Householder Projector for Unsupervised Latent Semantics Discovery

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

Abstract

Generative Adversarial Networks (GANs), especially the recent style-based generators (StyleGANs), have versatile semantics in the structured latent space. Latent semantics discovery methods emerge to move around the latent code such that only one factor varies during the traversal. Recently, an unsupervised method proposed a promising direction to directly use the eigenvectors of the projection matrix that maps latent codes to features as the interpretable directions. However, one overlooked fact is that the projection matrix is non-orthogonal and the number of eigenvectors is too large. The non-orthogonality would entangle semantic attributes in the top few eigenvectors, and the large dimensionality might result in meaningless variations among the directions even if the matrix is orthogonal. To avoid these issues, we propose Householder Projector, a flexible and general low-rank orthogonal matrix representation based on Householder transformations, to parameterize the projection matrix. The orthogonality guarantees that the eigenvectors correspond to disentangled interpretable semantics, while the low-rank property encourages that each identified direction has meaningful variations. We integrate our projector into pre-trained StyleGAN2/StyleGAN3 and evaluate the models on several benchmarks. Within only 1% of the original training steps for fine-tuning, our projector helps StyleGANs to discover more disentangled and precise semantic attributes without sacrificing image fidelity. Code is publicly available via https://github.com/KingJamesSong/HouseholderGAN.

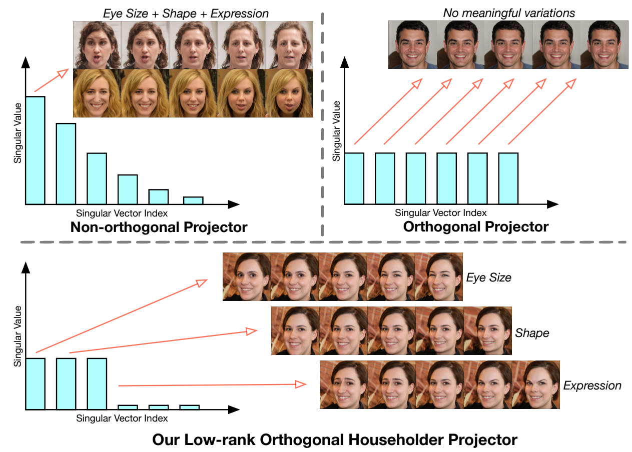

Figure 1: Motivation of our proposed Householder Projector. ***Here "Projector" denotes the projection matrix that maps latent codes to features,* i.e., *the modulation weight of StyleGANs.*** (*Top Left*) The singular value imbalance of the non-orthogonal projector would entangle multiple semantics in the top interpretable directions. (*Top Right*) Due to the large dimensionality of the projector, directly enforcing vanilla orthogonality would spread the data variations among all the eigenvectors, leading to imperceptible and meaningless traversal. ***(Bottom)*** Our Householder Projector equips the projection matrix with low-rank orthogonal properties, which simultaneously disentangles semantics into multiple equally-important eigenvectors and guarantees that each direction could correspond to semantically-meaningful variations.

1. Introduction

Generative Adversarial Networks (GANs) [15] and the recent style-based generative models (StyleGANs) [27, 28, 26] in particular have become the leading paradigm of generative modeling in the vision domain. The latent spaces of StyleGANs are known to embed rich and hierarchical semantics [14, 21]; moving the latent code in certain directions could trigger meaningful variations in the output images. Therefore, latent semantics discovery methods

The first two authors contribute equally; Wei Wang is the corresponding author.

emerge to identify such interpretable directions that each variation factor is disentangled and the generation process can be precisely controled [14, 47, 42, 61, 48, 1, 58, 37, 8].

Among the recent explorations of unsupervised interpretable semantics discovery methods [56, 48, 57, 41], SeFa [48] pointed out a promising direction to discover semantically meaningful concepts by computing the eigenvector of the projector. Here we refer to the projection matrix that maps latent codes to features as the projector. The key observation is that using the eigenvectors of the projector for latent perturbation would maximize the data variations. Such identified eigenvectors/directions would correspond to meaningful semantic concepts. However, as shown in Fig. 1 top left, this is likely to cause semantics entangled in the top few eigenvectors. This phenomenon stems from the fact that the variation caused by an eigenvector is actually determined by the associated eigenvalue. Due to the imbalanced eigenvalue distribution, the discovered directions are not equally-important, and the top few ones would simultaneously manipulate multiple attributes. This eigenvalue discrepancy can be mitigated by enforcing orthogonality constraint to the projector. Nonetheless, since standard orthogonal matrices have as many equally-important eigenvectors as the dimensionality, there would not be enough semantics to mine in practice when the method scales to large models such as StyleGANs whose projector dimension is 1024 or 512. Consequently, as an accompanying limitation, the data variations would be split among all the eigenvectors and none of them could produce meaningful output changes (see Fig. 1 top right).

To resolve the above issues, we propose Householder Projector, a low-rank orthogonal matrix representation based on Householder transformations, to parameterize the projection matrix that maps latent codes to features. The projector is first decomposed to its SVD form ($\mathbf{USV}^T$). Next, the orthogonal singular vectors $\mathbf{U}$ and $\mathbf{V}$ are represented by a series of Householder reflectors, respectively. Thanks to the normalization of Householder reflections, the orthogonality is also preserved during backpropagation. For the singular value $\mathbf{S}$, we explicitly set it as a low-rank identity matrix (*i.e.,* $\mathbf{S}$=$\mathrm{diag}(1,\dots,1,0,\dots,0)$) whose rank defines exactly the number of semantic concepts. As shown in Fig. 1 bottom, the low-rank property guarantees that the identified directions would cause meaningful variations, while the orthogonality encourages that each semantic attribute is disentangled from the others. Moreover, a proper initialization scheme is proposed to leverage the statistics of pre-trained weights, and an acceleration technique is applied to speed up the computation. We also propose a metric dedicated to measuring the smoothness of latent space to interpretable directions based perturbations. Our Householder Projector is integrated into pre-trained StyleGANs [28, 26] at multiple different layers to mine the diverse and hierarchical semantics. Since our projector inevitably changes the pre-trained parameters, the modified models incorporated with our projector are fine-tuned for limited steps to maintain the original image fidelity. Both quantitative and qualitative results on several widely used benchmarks [29, 63, 10, 25, 13] show that **within marginal fine-tuning steps (1% of the training steps), our Householder Projector improves the latent semantics discovery of StyleGANs to have more precise attribute control while not impairing the quality of generated images.**

The key results and main contributions are as follows:

- We propose Householder Projector, a flexible and general low-rank orthogonal matrix representation based on Householder transformations, to parameterize the projector of generative models for latent semantics discovery.
- Our Householder Projector can be easily integrated into pre-trained GAN models. With the acceleration technique, it can be efficiently fine-tuned for very limited additional steps, which paves the way for applying latent semantics discovery and orthogonality techniques to large-scale generative models such as StyleGANs.
- Extensive experiments on two popular backbones (StyleGAN2 [28] and StyleGAN3 [26]) and six benchmarks (FFHQ [29], LSUN Church and Cat [63], MetFaces [25], AFHQ [10], and SHHQ [13]) demonstrate that within marginal extra fine-tuning steps (1% of the original training steps), our method could both achieve precise attribute control and preserve the original image quality.

2. Related Work

Generative Adversarial Networks. In the past few years, GAN-based generative models [15] have achieved remarkable progress in high-fidelity image synthesis [43, 24, 5, 27, 28, 26, 25, 39, 6, 16, 50, 46]. The generation process usually takes the following procedure: a randomly-sampled latent code is fed into the generator through a projection step, and then the generator outputs realistic-looking images. Recently, the style-based generators [27, 28, 26] that gradually absorb layer-wise latent style codes are becoming the *de facto* GAN backbones. StyleGAN2 [28] improves the original StyleGAN [27] by redesigning generator normalization and training techniques, and StyleGAN3 [26] further explores some equivariance properties. We use the popular StyleGAN2 and StyleGAN3 as our backbones in this paper.
Latent Semantics Discovery. Recently, lots of methods explore disentangling the latent space to achieve image editing by moving the latent code in the identified interpretable directions [7, 3, 21, 47, 17, 56, 67, 48, 18, 41, 54, 57, 55, 30, 55, 31, 60, 22, 44, 9, 53, 51]. Supervised methods rely on human annotations (*i.e.*, segmentation masks, attribute categories, 3D priors, and text descriptions) to de-

fine the semantic labels [14, 47, 61, 21, 42, 33, 8, 12, 49, 40, 34, 62, 59, 23]. Here we mainly highlight some relevant unsupervised approaches that are free of annotations. Voynov *et al.* [56] proposed to jointly learn a set of directions and a classifier such that the interpretable directions can be recognized by the classifier. More recently, Peebles *et al.* [41] and Wei *et al.* [57] proposed to add orthogonal Hessian and Jacobian regularization to encourage disentangled representations, respectively. Song *et al.* [51] proposed to use wave equations to model the spatiotemporal dynamic non-linear latent traversals in generative models. SeFa [48] showed that the eigenvectors of the projector after the latent code could maximize the data variations and proposed to directly use them as the interpretable directions. However, due to the imbalanced eigenvalues, the image attributes would be entangled in the top few eigenvectors. Our proposed Householder Projector solved this issue by parameterizing the projector to a low-rank but orthogonal matrix. Notice that the used orthogonality technique in [56] is different from ours. In [56], the authors use matrix exponential $\exp(\mathbf{A}-\mathbf{A}^T)$ to generate skew orthogonal matrices where the skew-symmetry could limit the representation power. Also, the technique cannot parameterize given matrices and cannot explicitly control the rank. Our Householder representation is more general and flexible, allowing for controllable rank and parameterization of given matrices. Compared with the *soft* orthogonality regularization used in [41, 57, 18], the *hard* orthogonality of our method helps the model to learn more uncorrelated attributes within less fine-tuning steps.

In contrast to global editing approaches discussed above, some methods can perform local image editing in a *post hoc* way: they first define or learn a segment of regions of interests, and then manipulate the masked intermediate features for local editing [2, 11, 65, 66, 38]. Empowered by the precise control of attributes, our method can also achieve competitive performance in local image editing (see Sec. 4.2).

3. Householder Projector

This section starts with the preliminary introduction of the closed-form latent semantics discovery. Next, we analyze its inherent limitation on entangled semantics and then illustrate our proposed Householder Projector in detail.

3.1. Preliminary: Closed-form Latent Discovery

Previous latent semantics discovery approaches [14, 56, 17, 47, 48] consider the GAN manipulation as $\texttt{edit}(G(\mathbf{z}))=G(\mathbf{z}+\alpha\mathbf{n})$ where $G(\cdot)$ represents the generator, $\mathbf{z}\in\mathbb{R}^d$ denotes the latent code of dimension d, $\mathbf{n}\in\mathbb{R}^d$ is the identified semantically meaningful direction, and α represents the perturbation strength. If one views the GAN as a multi-step projection function, the first projection can be expressed as:

$$G_1(\mathbf{z}+\alpha\mathbf{n}) = \mathbf{A}\mathbf{z}+\mathbf{b}+\alpha\mathbf{A}\mathbf{n}, \tag{1}$$

where $\mathbf{A}$ and $\mathbf{b}$ denote the weight and bias of the projection step (*e.g.,* convolution or linear transform), respectively. As can be seen from eq. (1), the resultant manipulation depends on the term $\alpha\mathbf{A}\mathbf{n}$. Intuitively, an interpretable direction $\mathbf{n}$ should cause large variations of $G_1(\mathbf{z}{+}\alpha\mathbf{n})$, which is equivalent to maximizing the impact of $\alpha\mathbf{A}\mathbf{n}$. Motivated by this observation, SeFa [48] proposed to consider the formulation as the following constrained optimization problem:

$$\mathbf{n}^\star = \arg\max ||\mathbf{A}\mathbf{n}||_2^2 \; s.t. \; \mathbf{n}^T\mathbf{n}=1, \tag{2}$$

where the constraint $\mathbf{n}^T\mathbf{n}{=}1$ is set to satisfy vector orthogonality, and $||\cdot||$ denotes the l_2 norm. Introducing a Lagrange multiplier λ leads to $2\mathbf{A}^T\mathbf{A}\mathbf{n}{-}2\lambda\mathbf{n}{=}0$. The closed-form solutions all correspond to the eigenvectors of $\mathbf{A}^T\mathbf{A}$. This presents a promising approach to identify the semantics by exploiting the projector $\mathbf{A}$ that projects latent codes. However, one fact overlooked by [48] is that the eigenvectors would cause different extents of variations due to the discrepancy of associated eigenvalues. Supposing that $\mathbf{n}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$, then we would have $||\mathbf{A}\mathbf{n}||_2^2{=}\sigma^2$ where σ is the corresponding singular value of $\mathbf{A}$. For non-orthogonal matrices, the singular values are exponentially decreasing, *i.e.,* $\sigma_1{\geq}\sigma_2{\geq}\dots{\geq}\sigma_d$. This would cause most variations captured by the first few interpretable directions. The imbalance is thus likely to make semantic attributes entangled in the top eigenvectors (see Fig. 1 top left).

3.2. Householder Low-rank Orthogonal Projector

The eigenvalue discrepancy can be eliminated by enforcing *strict* orthogonality. Orthogonal matrices have the property of identical eigenvalues, which assigns equal importance to each eigenvector. The low-rank constraint could further limit the number of semantics to mine. Therefore, we propose to use Householder representation, a flexible and general framework to parameterize matrices, to endow the projection matrix with low-rank orthogonality.

Householder Parameterization. Householder representations can parameterize any matrices by using a series of Householder reflectors to represent the orthogonal singular vectors of its Singular Value Decomposition (SVD) form. In the field of deep learning, it has been used to parameterize the transition matrix and to stabilize gradients of recurrent neural networks [36, 64, 35]. The key to the orthogonality representation relies on the following theorem:

Theorem 1 (Householder representation [20, 32])**.** *Given any square orthogonal matrix* $\mathbf{M}\in\mathbb{R}^{m\times m}$*, it can be represented by the product of Householder matrices* $\mathbf{M} = \mathbf{H}_1\mathbf{H}_2\dots\mathbf{H}_m$ *where each Householder matrix is parameterized by a vector as* $\mathbf{H}_i = \mathbf{I} - 2\frac{\mathbf{h}_i\mathbf{h}_i^T}{||\mathbf{h}_i||_2^2}$.

Let $\mathbf{U}\mathbf{S}\mathbf{V}^T$ denote the SVD of the projector $\mathbf{A}$ where $\mathbf{S}$

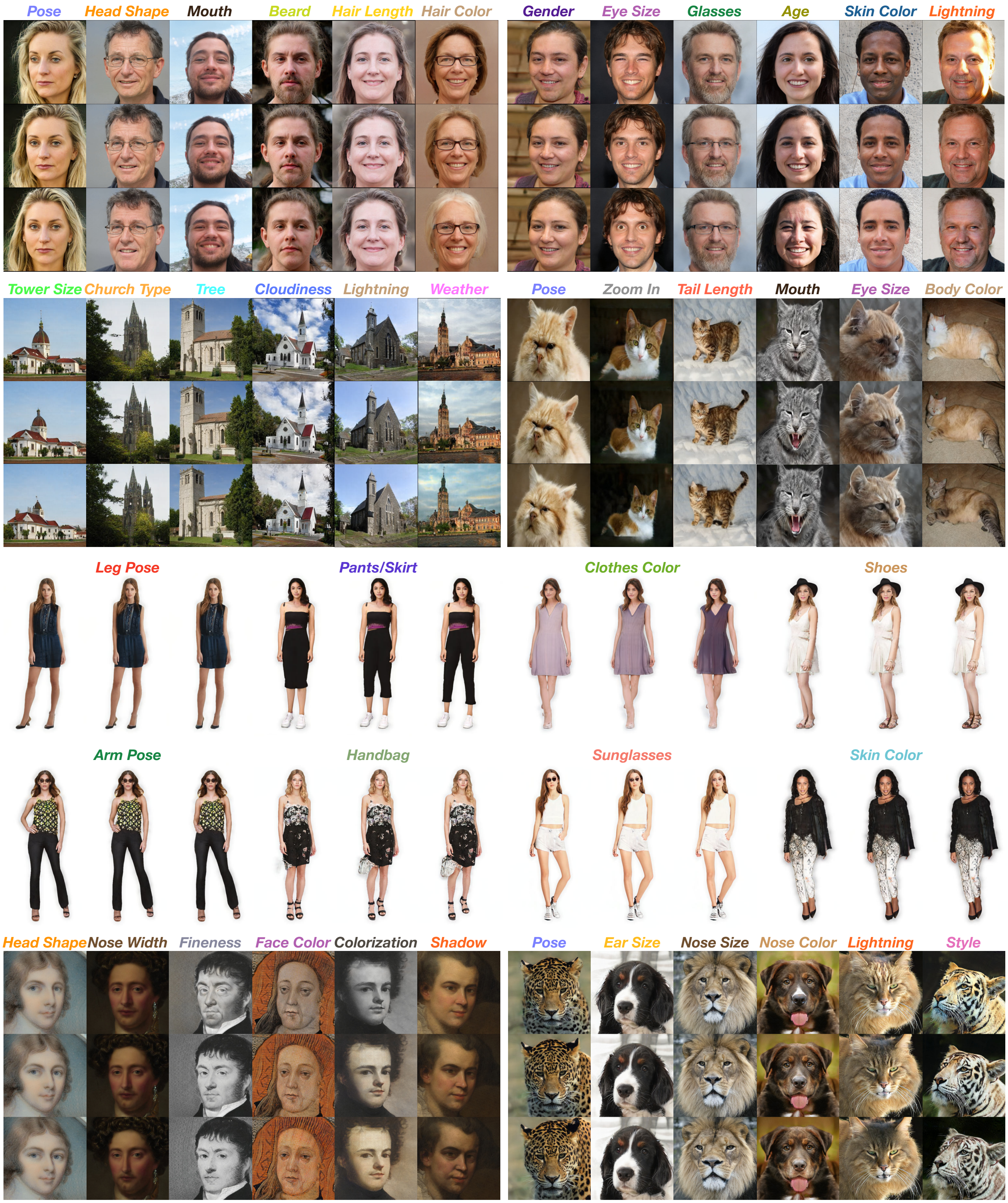

Figure 2: Gallery of some semantic attributes discovered by our Householder Projector across all used datasets (FFHQ [29] in the top row, LSUN Church [63] and LSUN Cat [63] in the 2_{nd} row, SHHQ [13] in the 3_{rd} row, and MetFaces [25] and AFHQ [10] in the bottom row). These semantic attributes are sorted from low-level layers (left) to high-level layers (right).

denotes the diagonal singular value, and $\mathbf{U}$ and $\mathbf{V}$ are left and right orthogonal singular vectors. We use the accumulation of Householder reflectors (*i.e.,* $\prod_{i=0}^{w}\left(\mathbf{I}-2\frac{\mathbf{u}_i\mathbf{u}_i^T}{||\mathbf{u}_i||_2^2}\right)$ and $\prod_{i=0}^{h}\left(\mathbf{I}-2\frac{\mathbf{v}_i\mathbf{v}_i^T}{||\mathbf{v}_i||_2^2}\right)$ where w, h denotes the width and height of $\mathbf{A}$) to parameterize $\mathbf{U}$ and $\mathbf{V}$, respectively. Notice that only $\mathbf{u}_i$ and $\mathbf{v}_i$ are the actual learnable parameters. For $\mathbf{S}$, we explicitly set it to a diagonal matrix and keep the weight fixed. The projector is thus represented by our Householder parameterizations.

Low-rank Constraint. To achieve the property of disentangled attributes, one straightforward approach is to parameterize $\mathbf{A}$ as an orthogonal matrix, *i.e.,* to set $\mathbf{S}$ to an identity matrix $\mathbf{I}$ where $\mathbf{I}_{i,j}{=}1$ for $i{=}j$ and $\mathbf{I}_{i,j}{=}0$ otherwise. This would lead to equally-important semantic attributes whose number is exactly the projector dimension. However, for large generative models such as StyleGANs, the projector dimension is typically very large (*e.g.,* 512 or 1024). It is not likely to have enough attributes to edit in practice. Setting $\mathbf{S}$ to a full-rank diagonal matrix would spread data variations among all the eigenvectors, resulting in trivial and imperceptible traversal (see Fig. 1 top right). To avoid this issue, we propose to define $\mathbf{S}$ as a low-rank identity matrix:

$$\mathbf{S} = \mathrm{diag}(\underbrace{1,\dots,1}_{N},0,\dots,0), \tag{3}$$

where N defines the rank and also the number of semantic attributes to mine. By restricting the rank of the projector, we explicitly limit the number of interpretable directions. As shown in Fig. 1 bottom, this would benefit the latent traversal for more meaningful output variations.

Orthogonality Preservation. One advantage of our Householder representation is that the orthogonality can be kept during backpropagation. Since we have the vector normalization $\frac{\mathbf{h}_i\mathbf{h}_i^T}{||\mathbf{h}_i||_2^2}$, the impact of any gradient descent step $\mathbf{h}_i{-}\eta\nabla\mathbf{h}_i$ on the orthogonality would be cancelled, *i.e.,* the updated vector remains orthogonal after normalization.

Initialization. When our method is applied to pre-trained models, the well-trained network weights could be leveraged to initialize our Householder Projector. To this end, we propose to use the nearest-orthogonal mapping [52] to project the weight matrix into its orthogonal form that has the nearest distance in the Frobenius norm (*i.e.,* $\min||\mathbf{R}-\mathbf{A}||_{\mathrm{F}}$ where $\mathbf{R}^T\mathbf{R}=\mathbf{I}$). The closed-form solution is given by $\mathbf{U}\mathbf{V}^T$ where $\mathbf{U}\mathbf{S}\mathbf{V}^T$ is the SVD of the original weight matrix $\mathbf{A}$. Next, we decompose $\mathbf{U}$ and $\mathbf{V}$ into their Householder reflectors and use them to initialize our projector. Such an initialization scheme leverages the statistics of the original weight matrix, which might give our projector a good starting point and improve the performance (see the ablation study of Sec. D.3 in the supplementary).

Acceleration. The accumulated product of elementary Householder matrices can be accelerated via the theorem:

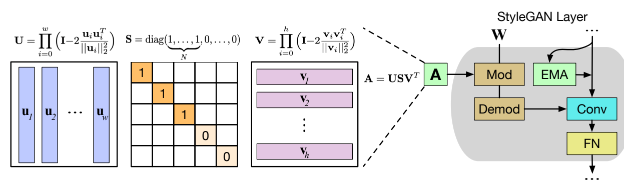

Figure 3: Illustration on how our Householder Projector represents the modulation weight $\mathbf{A}$ of StyleGANs. Here "Demod", "EMA", and "FN" denote Demodulation, Exponential Moving Average, and Filtered Non-linearities, respectively. The projector is parameterized by its SVD form where $\mathbf{U}$ and $\mathbf{V}$ are represented by the accumulation of Householder reflectors, and $\mathbf{S}$ is set to a low-rank identity matrix. Our projector is applied at multiple different layers of StyleGANs to explore the diverse and hierarchical semantics. The actual learnable parameters are $\mathbf{u}_i$ and $\mathbf{v}_i$.

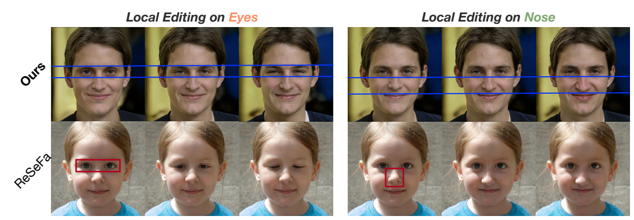

Figure 4: Comparison with ReSeFa [66]. The blue lines indicate the specific regions changed by our method, and the red box indicates the region of interest that is needed as input to ReSeFa [66].

Theorem 2 (Compact WY representation [4])**.** *For any accumulation of m Householder matrices $\mathbf{H}_1 \dots \mathbf{H}_m$, there exists $\mathbf{W},\mathbf{Y}{\in}\mathrm{R}^{d\times m}$ such that $\mathbf{I}{-}2\mathbf{W}\mathbf{Y}^T{=}\mathbf{H}_1\dots\mathbf{H}_m$ where computing $\mathbf{W}$ and $\mathbf{Y}$ takes $O(dm^2)$ time and m sequential Householder multiplications.*

This theorem indicates the possibility to repeatedly split the accumulation $\mathbf{H}_1\mathbf{H}_2\dots\mathbf{H}_m$ into multiple subsequences until irreducible. Then these sub-sequences can be computed in parallel and gradually merged. As revealed in the ablation of Sec. D.3 in the supplementary, this technique could fully exploit the parallel computational power of GPUs and greatly speed up the aggregation, particularly in our case where the projector dimension is large.

Implementation in StyleGANs. Fig. 3 depicts how our Householder Projector modifies the original StyleGAN architectures. The projector used in the weight modulation module is represented by our proposed projector. The weight matrix is thus endowed with low-rank orthogonal properties. We insert the proposed projector at every layer of StyleGAN2 and every four layers of StyleGAN3. Since StyleGAN3 has 15 intermediate layers, the adjacent layers have very similar and even repeated semantics. Therefore,

	Identity	Pose	Age	Gender	Glasses	Smile
Identity	**0.65**	0.24	0.03	0.04	0.01	0.03
Pose	0.11	**0.57**	0.04	0.04	0.11	0.13
Age	0.02	0.05	**0.67**	0.03	0.19	0.03
Gender	0.03	0.32	0.02	**0.52**	0.10	0.00
Glasses	0.01	0.08	0.00	0.02	**0.88**	0.01
Smile	0.02	0.01	0.01	0.00	0.01	**0.95**

(a) Our method

	Identity	Pose	Age	Gender	Glasses	Smile
Identity	**0.56**	0.06	0.05	0.02	0.05	0.26
Pose	0.44	**0.48**	0.05	0.01	0.01	0.00
Age	**0.43**	0.27	0.22	0.05	0.03	0.01
Gender	0.33	0.18	0.04	**0.40**	0.04	0.02
Glasses	**0.27**	0.14	0.04	0.10	0.23	0.22
Smile	0.20	0.08	0.09	0.00	0.03	**0.61**

(b) SeFa

	Identity	Pose	Age	Gender	Glasses	Smile
Identity	**0.51**	0.27	0.04	0.05	0.03	0.10
Pose	**0.39**	0.35	0.05	0.02	0.07	0.11
Age	0.26	0.14	**0.32**	0.10	0.05	0.12
Gender	0.20	0.06	**0.34**	0.29	0.04	0.07
Glasses	0.18	0.06	0.12	0.07	**0.42**	0.14
Smile	0.13	0.05	0.07	0.02	0.03	**0.70**

(c) OrJaR

	Identity	Pose	Age	Gender	Glasses	Smile
Identity	**0.54**	0.25	0.04	0.06	0.02	0.08
Pose	**0.42**	0.30	0.08	0.04	0.05	0.12
Age	0.21	0.15	**0.28**	0.11	0.08	0.17
Gender	**0.27**	0.07	0.23	0.25	0.04	0.13
Glasses	0.15	0.12	0.16	0.04	**0.37**	0.16
Smile	0.11	0.09	0.12	0.01	0.01	**0.66**

(d) HP

Table 1: The l_1 normalized attribute correlations based on $2K$ same samples generated by GAN inversion.

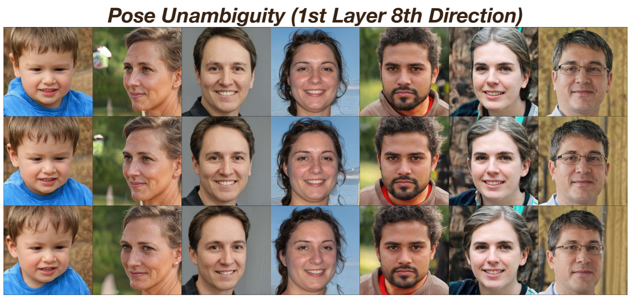

Figure 5: Interpretable directions identified by our method are semantically consistent among different samples.

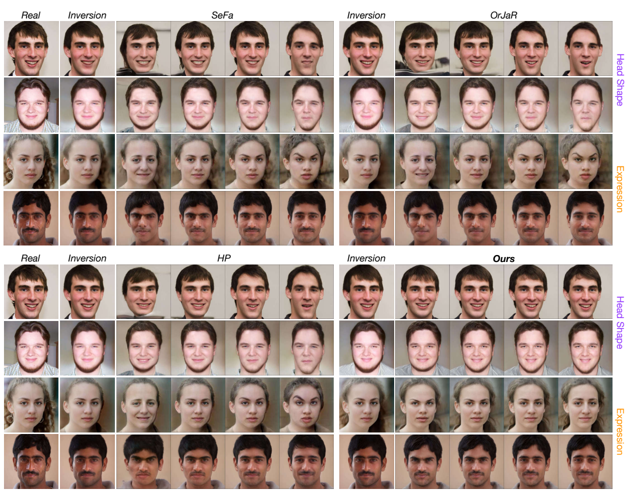

Figure 6: Qualitative comparisons of the same samples.

we choose to integrate our projector every few layers to obtain interpretable directions of different semantics.

4. Experiments

In this section, we first introduce the experimental setup, followed by the quantitative and qualitative evaluation. ***We defer the full ablation studies to Sec. D of supplementary.***

4.1. Setup

Models. We evaluate our Householder Projector on StyleGAN2 [28] and StyleGAN3 [26], *i.e.,* the challenging *state-of-the-art* GAN backbones in the field of computer vision.

Datasets and Baselines. For StyleGAN2, we conduct experiments on FFHQ [29], LSUN Church [63], and LSUN Cat [63]. The experiments of StyleGAN3 are performed on SHHQ [13], MetFaces [25] and AFHQv2 [10]. We mainly compare our method with representative unsupervised latent semantics discovery approaches, including SeFa [48], Orthogonal Jacobian Regularization (OrJaR) [57], and Hessian Penalty (HP) [41]. SeFa [48] can be directly applied to pre-trained models, while OrJaR and HP need extra fine-tuning or training from scratch due to the regularization.

Metrics. We conduct quantitative evaluation using (1) **Fréchet Inception Distance (FID)** [19], (2) **Perceptual Path Length (PPL)** [27], (3) **Perceptual Interpretable Path Length (PIPL)**, and (4) **Face Attribute Correlation**. FID aims to measure the image quality and diversity by computing the distance between the real and fake distributions, and PPL is designed to assess the perceptual smoothness of the latent space where the smoothness can reflect the disentanglement ability. Our proposed PIPL is a natural modification of PPL by adapting the latent manipulation from random interpolation-based perturbations to vector-based perturbations using interpretable directions. Compared with PPL, our PIPL can better measure the latent space smoothness when the latent code is perturbed along with specific vectors, which suits more vector-based semantic discovery methods like SeFa [48] and ours. Furthermore, for StyleGAN2 trained on FFHQ, we rely on pre-trained face attribute estimators to compute the correlation coefficient between the traversal steps and the face attributes. Besides the above four metrics, we also assess the disentanglement through visual observation. We defer the details of used datasets and metrics to Sec. C of the supplementary.

Implementation Details. We adopt the widely used Pytorch implementation of StyleGAN2[1] and convert the official TensorFlow pre-trained models into PyTorch for FFHQ [29], LSUN Church [63], and LSUN Cat [63]. For StyleGAN3, we use the official code and pre-trained models of AFHQv2 [10] and MetFaces [25]. As for SHHQ [13], we also use the official pre-trained model[2]. Following the original optimization strategy, we finetune all the parameters of the pre-trained generator and discriminator within 1% of the total training steps (kimgs for StyleGAN3). For instance, the fine-tuning process takes $5K$ steps for StyleGAN2 with FFHQ and 250 kimgs for StyleGAN3 with AFHQ. The fine-tuning time is thus very limited due to the small number

[1] https://github.com/rosinality/stylegan2-pytorch
[2] https://github.com/stylegan-human/StyleGAN-Human

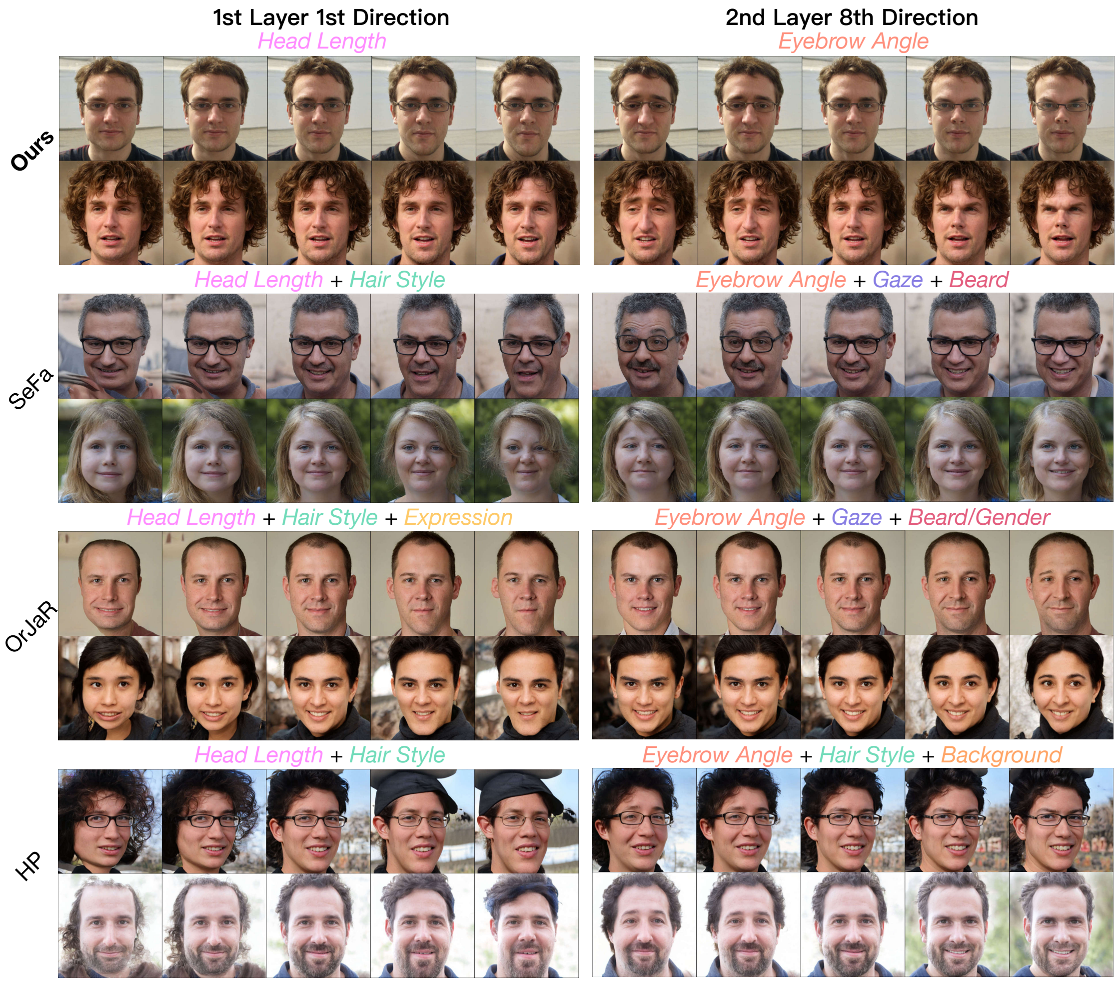

Figure 7: Exemplary qualitative comparison of two different semantics on FFHQ [29] with StyleGAN2 [28]. Our Householder Projector can precisely control the image attributes without changing the face identity. The direction index denotes the index of eigenvectors.

of training steps. To give concrete examples, fine-tuning models on FFHQ and AFHQ takes 1.5 and 2.5 hours, respectively. The rank N of $\mathbf{S}$ is set to 10 for all experiments based on our empirical observation of the number of semantics of StyleGANs. Our editing strength is set the same as SeFa. We use 4 RTX A6000 GPUs for the training. For the comparison fairness, the baseline methods OrJaR [57] and HP [41] are also fine-tuned with the same steps.

4.2. Qualitative Evaluation and Discussion

Diverse and Precise Attributes. Fig. 2 exhibits some semantic attributes discovered by our Householder Projector on all the datasets. Our method mines a diverse set of disentangled semantics, enabling a wide range of attribute manipulation. Take as an example the first row of attributes discovered on FFHQ [29]. The left columns present diverse high-level semantic concepts such as "Pose" and "Shape", while the right columns show low-level attributes such as "Color" and "Lighting". This hierarchy also meets the same trend of original StyleGANs. The manipulation of the diverse and highly-disentangled semantics would give users more precise control on the image generation process.

Datasets	Methods	FID (↓)	PPL (↓)	PIPL (↓)
FFHQ [29] 1024×1024	SeFa [48]	4.48	1579.76	0.227
	OrJaR [57]	4.51	987.80	0.204
	HP [41]	4.66	993.17	0.207
	Ours	**4.40**	**966.23**	**0.141**
LSUN Church [63] 256×256	SeFa [48]	4.61	530.68	0.069
	OrJaR [57]	3.77	474.77	0.065
	HP [41]	3.78	486.93	0.058
	Ours	**3.72**	**457.52**	**0.030**
LSUN Cat [63] 256×256	SeFa [48]	8.37	722.24	0.141
	OrJaR [57]	8.39	562.98	0.134
	HP [41]	**8.31**	573.71	0.136
	Ours	8.46	**526.26**	**0.057**

Table 2: Evaluation results on StyleGAN2.

Semantic Unambiguity. Importantly, our interpretable di-

Figure 8: Exemplary visual comparison of three different semantics on SHHQ [13] with StyleGAN3 [26]. Our method is able to mine more disentangled interpretable directions and have more precise control on the attributes. The direction index denotes the index of eigenvectors.

rections are unambiguous to different samples. As shown in Fig. 5, the image variations manipulated by our discovered directions all correspond to the same semantic attribute, *i.e.*, the head pose. The other non-target attributes are untouched, such as the background and face identity.

Comparison against Other Methods. Fig. 7 and Fig. 8 compare the latent traversal of some directions against other baselines on FFHQ [29] and SHHQ [13], respectively. All the methods can discover similar attributes in the same layer, such as the head length in Fig. 7 left. However, the baselines suffer from entangled semantics and the other attributes vary during the traversal, such as hairstyle for SeFA [48] and OrJaR [57], and expression for HP [41]. By contrast, our method is able to discover more precise semantics and preserve other non-target attributes.

Comparison of Same Samples. To better compare the qualitative disentanglement performance, we further use a GAN inversion technique (PTI [45]) to create nearly the same images from FFHQ for different methods. Fig. 6 displays some qualitative comparisons. For the same samples, our method still has more precise attribute control.

Local Editing Applications. With precise attribute control, our method is even able to edit local regions by simply performing latent traversal. Fig. 4 displays such a use case. Our method achieves very competitive performance against local editing methods such as the recent ReSeFa [66]. In addition, our method is free of extra bounding box as input.

Please refer to Sec. E of the supplementary for more visualizations about comparisons on other datasets, and semantic diversity/unambiguity/hierarchy.

4.3. Quantitative Evaluation

StyleGAN2 Results. Table 2 presents the quantitative evaluation results on FFHQ [29], LSUN Church [63] and LSUN Cat [63] datasets. Our proposed Householder Projector outperforms other baselines in terms of both PPL and PIPL. This demonstrates that our method has a smoother and more structured latent space, which corresponds to more disentangled representations. In particular, our approach surpasses SeFa [48] by a large margin, which indicates the benefit of enforcing low-rank orthogonality to the projection matrix. Compared with the *soft* orthogonality regularization used in OrJaR [57] and HP [41], the *hard* orthogonality of our projector also has an advantage in latent smoothness due to the strict orthogonality preservation and the additional low-rank constraint. Moreover, our FID score is also very competitive, implying that our method could simultaneously improve the disentanglement performance while keeping the quality of generated images unharmed. For the attribute correlation, we also use GAN inversion to create a dataset of $2K$ identical images for each method. Table 1 compares the correlation results on FFHQ. Our method outperforms other unsupervised baselines and preserves the attribute well in particular.

Datasets	Methods	FID (↓)	PPL (↓)	PIPL (↓)
MetFaces [25] 1024×1024	SeFa [48]	**15.33**	5626.31	2.991
	OrJaR [57]	17.55	5754.44	3.700
	HP [41]	17.32	5323.27	3.465
	Ours	16.89	**4192.52**	**0.099**
AFHQv2 [10] 512×512	SeFa [48]	**4.40**	2193.74	0.470
	OrJaR [57]	5.45	2103.47	0.440
	HP [41]	5.33	2198.46	0.463
	Ours	4.98	**2052.38**	**0.070**
SHHQ [13] 512×256	SeFa [48]	**2.54**	1621.07	0.370
	OrJaR [57]	4.78	1614.56	0.245
	HP [41]	5.38	1648.74	0.216
	Ours	4.17	**1549.01**	**0.119**

Table 3: Evaluation results on StyleGAN3.

StyleGAN3 Results. Table 3 compares the performance of our method against other baselines on MetFaces [25], AFHQv2 [10], and SHHQ [13] datasets. The results are very coherent with those on StyleGAN2: our Householder Projector improves the latent space smoothness without harming the image fidelity. Our method as well as the two baselines that involve fine-tuning have slightly worse FID than the original StyleGAN3. This might stem from the fact that due to the limited computational resources, our used batch size (16) is actually smaller than the original setting of StyleGAN3 (32). As revealed in the ablation study of Sec. D.2 of the supplementary, the hyper-parameter batch size has a substantial impact on FID. We expect that increasing the batch size would further boost the image quality of our method and lead to a more competitive FID score.

5. Conclusion

This paper proposes a general and flexible low-rank orthogonal matrix representation coined as Householder Projector for unsupervised latent semantics discovery of generative models. The proposed method endows the projection matrix with low-rank orthogonality. This superiority helps pre-trained GANs to achieve precise and diverse semantics control within limited fine-tuning steps. Extensive experiments of StyleGANs on various benchmarks demonstrate that our method could simultaneously identify the disentangled attributes while maintaining image fidelity.

Acknowledgement

This work was supported by the MUR PNRR project FAIR (PE00000013) funded by the NextGenerationEU and by the PRIN project CREATIVE (Prot. 2020ZSL9F9).

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM TOG*, 2021. 2

[2] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM TOG*, 2020. 3

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 2

[4] Christian Bischof and Charles Van Loan. The wy representation for products of householder matrices. *SIAM Journal on Scientific and Statistical Computing*, 8(1):s2–s13, 1987. 5

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. 2

[6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 2016. 2

[8] Zikun Chen, Ruowei Jiang, Brendan Duke, Han Zhao, and Parham Aarabi. Exploring gradient-based multi-directional controls in gans. *ECCV*, 2022. 2, 3

[9] Jaewoong Choi, Junho Lee, Changyeon Yoon, Jung Ho Park, Geonho Hwang, and Myungjoo Kang. Do not escape from the manifold: Discovering the local coordinates on the latent space of gans. *ICLR*, 2022. 2

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2, 4, 6, 9

[11] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, 2020. 3

[12] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, 2020. 3

[13] 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, 4, 6, 8, 9

[14] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. 1, 2, 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 1, 2

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *ICLR*, 2022. 2

[17] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *NeurIPS*, 2020. 2, 3

[18] Zhenliang He, Meina Kan, and Shiguang Shan. Eigengan: Layer-wise eigen-learning for gans. In *ICCV*, 2021. 2, 3

[19] 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. 6

[20] Alston S Householder. Unitary triangularization of a non-symmetric matrix. *Journal of the ACM (JACM)*, 5(4):339–342, 1958. 3

[21] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. In *ICLR*, 2020. 1, 2, 3

[22] Adarsh Kappiyath, Silpa Vadakkeeveetil Sreelatha, and S Sumitra. Self-supervised enhancement of latent discovery in gans. In *AAAI*, 2022. 2

[23] Tejan Karmali, Rishubh Parihar, Susmit Agrawal, Harsh Rangwani, Varun Jampani, Maneesh Singh, and R Venkatesh Babu. Hierarchical semantic regularization of latent spaces in stylegans. In *ECCV*, 2022. 3

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 2

[25] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 2020. 2, 4, 6, 9

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 2021. 1, 2, 6, 8

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 6

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 2, 6, 7

[29] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 2, 4, 6, 7, 8

[30] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in gan for real-time image editing. In *CVPR*, 2021. 2

[31] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation. In *CVPR*, 2021. 2

[32] Richard B Lehoucq. The computation of elementary unitary matrices. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):393–400, 1996. 3

[33] Minjun Li, Yanghua Jin, and Huachun Zhu. Surrogate gradient field for latent space manipulation. In *CVPR*, 2021. 3

[34] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *NeurIPS*, 2021. 3

[35] Alexander Mathiasen, Frederik Hvilshøj, Jakob Rødsgaard Jørgensen, Anshul Nasery, and Davide Mottin. What if neural networks had svds? *NeurIPS*, 2020. 3

[36] Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections. In *ICML*, 2017. 3

[37] James Oldfield, Markos Georgopoulos, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Tensor component analysis for interpreting the latent space of gans. *BMVC*, 2021. 2

[38] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Panda: Unsupervised learning of parts and appearances in the feature maps of gans. *ICLR*, 2023. 3

[39] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, 2022. 2

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 3

[41] 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. 2, 3, 6, 7, 8, 9

[42] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. *ICLR*, 2020. 2, 3

[43] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 2

[44] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pre-trained generative models: A contrastive learning view. In *ICLR*, 2022. 2

[45] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2022. 8

[46] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*, 2022. 2

[47] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 2, 3

[48] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 2, 3, 6, 7, 8, 9

[49] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *CVPR*, 2022. 3

[50] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 2

[51] Yue Song, Andy Keller, Nicu Sebe, and Max Welling. Latent traversals in generative models as potential flows. In *ICML*. PMLR, 2023. 2, 3

[52] Yue Song, Nicu Sebe, and Wei Wang. Improving covariance conditioning of the svd meta-layer by orthogonality. In *ECCV*, 2022. 5

[53] Yue Song, Nicu Sebe, and Wei Wang. Orthogonal svd covariance conditioning and latent disentanglement. *IEEE TPAMI*, 2022. 2

[54] Nurit Spingarn-Eliezer, Ron Banner, and Tomer Michaeli. Gan" steerability" without optimization. *ICLR*, 2021. 2

[55] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. Warpedganspace: Finding non-linear rbf paths in gan latent space. In *ICCV*, 2021. 2

[56] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020. 2, 3

[57] 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. 2, 3, 6, 7, 8, 9

[58] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, 2021. 2

[59] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *CVPR*, 2021. 3

[60] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. Transeditor: Transformer-based dual-space gan for highly controllable facial editing. In *CVPR*, 2022. 2

[61] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *IJCV*, 2021. 2, 3

[62] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *CVPR*, 2021. 3

[63] 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, 4, 6, 7, 8

[64] Jiong Zhang, Qi Lei, and Inderjit Dhillon. Stabilizing gradients for deep neural networks via efficient svd parameterization. In *ICML*, 2018. 3

[65] Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, and Qifeng Chen. Low-rank subspaces in gans. *NeurIPS*, 2021. 3

[66] Jiapeng Zhu, Yujun Shen, Yinghao Xu, Deli Zhao, and Qifeng Chen. Region-based semantic factorization in gans. *ICML*, 2022. 3, 5, 8

[67] Xinqi Zhu, Chang Xu, and Dacheng Tao. Learning disentangled representations with latent variation predictability. In *ECCV*. Springer, 2020. 2