) for CelebA dataset [3]. “None” means uninterpretable.

L6 (Layer 6, shallowest)



L5 (Layer 5)

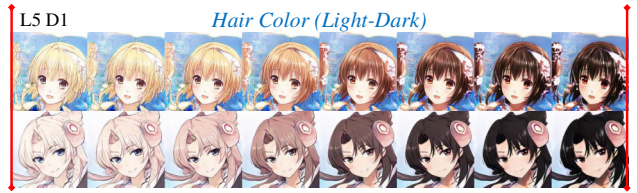
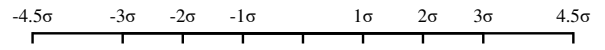
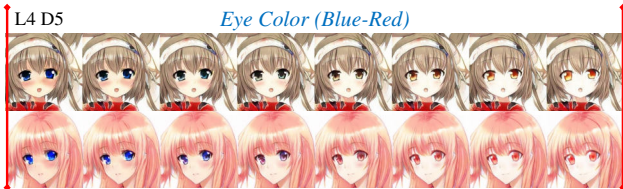
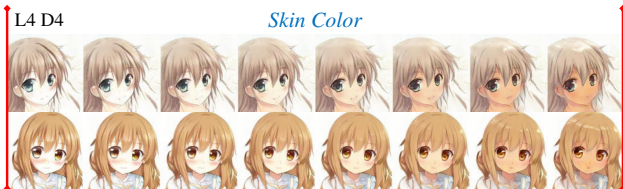
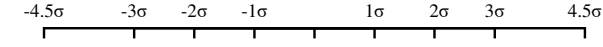


Figure 6. Interpretable dimensions of layer 6 (left, the shallowest) and layer 5 (right) for Anime dataset [1].

L4 (Layer 4)



L3 (Layer 3)

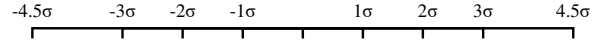


Figure 7. Interpretable dimensions of layer 4 (left) and layer 3 (right) for Anime dataset [1].

L2 (Layer 2)

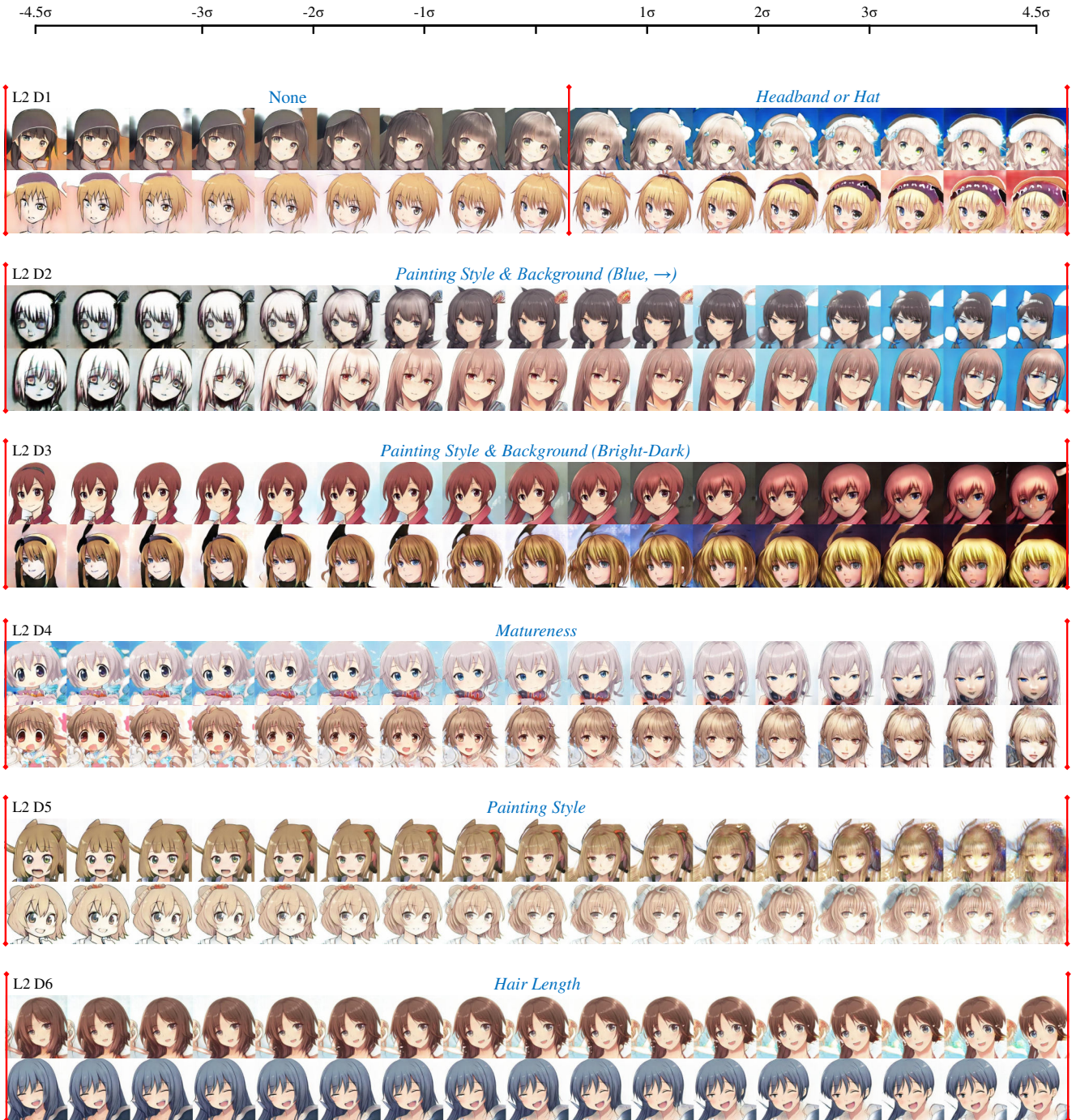


Figure 8. Interpretable dimensions of layer 2 for Anime dataset [1]. “None” means uninterpretable.

L1 (Layer 1, deepest)

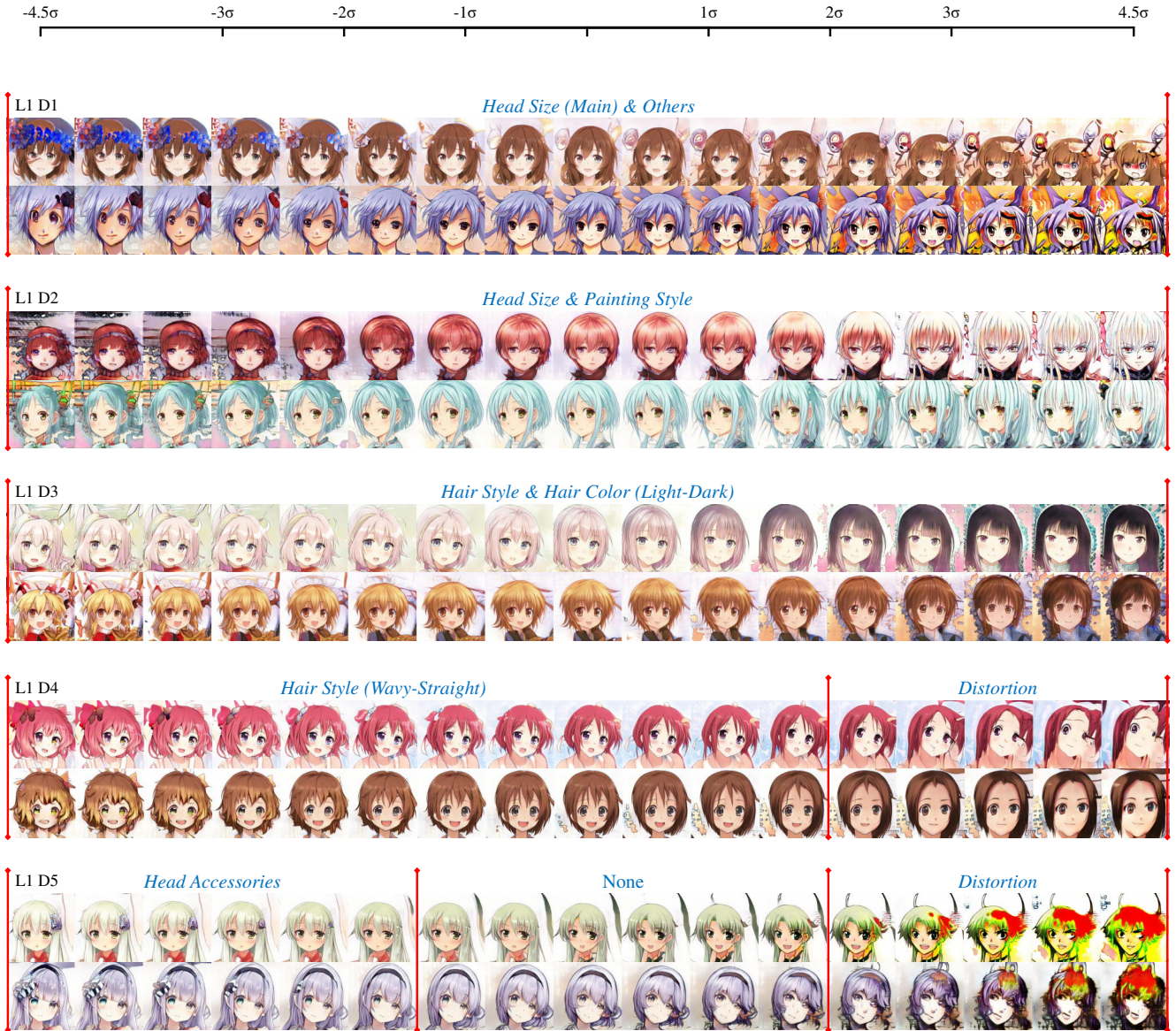


Figure 9. Interpretable dimensions of layer 1 (the deepest) for Anime dataset [1]. “None” means uninterpretable.

L6 (Layer 6, shallowest)

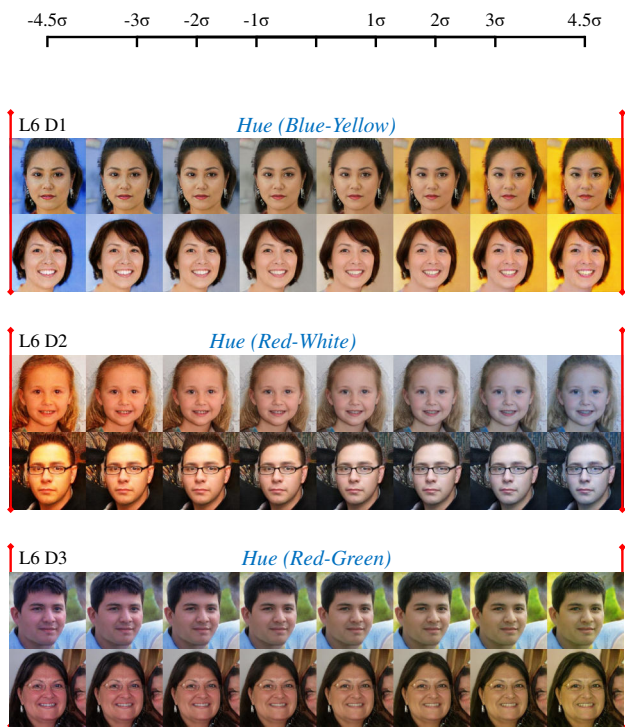
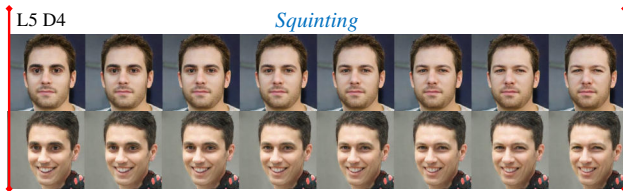
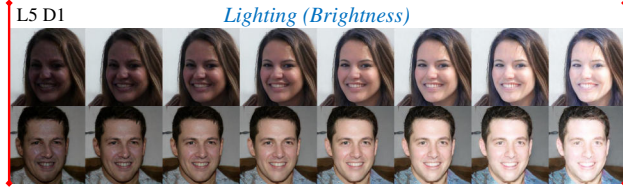
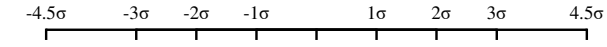


Figure 10. Interpretable dimensions of layer 6 (the shallowest) for FFHQ dataset [2].

L5 (Layer 5)



L4 (Layer 4)

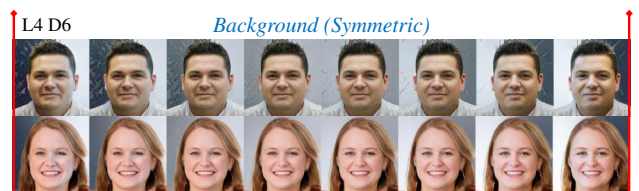
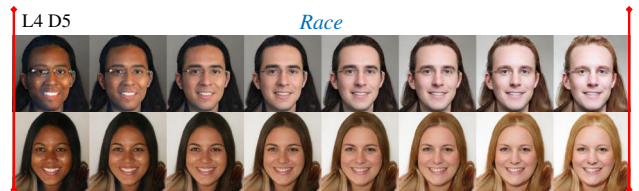
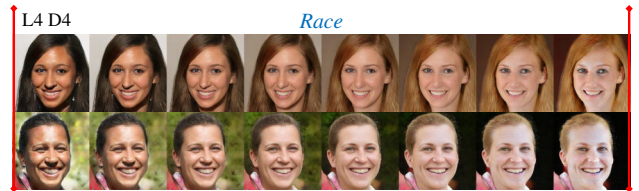
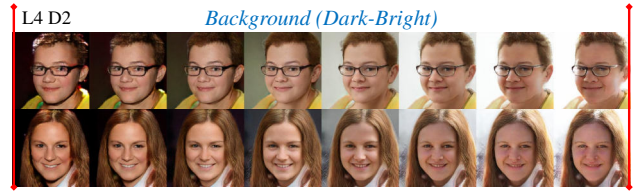
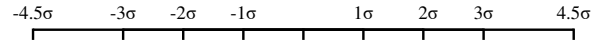


Figure 11. Interpretable dimensions of layer 5 (left) and layer 4 (right) for FFHQ dataset [2].

L3 (Layer 3)



Figure 12. Interpretable dimensions of layer 3 for FFHQ dataset [2].

L2 (Layer 2)



Figure 13. Interpretable dimensions of layer 2 for FFHQ dataset [2].

L1 (Layer 1, deepest)

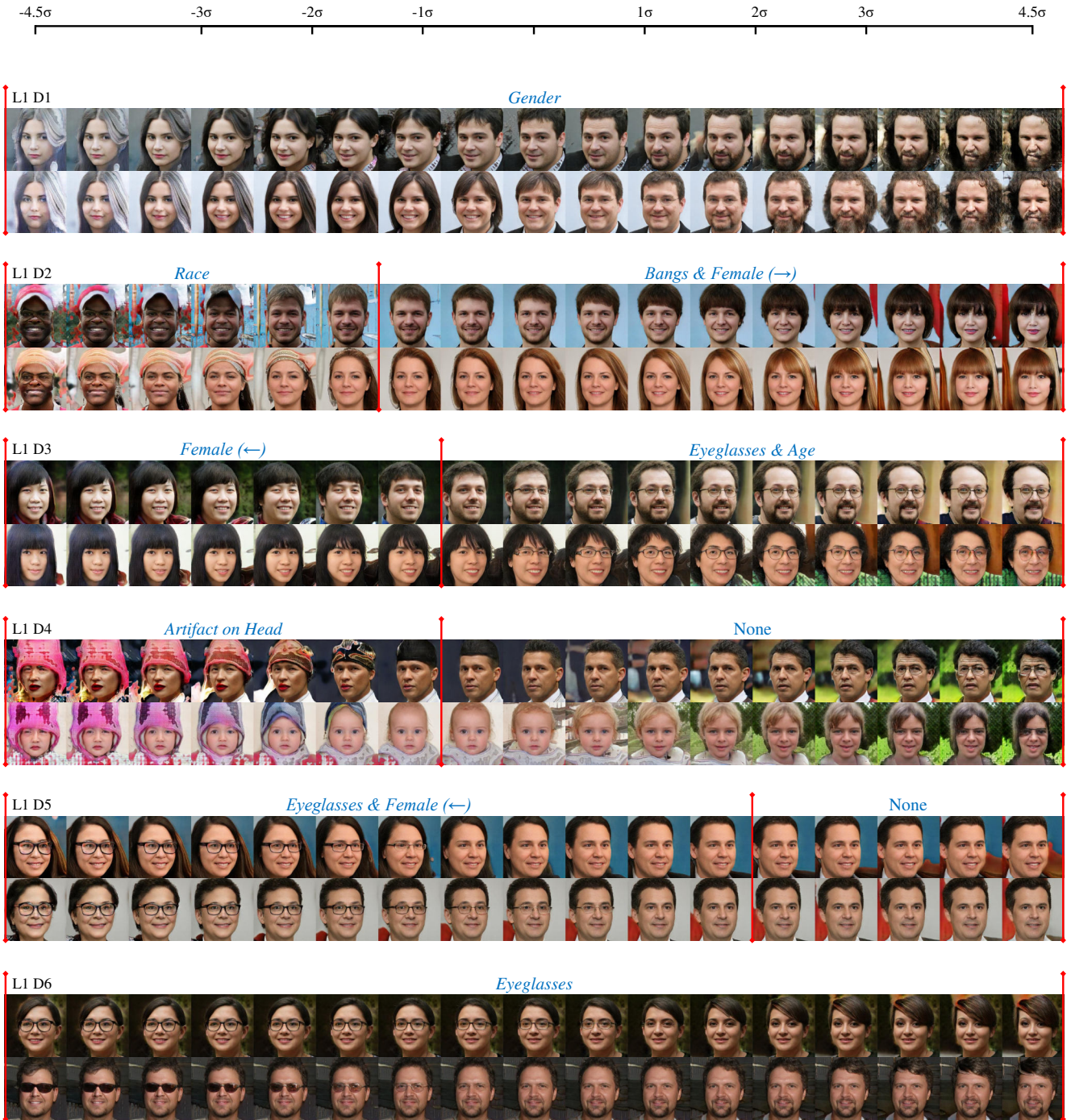


Figure 14. Interpretable dimensions of layer 1 (the deepest) for FFHQ dataset [2]. “None” means uninterpretable.

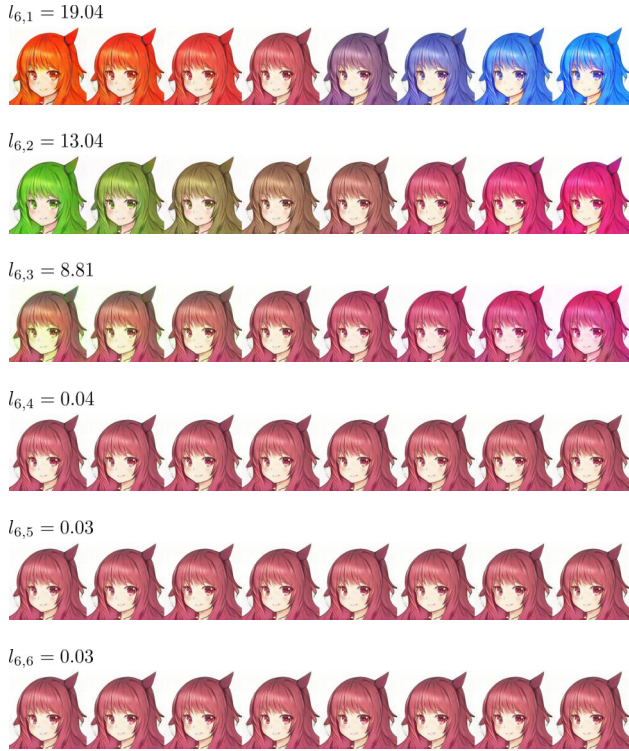


Figure 15. Effect of the importance values $\mathbf{L}_i = \text{diag}(l_{i1}, \dots, l_{iq})$. Dimensions with large importance value control large variations while dimensions with small importance value control small variations.

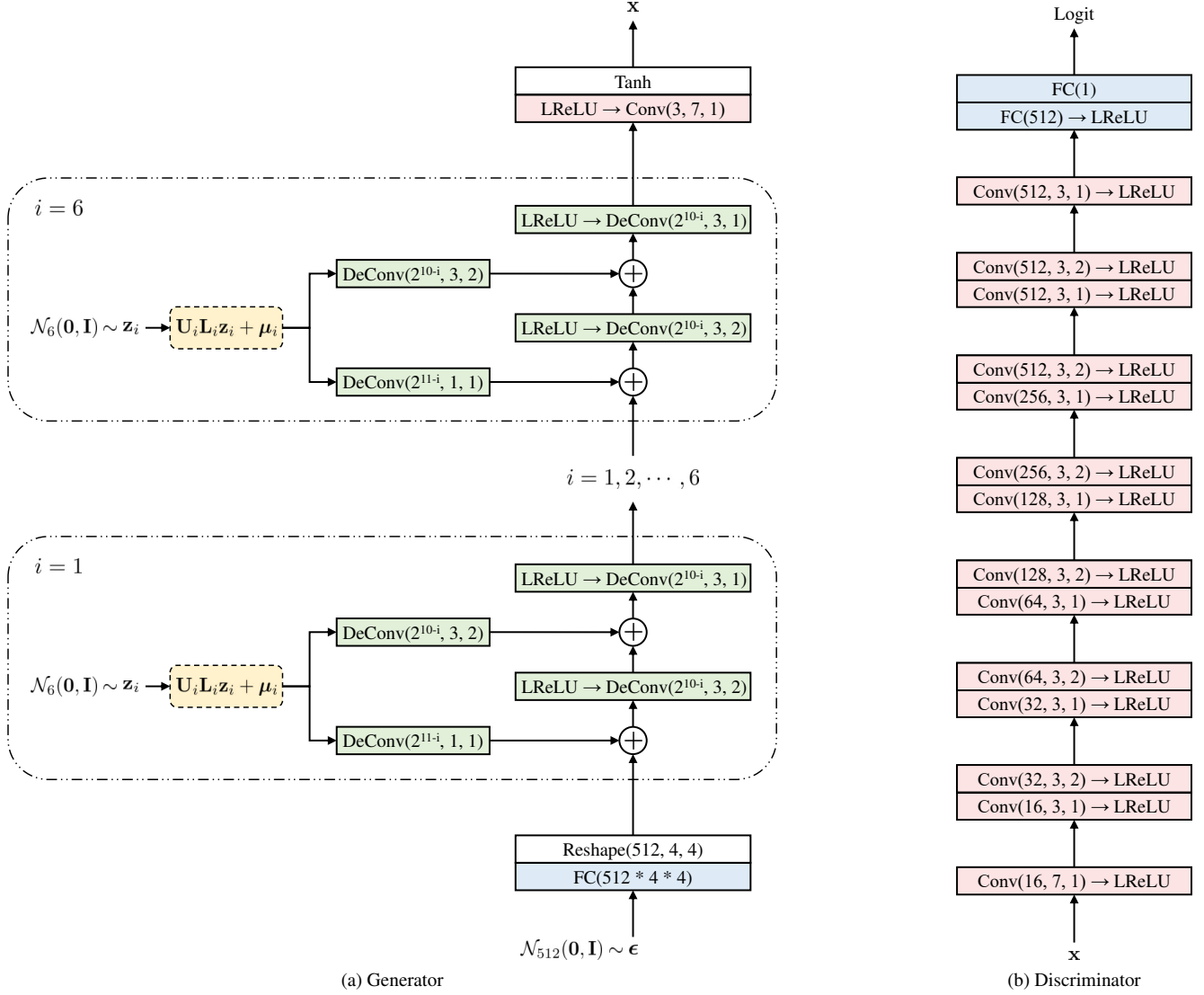


Figure 16. Network architectures of EigenGAN. Conv(d, k, s) and DeConv(d, k, s) denote convolutional layer and transposed convolutional layer with d as output dimensions, k as kernel size, and s as stride. FC(d) denotes fully connected layer with d as output dimensions. LReLU denotes Leaky ReLU [4].