Householder Projector for Unsupervised Latent Semantics Discovery –Supplementary Material–

Yue Song[1], Jichao Zhang[1], Nicu Sebe[1], and Wei Wang[2]

[1]Department of Information Engineering and Computer Science, University of Trento, Italy
[2]Beijing Jiaotong University, China
yue.song@unitn.it

A. Limitation and Future Work

Our current experiments only validate our approach in fine-tuning StyleGANs. Despite the easy usage, the image fidelity and disentanglement performance might be better if we could train StyleGANs equipped with our Householder Projector from scratch. However, due to the limited computational resources, this point cannot be validated for now. Additionally, in the current setting, we pre-define the number of semantics of each layer to a fixed number (the rank of the projector). Seeking an adaptive scheme to automatically mine the semantics would be also an important direction of our future work.

B. Mathematical Derivation

B.1. Decomposing $\mathbf{U}$ and $\mathbf{V}$

For $n\times n$ orthogonal matrix $\mathbf{U}$, there exists $\mathbf{H}_1\mathbf{H}_2\dots\mathbf{H}_n{=}\mathbf{U}$ where $\mathbf{H}_i$ is a Householder reflection matrix. The decomposition is achieved by the *n-reflections theorem*: each $\mathbf{H}_i$ can be designed to zero out the non-diagonal entries of $\mathbf{U}$ in the i-th column and row and to set the diagonal entry to 1. Such accumulation of n reflectors can transform $\mathbf{U}$ into an identity matrix ($\mathbf{U}\mathbf{H}_n\dots\mathbf{H}_2\mathbf{H}_1{=}\mathbf{I}$). Since $\mathbf{H}_i$ is a reflection ($\mathbf{H}_i\mathbf{H}_i{=}\mathbf{I}$), this theorem directly gives the relation $\mathbf{U}{=}\mathbf{H}_1\mathbf{H}_2\dots\mathbf{H}_n$.

B.2. Orthogonality Preservation

The orthogonality of a Householder matrix $\mathbf{H}_i$ can be easily verified by:

$$\begin{aligned}\mathbf{H}_i\mathbf{H}_i^T &= \Big(\mathbf{I}{-}2\frac{\mathbf{h}_i\mathbf{h}_i^T}{||\mathbf{h}_i||_2^2}\Big)\Big(\mathbf{I}{-}2\frac{\mathbf{h}_i\mathbf{h}_i^T}{||\mathbf{h}_i||_2^2}\Big)^T\\ &= \frac{1}{||\mathbf{h}_i||_2^4}(||\mathbf{h}_i||_2^2\mathbf{I}-2\mathbf{h}_i\mathbf{h}_i^T)(||\mathbf{h}_i||_2^2\mathbf{I}-2\mathbf{h}_i\mathbf{h}_i^T)^T\\ &= \frac{1}{||\mathbf{h}_i||_2^4}\Big(||\mathbf{h}_i||_2^4\mathbf{I}-4||\mathbf{h}_i||_2^2\mathbf{h}_i\mathbf{h}_i^T+4\mathbf{h}_i\mathbf{h}_i^T\mathbf{h}_i\mathbf{h}_i^T\Big)\\ &= \frac{1}{||\mathbf{h}_i||_2^4}\Big(||\mathbf{h}_i||_2^4\mathbf{I}-4||\mathbf{h}_i||_2^2\mathbf{h}_i\mathbf{h}_i^T+4||\mathbf{h}_i||_2^2\mathbf{h}_i\mathbf{h}_i^T\Big)\\ &= \frac{1}{||\mathbf{h}_i||_2^4}||\mathbf{h}_i||_2^4\mathbf{I}\\ &= \mathbf{I}\end{aligned} \tag{1}$$

Similarly, when a gradient descent step is performed (*i.e.,* $(\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i)$, we still have the relation:

$$\begin{aligned}&(\mathbf{I}-\frac{(\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i)(\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i)^T}{||\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i||_2^2})(\mathbf{I}-\frac{(\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i)(\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i)^T}{||\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i||_2^2})^T\\ &\qquad= \frac{1}{||\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i||_2^4}\Big(||\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i||_2^4\mathbf{I}\Big)=\mathbf{I}\end{aligned} \tag{2}$$

The orthogonality is preserved during the back-propagation and weight update phase.

B.3. Householder Representation

With the previous results on orthogonality preservation of a Householder matrix, we can proceed to show how an orthogonal matrix can be represented by the accumulation of elementary Householder reflectors. Given a square orthogonal eigenvector matrix defined as:

$$\mathbf{U}=\sum_{i=1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^T \tag{3}$$

where $\mathbf{u_i}$ denotes the eigenvector of $\mathbf{U}$, and $\lambda_i \subset \{-1, 1\}$ is the eigenvalue. Let $\prod_{j=1}^{d} \mathbf{H}_j$ be the accumulation of Householder reflectors as:

$$\prod_{j=1}^{d} \mathbf{H}_j = \prod_{j=1}^{d} (\mathbf{I} - 2\frac{\mathbf{h}_j \mathbf{h}_j^T}{||\mathbf{h}_j||_2^2}) \tag{4}$$

The eigenvector property directly gives

$$\prod_{j=1}^{d} \mathbf{H}_j \mathbf{u}_i = \prod_{j=1}^{d} (\mathbf{I} - 2\frac{\mathbf{h}_j \mathbf{h}_j^T}{||\mathbf{h}_j||_2^2}) \mathbf{u}_i \tag{5}$$

If we set $\mathbf{h}_j = \mathbf{u}_i$ for $i = j$, the orthogonality would naturally lead to

$$\begin{aligned} (\mathbf{I} - 2\frac{\mathbf{h}_j \mathbf{h}_j^T}{||\mathbf{h}_j||_2^2}) \mathbf{u}_i &= \mathbf{u}_i, \ i \neq j \\ (\mathbf{I} - 2\frac{\mathbf{h}_j \mathbf{h}_j^T}{||\mathbf{h}_j||_2^2}) \mathbf{u}_i &= \lambda_i \mathbf{u}_i, \ i = j \end{aligned} \tag{6}$$

Eq. (5) is further simplified as:

$$\prod_{j=1}^{d} \mathbf{H}_j \mathbf{u}_i = \lambda_i \mathbf{u}_i = \mathbf{U} \mathbf{u}_i \tag{7}$$

The above equation shows that the relation $\mathbf{U} = \prod_{j=1}^{d} \mathbf{H}_j$ holds. This indicates that any orthogonal matrices can be represented by a series of Householder accumulations.

B.4. Semi-orthogonality of Non-Square Matrices

For the fluency of text flow, we do not differentiate the projector $\mathbf{A}$ from square or non-square matrices in the paper. Strictly speaking, non-square matrices with orthonomal rows or columns (depending on whether $\mathbf{A}$ is a flat matrix or tall matrix) should be called semi-orthogonal matrices more precisely. Here we give a special note for the strictness of math definitions, but this does not influence the core contribution of our method or any experimental results.

C. Details of Datasets and Metrics

C.1. Datasets

StyleGAN2 Datasets. FFHQ [11] consists of $70,000$ high-quality face images that have considerable variations in identifies and have good coverage in common accessories. LSUN Church [16] has $126,227$ scenes images of outdoor churches, and LSUN Cat [16] is comprised of $1,657,266$ different cat images collected online.

StyleGAN3 Datasets. MetFaces [7] contains $1,336$ high-quality human faces extracted from works of arts. AFHQv2 [1] is a dataset consisting of $15,803$ animal faces from three different domains, including cat, dog, and wildlife. SHHQv1 [3] covers $40,000$ images of diverse full-body clothed humans in its current version. Notice that their pre-trained models use $230,000$ images for training but only a subset of the training set is released. We expect that using the complete set for training would further improve the FID score of our method on SHHQ.

C.2. Metrics

Fréchet Inception Distance (FID) [5]. FID assesses the Fréchet distance of deep features between the set of generated images and the set of real images. More formally, given the feature distribution $\mathcal{N}(\mu, \Sigma)$ of real images and the feature distribution $\mathcal{N}(\mu', \Sigma')$ of fake images, the distance is computed as:

$$d_F = \sqrt{||\mu - \mu'||_2^2 + \mathrm{tr}\Big(\Sigma + \Sigma' - 2(\Sigma^{\frac{1}{2}} \Sigma' \Sigma^{\frac{1}{2}})^{\frac{1}{2}}\Big)} \tag{8}$$

A small value would indicate that the distance between distributions is close and the generated images are realistic. Our FID score is computed based on $50,000$ samples.

Perceptual Path Length (PPL) [9] and Perceptual Interpretable Path Length (PIPL). PPL subdivides the interpolation path into linear segments and measures the perceptual image distance of the segmented path. Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be the randomly sampled latent code in the $\mathcal{W}$ space of StyleGANs. Then PPL defined in the $\mathcal{W}$ space is calculated as:

$$\begin{aligned} \mathrm{PPL}_{\mathcal{W}} = \mathbb{E}\Big[\frac{1}{\epsilon^2} d(&G(\mathrm{lerp}(\mathbf{w}_1, \mathbf{w}_2, t), \\ &G(\mathrm{lerp}(\mathbf{w}_1, \mathbf{w}_2, t + \epsilon)))\Big] \end{aligned} \tag{9}$$

where $d(\cdot)$ represents the LPIPS [17] distance, $\mathrm{lerp}(\cdot)$ denotes the spherical interpolation function, t is a random variable sampled from $U(0, 1)$, and ϵ is the subdivision constant, respectively. The division coefficient ϵ is set to $1e-4$ for all the experiments.

The metric PPL suits use cases where the latent code is randomly interpolated. However, when the latent code is moved around as $\mathbf{z} + \mathbf{n}$ where $\mathbf{n} \in \mathbb{R}^d$ is an interpretable direction sampled from a given vector set (*i.e.,* the eigenvectors extracted by SeFa [14]), the PPL score can not reflect the smoothness of latent space. To make the score adapt to such vector-based manipulations, we propose our PIPL metric by naturally incorporating orthogonal vector perturbations into PPL. Formally, the PIPL is defined as:

$$\begin{aligned} \mathrm{PIPL}_{\mathcal{W}} = \mathbb{E}\Big[\frac{1}{\epsilon^2} d(&G(\mathrm{lerp}(\mathbf{w}_1, \mathbf{w}_2, t), \\ &G(\mathrm{lerp}(\mathbf{w}_1, \mathbf{w}_2, t) + \epsilon \mathbf{n}))\Big] \end{aligned} \tag{10}$$

where $\mathbf{n}$ is an orthogonal vector (*i.e.,* $\mathbf{n}^T \mathbf{n} = 1$) sampled from the given vector set. Here different vector sets are

Steps	FID (↓)	PPL (↓)	PIPL (↓)
0%	18.97	799.38	0.101
0.25%	10.56	427.90	0.057
0.5%	9.31	474.12	0.060
1%	8.46	526.26	0.057
2%	8.10	544.31	0.056
Original StyleGAN2	8.37	722.24	0.141

Table 1. Impact of different fine-tuning steps (% of the original training steps) on LSUN Cat [16] with StyleGAN2 [10].

used because each model is fine-tuned and the interpretable directions are changed. It is thus more reasonable to use the corresponding directions of each method for evaluation. Since the impact of orthogonal vector perturbation is very small in the perceptual distance change, we set ϵ as 1 for StyleGAN2 and as $1e{-}2$ for StyleGAN3 to avoid the magnification by $1/\epsilon^2$. We use different ϵ for StyleGAN2 and StyleGAN3 because these two models have different levels of sensitivities to the latent perturbation. StyleGAN3 is less sensitive due to the intrinsic equivariance properties and also the fact that we insert fewer layers. Compared with PPL, our proposed PIPL can better assess the vector-based latent disentanglement approaches. Both PPL and PIPL are computed with $10,000$ samples.

Face Attribute Correlation. For the attribute correlation, we first use S3FD [18] to extract the face region and then compute the normalized Pearson's correlation between the traversal steps and the predictions using several pre-trained attributes estimators, including FairFace [6] for face attributes (age, race, glasses, and gender) and HopeNet [2] for face poses. Among the pool of interpretable directions, the direction with the highest correlation is deemed to control the attribute. The results are averaged based on $2K$ same samples generated by PTI [13].

D. Ablation Studies

This section presents the ablations on studying the impact of fine-tuning steps, batch size, initialization schemes, low-rank orthogonality, and acceleration techniques.

D.1. Impact of Fine-tuning Steps.

Table 1 evaluates the impact of fine-tuning steps on the performance. When the number of fine-tuning steps increases, the FID score and the image fidelity improve. However, the PPL smoothness deteriorates as FID improves. This can be considered as a trade-off between image quality and latent smoothness. We choose 1% fine-tuning steps to avoid incurring large computational burdens. Nonetheless, one can always choose an appropriate step if a better FID score is required.

BS \ Metrics	FID (↓)	PPL (↓)	PIPL (↓)
8	5.78	468.99	0.029
16	4.94	473.19	0.031
32	3.72	457.52	0.030

Table 2. Impact of Batch Size (BS) on the quality of generated images on LSUN Church [16] with StyleGAN2 [10].

Initialization Scheme	FID (↓)	PPL (↓)	PIPL (↓)
Random Initialization	4.89	978.79	0.160
Nearest-orthogonal Mapping	**4.40**	**966.23**	**0.141**

Table 3. Impact of initialization schemes on FFHQ [11].

Computation Method	Vanilla Accumulation	Accelerated Accumulation
Time (ms)	68.02	**2.67**

Table 4. Computation time cost for Householder accumulation of representing $512{\times}512$ matrices measured on a RTX A6000 GPU.

D.2. Impact of Batch Size

Table 2 presents the image fidelity and the latent space smoothness when different batch sizes are used for fine-tuning. When the batch size increases, the FID score has also steady improvements, while the latent space smoothness is mildly influenced. This indicates that the batch size can greatly affect image quality. We believe that using a larger batch size can further boost the FID score of our method, particularly in StyleGAN3 experiments where our batch size is actually smaller than the original setting due to computational resources.

D.3. Impact of Initialization and Acceleration.

Table 3 compares the performance of different initialization schemes. The proposed nearest-orthogonal initialization maps the pre-trained projector into the nearest orthogonal form, which leverages the statistic of well-trained network weights. It thus outperforms the ordinary random initialization. Table 4 shows the computational time of our accelerated Householder aggregation. The acceleration technique significantly improves 25 times the speed of vanilla accumulation, enabling efficient implementation of our Householder Projector in deep neural networks. The marginal time cost would not bring much computational overhead to generative models.

D.4. Impact of Low-rank Orthogonality

Table 5 presents the quantitative evaluation results on the impact of projector rank. The FID score of the full-rank projector falls behind that of the low-rank projector. This stems from the fact the full-rank projector might be slower to converge and harder to optimize within the very limited fine-tuning steps. In terms of latent smoothness, the

Rank	FID (↓)	PPL (↓)	PIPL (↓)
512 (full rank)	4.34	390.89	0.025
10 (low rank)	3.72	457.52	0.030
5 (low rank)	3.65	461.76	0.032

Table 5. Impact of different matrix rank for our Householder Projector on LSUN Church [16] with StyleGAN2.

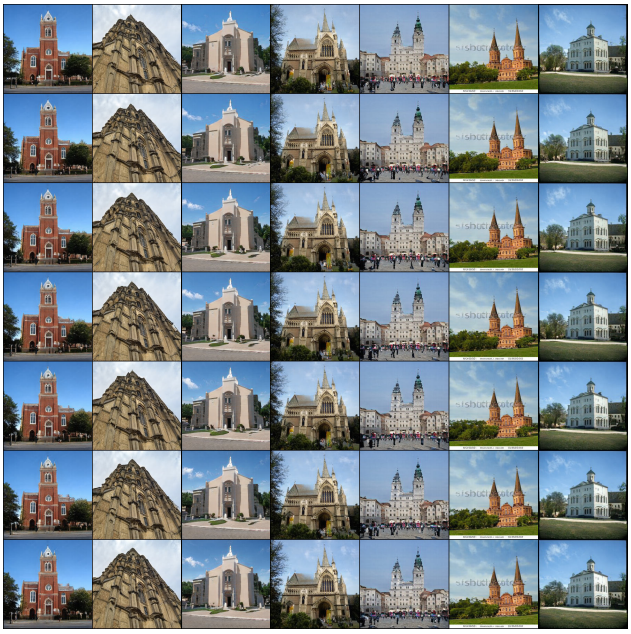

Figure 1. Exemplary latent traversal results of full-rank Householder Projector on LSUN Church [16] with StyleGAN2 [10]. Due to the large dimensionality, using the full-rank projector would split data variations among the eigenvectors. The output changes are thus imperceptible and it is unlike to inspect the concrete semantic attribute of each traversal direction.

full-rank projector seems to outperform the low-rank projector. However, as shown in Fig. 1, there are not much variations in the traversal results and it is hard to inspect the specific semantic attributes of the identified directions. In this case, the advantage of full-rank projector on PPL and PIPL might come from less meaningful variations instead of the improved latent smoothness. Setting the matrix rank to 5 and 10 leads to very competitive performance. We set the rank to 10 throughout the experiments because we empirically observed that each layer of StyleGANs has approximately 10 semantic concepts. Nonetheless, the readers are encouraged to set different ranks for other datasets and architectures if more semantics are observed.

E. More Visualizations and Discussions

E.1. Semantic Unambiguity

Fig. 2 displays some examples of semantics unambiguity. The interpretable directions identified by our Householder Projector are unambiguous: different samples would have consistent semantic attribute changes when the latent code is moved by the discovered directions.

E.2. Semantic Hierarchy

Fig. 3 shows the layer hierarchy of different semantics on FFHQ [11]. The shallow layers mainly focus on some geometric changes of the input images. Then the middle layers proceed to manipulate local details such as mouths and eyes. Finally, the deep layers target the global style and appearance of the images. Overall, the semantics hierarchy meets the same trend of StyleGANs. This indicates that our Householder Projector does not modify the semantics hierarchy of pre-trained models but tunes the model to mine more disentangled semantic concepts.

E.3. Semantic Diversity

Fig. 4 displays some more semantic attributes discovered on the used datasets. Different from the paper, here we exhibit more style semantics, *i.e.,* the global appearance changes in the high-level layers of StyleGANs. Specific to each datatset, the style semantics correspond to different global variations that frequently occur in the datasets. For example, the style variations in MetFaces [7] are mainly different painting and colorization styles, and the style variations in FFHQ [11] mainly concern global color contrast, image sharpness, and different color temperatures.

E.4. Visual Comparison on Other Datasets

Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9 present the exemplary attribute comparison across all used datasets. The results are consistent with the visualizations in the paper. Our Householder Projector is able to identify more disentangled semantic attributes and gives users more precise control of the image attributes in the generation process.

E.5. Comparison with EigenGAN

EigenGAN [4] is a small-scale GAN architecture that progressively injects orthogonal subspace into each layer of the generator to achieve disentanglement. Similar with HP [12] and OrJaR [15], the *soft* orthogonality regularization is also used in EigenGAN to preserve the approximate orthogonality. Fig. 10 compares some semantics learned by our method and EigenGAN [4] on FFHQ [11]. Our method can discover more precise image attributes.

E.6. Generated Samples

Fig. 11 displays some samples randomly generated by our method across datasets. The image quality of the original StyleGANs [10, 8] is maintained by our Householder Projector.

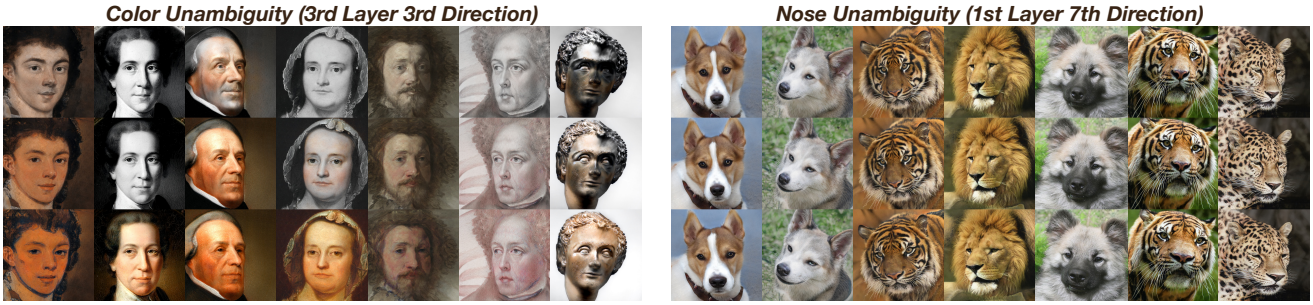

Figure 2. Illustration of semantic unambiguity on MetFaces [7] and AFHQ [1] based on StyleGAN3 [8] equipped with our Householder Projector. The discovered interpretable directions are semantically consistent among different samples.

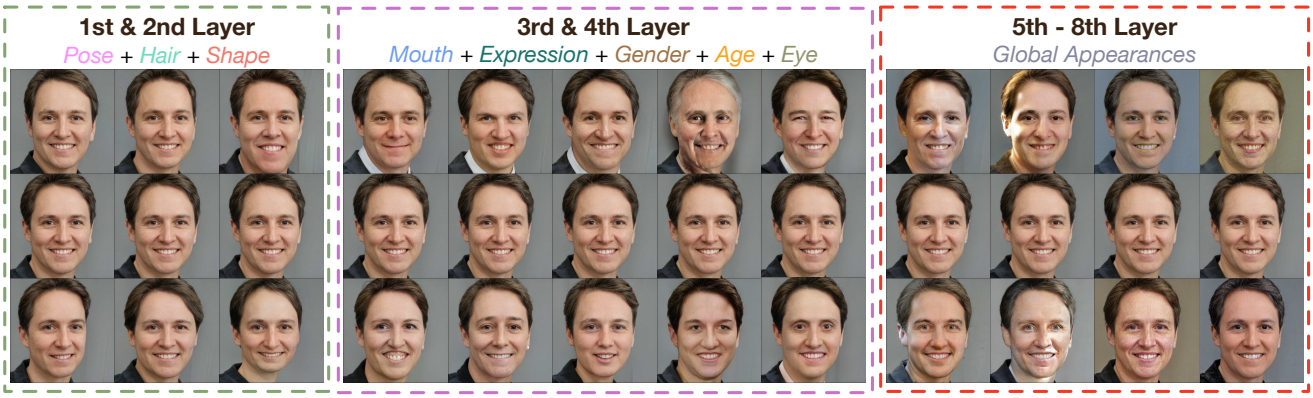

Figure 3. The layer hierarchy of semantic attributes identified by our Householder Projector based on FFHQ [11] with StyleGAN2 [10].

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2, 5, 9

[2] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, 2020. 3

[3] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *ECCV*, 2022. 2, 10

[4] Zhenliang He, Meina Kan, and Shiguang Shan. Eigengan: Layer-wise eigen-learning for gans. In *ICCV*, 2021. 4, 11

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 2

[6] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021. 3

[7] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 2020. 2, 4, 5

[8] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 2021. 4, 5, 12

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3, 4, 5, 12

[11] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 2, 3, 4, 5, 11

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

[13] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2022. 3

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

[15] 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. 4

[16] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2, 3, 4, 7, 8

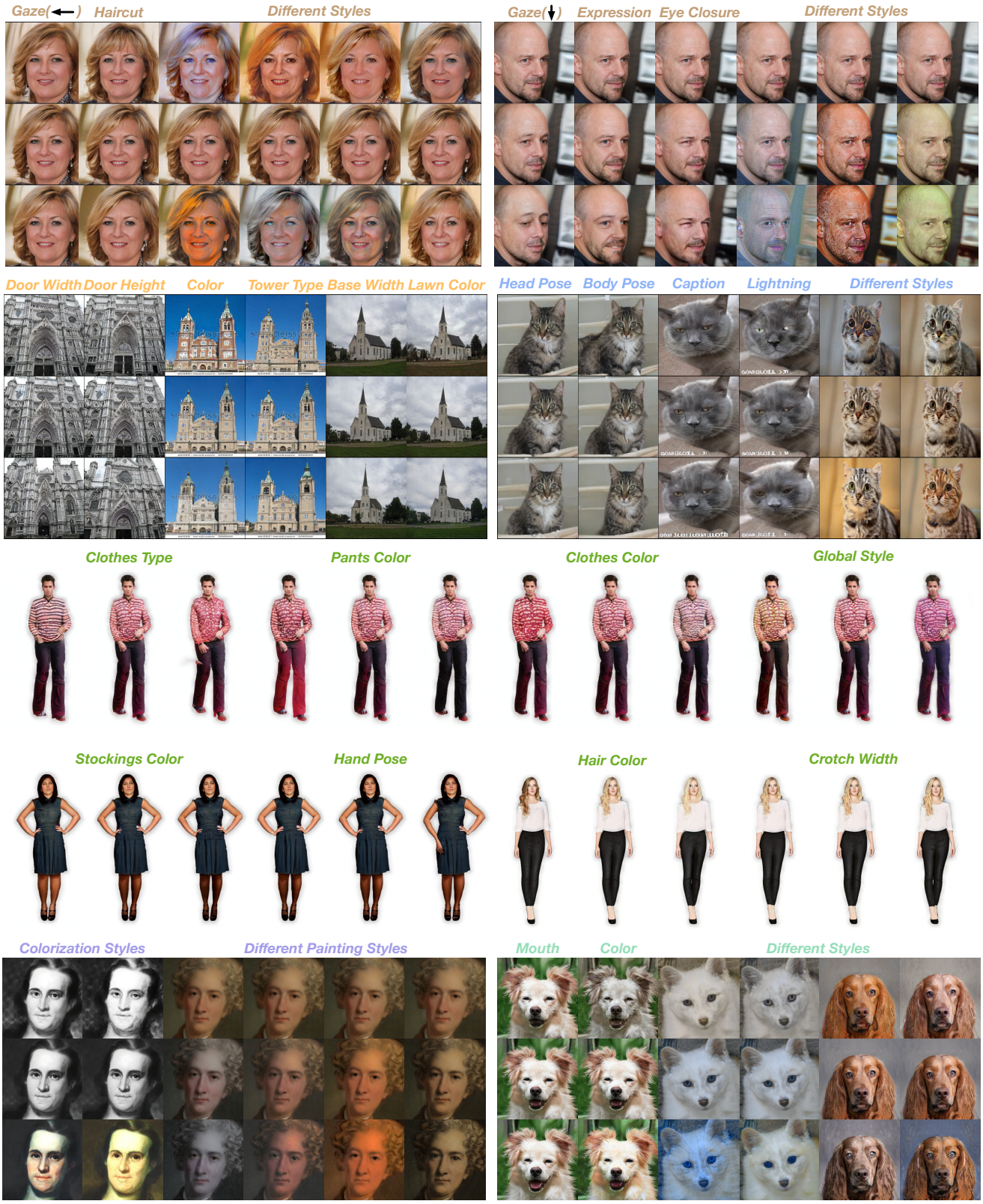

Figure 4. Gallery of more semantic attributes discovered on the used datasets. Here we display more style-related semantics.

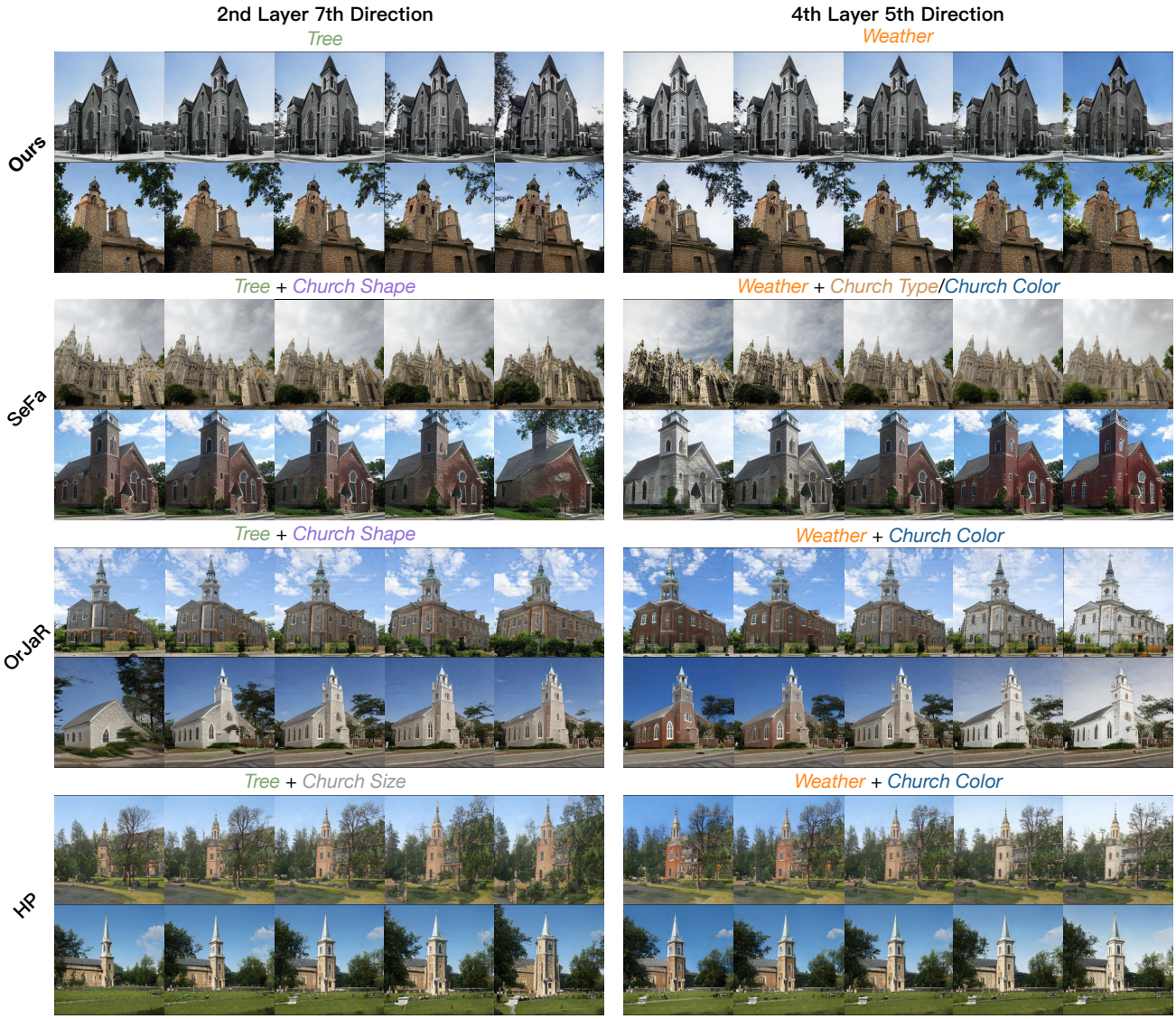

Figure 5. Exemplary latent traversal comparison of two attributes on LSUN Church [16].

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[18] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, 2017. 3

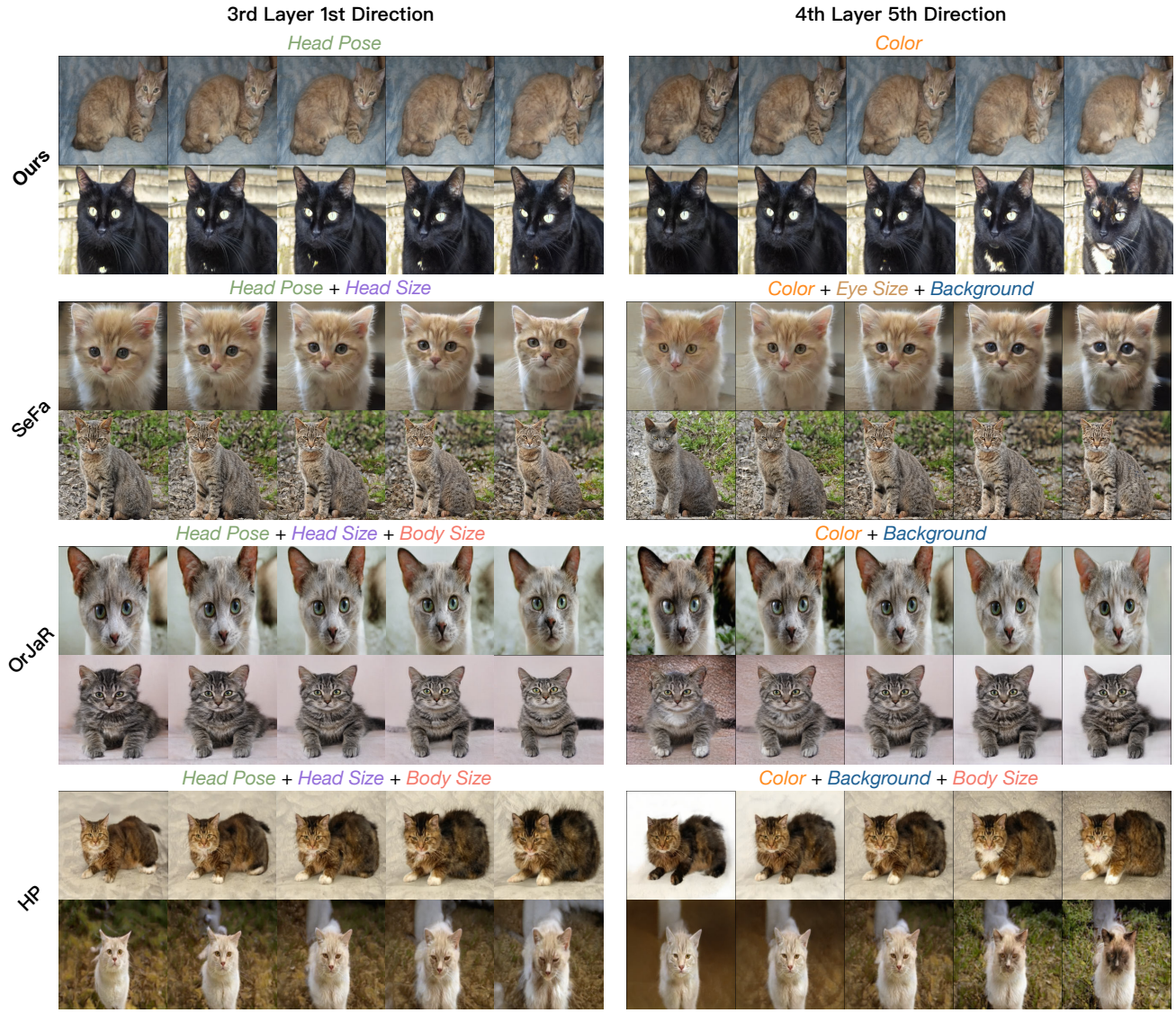

Figure 6. Exemplary latent traversal comparison of two attributes on LSUN Cat [16].

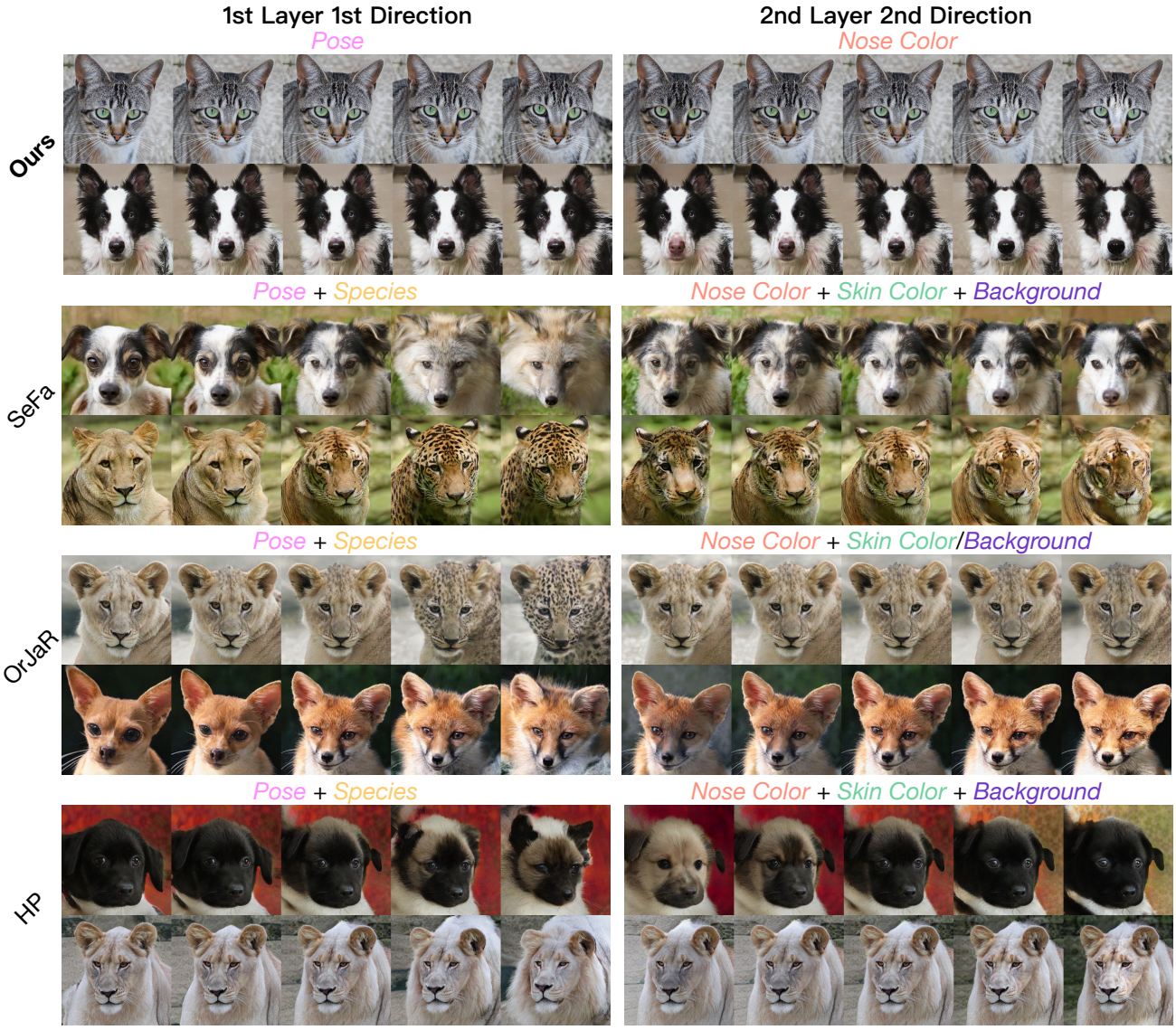

Figure 7. Exemplary latent traversal comparison of two attributes on AFHQv2 [1].

Figure 8. Exemplary latent traversal comparison of three attributes on SHHQ [3].

Figure 9. Exemplary latent traversal comparison of two attributes on FFHQ [11].

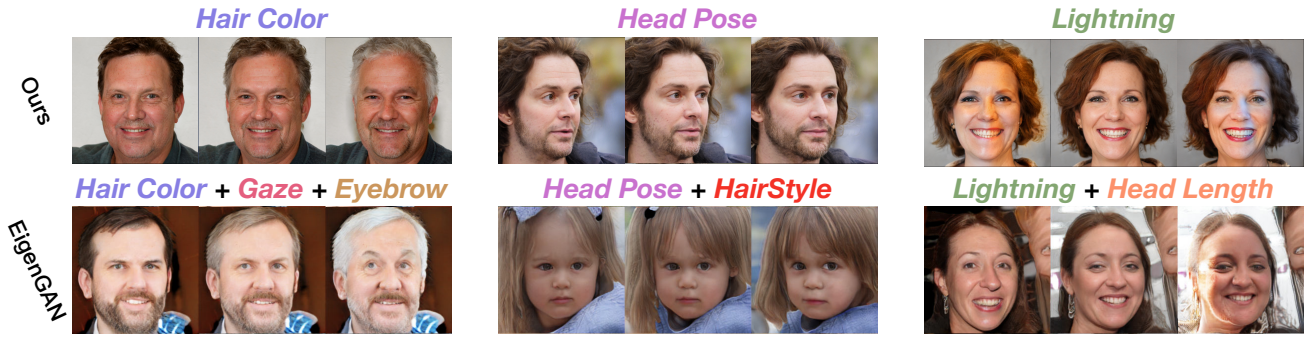

Figure 10. Comparison against EigenGAN [4] on some learned attributes with FFHQ [11].

Figure 11. Random samples generated by StyleGANs [10, 8] equipped with our Householder Projector. Our method does not harm the original quality of generate images.