Analyzing the Latent Space of GAN through Local Dimension Estimation

Jaewoong Choi 1 Geonho Hwang 1 Hyunsoo Cho 2 Myungjoo Kang 2

Abstract

The impressive success of style-based GANs (StyleGANs) in high-fidelity image synthesis has motivated research to understand the semantic properties of their latent spaces. In this paper, we approach this problem through a geometric analysis of latent spaces as a manifold. In particular, we propose a local dimension estimation algorithm for arbitrary intermediate layers in a pre-trained GAN model. The estimated local dimension is interpreted as the number of possible semantic variations from this latent variable. Moreover, this intrinsic dimension estimation enables unsupervised evaluation of disentanglement for a latent space. Our proposed metric, called Distortion, measures an inconsistency of intrinsic tangent space on the learned latent space. Distortion is purely geometric and does not require any additional attribute information. Nevertheless, Distortion shows a high correlation with the global-basis-compatibility and supervised disentanglement score. Our work is the first step towards selecting the most disentangled latent space among various latent spaces in a GAN without attribute labels.

1. Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have achieved remarkable success in generating realistic high-resolution images (Karras et al., 2018; 2019; 2020b; 2020a; Brock et al., 2018). Nevertheless, understanding how GAN models represent the semantics of images in their latent spaces is still a challenging problem. To this end, several recent works investigated the disentanglement (Bengio et al., 2013) properties of the latent space in GAN (Goetschalckx et al., 2019; Jahanian et al., 2019; Plumerault et al., 2020; Shen et al., 2020). A latent space is called disentangled if there is a bijective correspondence between each semantic attribute and each axis of latent space when represented with the optimal basis. (See the appendix for detail.) Due to this bijective correspondence, the disentangled latent space is a concise and interpretable representation of data (Bengio et al., 2013).

The style-based GAN models (Karras et al., 2019; 2020b) have been popular in previous studies for identifying a disentangled latent space in a pre-trained model. First, the space of style vector, called W-space, was shown to provide a better disentanglement property compared to the latent noise space Z (Karras et al., 2019). After that, several attempts have been made to discover other disentangled latent spaces, such as W+space (Abdal et al., 2019) and S-space (Wu et al., 2020). However, the degree of disentanglement was assessed by the manual inspection (Karras et al., 2019; Abdal et al., 2019; Wu et al., 2020) or by the quantitative scores employing a pre-trained feature extractor (Karras et al., 2019) or an attribute annotator (Karras et al., 2019; Wu et al., 2020). The manual inspection is vulnerable to sample dependency, and the quantitative scores depend on the pre-trained models and the set of selected target attributes. Therefore, we need an unsupervised quantitative evaluation scheme for the disentanglement of latent space that does not rely on pre-trained models.

In this paper, we investigate the semantic property of a latent space in pre-trained GAN models by analyzing its geomet-
rical property. Specifically, we propose a method for estimating the local (manifold) dimension of the learned latent space. Recently, Choi et al. (2022b) demonstrated that, in the learned latent space, the basis of tangent space describes the disentangled semantic variations. In this regard, the local intrinsic dimension represents the number of disentangled semantic variations from the given latent variable. Furthermore, this intrinsic dimension estimate leads to an unsupervised quantitative score for the disentanglement property of the latent space, called Distortion. Our experiments demonstrate that our proposed metric shows a high correlation with the global-basis-compatibility and supervised disentanglement score. (The global-basis-compatibility will be defined in Sec 4.) Our contributions can be summarized as follows:

1. We propose a local intrinsic dimension estimation scheme for an intermediate latent space in pre-trained GAN models. The estimated dimension and the corresponding refined latent manifold describe the principal variations of generated images.

2. We propose a layer-wise disentanglement score, called Distortion, that measures the inconsistency of intrinsic tangent space. The proposed metric shows a high correlation with the global-basis-compatibility and supervised disentanglement score.

3. We analyze the intermediate layers of the mapping network through Distortion metric. Our analysis elucidates the superior disentanglement of $W$-space compared to the other intermediate layers and suggests a criterion for finding a similar-or-better alternative.

### 2. Related Works

**Style-based Generator** Recently, GANs with style-based generator architecture (Karras et al., 2019; 2020b; 2021; Sauer et al., 2022) have achieved state-of-the-art performance in realistic image generation. In conventional GAN architecture, such as DCGAN (Radford et al., 2016) and ProGAN (Karras et al., 2018), the generator synthesizes an image by transforming a latent noise with a sequence of convolutional layers. On the other hand, the style-based generator consists of two subnetworks: mapping network $f: Z \rightarrow W$ and synthesis network $g: \mathbb{R}^{n_0} \times W^L \rightarrow X$. The synthesis network is similar to conventional generators in that it is composed of a series of convolutional layers $\{g_i\}_{i=1,\cdots,L}$. The key difference is that the synthesis network takes the learned constant feature $y_0 \in \mathbb{R}^{n_0}$ at the first layer $g_0$, and then adjusts the output image by injecting the layer-wise styles $w$ and noise (Layer-wise noise is omitted for brevity):

$$y_i = g_i(y_{i-1}, w) \text{ with } w = f(z) \text{ for } i = 1, \cdots, L, (1)$$

where the style vector $w$ is attained by transforming a latent noise $z$ via the mapping network $f$.

**Understanding Latent Semantics** The previous attempts to understand the semantic property of latent spaces in StyleGANs are categorized into two topics: (i) finding more disentangled latent space in a model; (ii) discovering meaningful perturbation directions in a latent space corresponding to disentangled semantics. More precisely, the semantic perturbation in (ii) is defined as a latent perturbation that changes a generated image in only one semantics. Several studies on (i) suggested various disentangled latent spaces in StyleGAN models, for example, $W$ (Karras et al., 2019), $W^+$ (Abdal et al., 2019), $P_N$ (Zhu et al., 2020), and $S$-space (Wu et al., 2020). However, the disentanglement of each latent space was compared by observing samples manually (Karras et al., 2019; Abdal et al., 2019; Wu et al., 2020) or by quantitative metrics relying on pre-trained models (Karras et al., 2019; Abdal et al., 2019; Wu et al., 2020). Our Distortion metric achieves a high correlation with the supervised metric, e.g., DCI metric (Eastwood & Williams, 2018), while not requiring a pre-trained model.

Also, the previous works on (ii) are classified into local and global methods. The local methods find sample-wise perturbation directions (Ramesh et al., 2018; Patashnik et al., 2021; Abdal et al., 2021; Zhu et al., 2021; Choi et al., 2022b). On the other hand, the global methods search sample-independent perturbation directions that perform the same semantic manipulation on the entire latent space (Härkönen et al., 2020; Shen & Zhou, 2021; Vojnov & Babenko, 2020). **Throughout this paper, we refer to these local methods as local basis and these global methods as global basis.** GANSpace (Härkönen et al., 2020) showed that the principal components obtained by PCA can serve as the global basis. SeFa (Shen & Zhou, 2021) suggested the singular vectors of the first weight parameter applied to latent noise as the global basis. These global basis showed promising results, but they were successful in a limited area. Depending on the sampled latent variables, these methods exhibited limited semantic factorization and sharp degradation of image fidelity (Choi et al., 2022b:a). In this regard, Choi et al. (2022b) suggested the need for diagnosing a global-basis-compatibility of latent space. If there is no ideal global basis for semantic attributes in the target latent space in the first place, all global basis can only attain limited success on it. Our Distortion metric can serve as a meaningful score for testing the global-basis-compatibility of latent spaces (Sec 4).

**Semantic Analysis via Latent Manifold** Choi et al. (2022b) proposed an unsupervised method for finding local semantic perturbations based on the local geometry, called Local Basis (LB). Throughout this paper, we denote Local Basis as LB to avoid confusion with the general term "local basis" in the previous paragraph. Assume the support $Z$ of input prior distribution $p(z)$ is the entire Euclidean space, i.e., $Z = \mathbb{R}^d_z$, for example, Gaussian prior $p(z) = N(0, I)$. 

---

**Analyzing the Latent Space of GAN through Local Dimension Estimation**

---

**Semantic Analysis via Latent Manifold** Choi et al. (2022b) proposed an unsupervised method for finding local semantic perturbations based on the local geometry, called Local Basis (LB). Throughout this paper, we denote Local Basis as LB to avoid confusion with the general term "local basis" in the previous paragraph. Assume the support $Z$ of input prior distribution $p(z)$ is the entire Euclidean space, i.e., $Z = \mathbb{R}^d_z$, for example, Gaussian prior $p(z) = N(0, I)$.
We denote the target latent space by \( M = f(Z) \subseteq \mathbb{R}^d \) and refer to the subnetwork between them by \( f \). Note that \( M \) is defined as an image of the trained subnetwork \( f \). Thus, we call \( M \) the learned latent space or latent manifold following the manifold interpretation of Choi et al. (2022b).

LB is defined as the ordered basis of tangent space \( T_w M^k_w \) at \( w = f(z) \in M \) for the \( k \)-dimensional local approximating manifold \( M^k_w \). Here, \( M^k_w \subseteq M \) indicates a \( k \)-dimensional submanifold of \( M \) that approximates \( M \) around \( w \) (Fig 1) with \( k \leq d_M \):

\[
M^k_w \approx M_w
\]

where \( M_w = \{ f(z) \mid \| z - z_i \| < \epsilon \} \subseteq M \). (2)

Using the fact that \( M = f(Z) \), the local approximating manifold \( M^k_z \) can be discovered by solving the low-rank approximation problem of \( df_z \), i.e., the Jacobian matrix \( \nabla_z f \) of \( f \) because of the assumption \( T_z Z = \mathbb{R}^d \):

\[
\min_{L} \| df_z - L \|_2, \quad \text{where rank}(L) \leq k. \quad (3)
\]

The analytic solution of this low-rank approximation problem is obtained in terms of Singular Value Decomposition (SVD) by Eckart–Young–Mirsky Theorem (Eckart & Young, 1936). From that, \( M^k_w \) and the corresponding LB are given as follows: For the \( i \)-th singular vector \( u_i^w \in \mathbb{R}^{d_z} \), \( v_i^w \in \mathbb{R}^d \), and \( i \)-th singular value \( \sigma_i^w \in \mathbb{R} \) of \( df_z \) with \( \sigma_1^w \geq \cdots \geq \sigma_n^w \),

\[
\text{(SVD)} \quad df_z(u_i^w) = \sigma_i^w \cdot v_i^w \quad \text{for } i,
\]

\[
\text{LB}(w) = \{ v_i^w \}_{1 \leq i \leq n} \quad \text{where } w = f(z),
\]

\[
M^k_w = \left\{ f \left( z + \sum_{1 \leq i \leq k} t_i \cdot u_i^w \right) \mid t_i \in (-\epsilon_i, \epsilon_i) \right\}. \quad (6)
\]

Note that the tangent space of \( M^k_w \) is spanned by the top-\( k \) LB, i.e., \( T_w M^k_w = \text{span}\{ v_i^w : 1 \leq i \leq k \} \). Choi et al. (2022b) demonstrated that LB serves as the local semantic perturbations from the base latent variable \( w \). However, estimating the number of meaningful perturbations remains elusive. Since LB is defined as singular vectors, LB presents the candidates as much as \( \min(d_z, d_M) \), e.g., 512 for \( W \)-space in StyleGANs. To address this problem, we propose the local dimension estimation that can refine these candidates up to 90%. Moreover, this local dimension estimation leads to an unsupervised disentanglement metric (Sec 4).

### 3. Latent Dimension Estimation

In this section, we propose our local dimension estimation scheme for a learned latent manifold in a pre-trained GAN model. We interpret the latent manifold \( M = f(Z) \) as a noisy manifold, and hence the subnetwork differential \( df_z \) as a noisy linear map. The local dimension at \( w \in M \) is estimated as the intrinsic rank of \( df_z \). Then, we evaluate the validity of this estimation scheme. In this section, our analysis of learned latent manifold is focused on the intermediate layers in the mapping network of StyleGAN2 trained on FFHQ (See Fig 10 for architecture). However, our scheme can be applied to any \( z \)-differentiable intermediate layers for an input latent noise \( z \).

#### 3.1. Method

Throughout this work, we follow the notation presented in Sec 2. Consider a target latent manifold \( M \) given by a subnetwork \( f \), i.e., \( M = f(Z) \). Our goal is to estimate the intrinsic local dimension of \( M \) around \( w = f(z) \). Formally, the local dimension at \( w \in M \) is the dimension of Euclidean space \( \mathbb{R}^k \) whose open set is homeomorphic to the \( \epsilon \)-neighborhood of \( w \). However, the latent manifold exhibits some noise in practice. The bijective condition of homeomorphism is too strict under the presence of noise. Therefore, we define the local intrinsic dimension of learned latent space as the local dimension of denoised latent space (Fig 1). In particular, we discover this intrinsic dimension by estimating the intrinsic rank of noisy subnetwork differential \( df_z \), i.e., the Jacobian \( (\nabla_z f)(z) \). The correspondence between the local dimension and rank of \( df_z \) is described in Eq 6 because the rank of a linear map is the same as the number of singular vectors with non-zero singular values.

**Motivation** Before presenting our dimension estimation method, we provide motivation for introducing the lower-dimensional approximation to \( M \). Figure 2a shows the singular value distribution of Jacobian matrices evaluated for the subnetworks of the mapping network in StyleGAN2. Last layer \( i \) denotes the subnetwork from the input noise space \( Z \) to the \( i \)-th fully connected layer. The distribution of singular values \( \{ \sigma_i^w \} \), gets monotonically sparser as the subnetwork gets deeper. In particular, \( W \)-space, i.e., Last layer 8, is extremely sparse as much as \( \sigma_{150}^w/\sigma_1^w \approx 0.005 \). Therefore, it is reasonable to prune the singular values with negligible magnitude and consider the lower-dimensional approximation of the learned latent manifold.

**Local Dimension Estimation** Our intrinsic rank estimation algorithm distinguishes the large meaningful components and the small noise-like components given the singular values \( \{ \sigma_i^w \} \) of \( (\nabla_z f)(z) \). The Pseudorank algorithm (Kritchman & Nadler, 2008) estimates the number of meaningful components in a linear mixture model from noisy samples. We reinterpret this algorithm to find the intrinsic rank of noisy \( (\nabla_z f)(z) \). Assume the isotropic Gaussian noise on the Jacobian \( (\nabla_z f)(z) \in \mathbb{R}^{d_M \times d_Z} \):

\[
(\nabla_z f)(z) = L(z) + \sigma \cdot (\epsilon_1, \cdots, \epsilon_{d_M})^T,
\]

with \( \epsilon_i \sim \mathcal{N}(0,I_{d_Z}) \) where \( L(z) \) denotes the denoised low-rank representation of \( (\nabla_z f)(z) \). Then, taking the expecta-
tion over the noise distribution gives:
\[ \mathbb{E}_c \left( (\nabla_x f)^\top (z) \cdot (\nabla_x f)(z) \right) = L^\top (z) \cdot L(z) + \sigma^2 \cdot I_{d_z}. \]
(8)
The eigenvalues of \((\nabla_x f)^\top (z) \cdot (\nabla_x f)(z)\) are the squares of singular values \(\{\sigma_i^2\}_{i = 1}^9\), and the noise covariance term \(\sigma^2 \cdot I_{d_z}\) increases all eigenvalues by \(\sigma^2\). This observation explains our intuition that large singular values correspond to signals and small ones correspond to noise. Therefore, determining the intrinsic rank of \((\nabla_x f)(z)\) is closely related to the largest eigenvalue \(\lambda_1\) of the empirical covariance matrix \(S = \frac{1}{d_z} \sum_{i = 1}^n \epsilon_i \epsilon_i^\top\), which is the threshold for distinguishing between signal and noise. In the random matrix theory, it is known that the distribution of this largest eigenvalue \(\lambda_1\) for \(n\)-samples of \(\mathcal{N}(0, I_p)\) converges to a Tracy-Widom distribution \(F_\beta\) of order \(\beta = 1\) for real-valued observations (Johnstone, 2001) (See the appendix for detail):
\[ P \left( \lambda_1 < \sigma^2 (\mu_{n, p} + s \cdot \sigma_{n, p}) \right) \to F_\beta(s) \quad \text{as } n, p \to \infty \text{ with } c = p/n \text{ fixed}. \]
(9)
Using the above theoretical results, our algorithm applies a sequence of nested hypothesis tests. Given the Jacobian \((\nabla_x f)(z) \in \mathbb{R}^{d_M \times d_z}\), let \(p = \min(d_M, d_z)\). Then, for \(k = 1, 2, \cdots, p - 1\),
\[ H_0^k : \text{rank at least } k \quad \text{vs.} \quad H_1^k : \text{rank at most } (k - 1) \]
(10)
For each \(k\), the hypothesis test consists of two parts. First, the noise level \(\sigma\) of \((\nabla_x f)(z)\) in Eq 7 should be estimated to perform a hypothesis test. We adopted the consistent noise estimate \(\sigma_{est}(k)\) from Kritchman & Nadler (2008), which is derived by assuming that \(\lambda_{k+1}, \lambda_{k+2}, \cdots, \lambda_p\) are the noise components for \(\lambda_i = (\sigma_i^2)\). Second, we test whether \(\lambda_k\) belongs to the noise components based on the corresponding Tracy-Widom distribution as follows:
\[ \lambda_k \leq \sigma_{est}(k) (\mu_{n, p - k} + s(\alpha) \cdot \sigma_{n, p - k}), \]
(11)
where \(\alpha\) denotes a confidence level. We chose \(\alpha = 0.1\) in our experiments. The above test is repeated until Eq 11 is satisfied. Then, the estimated rank \(K\) becomes \(K = k - 1\) because Eq 11 means that \(\lambda_k\) is not large enough to be judged as a signal component.

**Preprocessing** This algorithm supposes the isotropic Gaussian noise on the Jacobian matrix. However, even considering the randomness of empirical covariance, the observed singular values of Jacobian matrix are too sparse \((\sigma_{min}/\sigma_{max} \approx 10^{-9})\). Hence, the isotropic Gaussian assumption leads to the underestimation of the noise level, which causes the overestimation of intrinsic rank (In our experiments, estimated rank > 200 and \(\sigma_{rank} \approx 0.0003\)). To address this problem, we introduce a simple preprocessing on the singular values of the Jacobian. Before applying the rank estimation algorithm, we filter out the singular values \(\{\sigma_i^2\}\) with \(\{(\sigma_i^2) \leq \theta_{pre} \cdot \max_i (\sigma_i^2)\}\). We set \(\theta_{pre} \in \{0.0005, 0.001, 0.005, 0.01\}\).

**Validity of the estimated local dimension** Our local dimension estimate is based on the Jacobian \((\nabla_x f)\), where \(f\) defines the latent manifold \(M = f(Z)\). Because \((\nabla_x f)\) is the first order approximation of \(f\), the estimated dimension might be improper for \(M\). Therefore, we designed and conducted the Off-manifold experiment to assess the validity of the estimated local dimension. Intuitively, the intrinsic local dimension at \(w \in M\) is discovered by refining minor variations of the noisy latent manifold. The goal of Off-manifold experiment is to test whether the latent perturbation along the \(k\)-th coordinate axis stays inside \(M\). If the margin of \(M\) at \(w\) in the \(k\)-th axis is large enough, then the \(k\)-th axis is needed to locally approximate \(M\). To be more specific, we solve the following optimization problem by Adam optimizer (Kingma & Ba, 2015) on MSE loss with a learning rate 0.005 for 1000 iterations for each \(k\) with \(z_0 = z_{init}\):
\[ w_{init} = f(z_{init}), \quad w_{ptb} = w_{init} + c \cdot v_{k, init}^{w_{init}}, \]
\[ z_{opt} = \arg \min_z ||w_{ptb} - f(z)||^2. \]
(12)
Here, \(v_{k, init}^{w_{init}}\) denotes the \(k\)-th LB at \(w_{init}\). \(w_{init}\) is perturbed along \(v_{k, init}^{w_{init}}\) because it is the tangent vector at \(w_{init}\) along \(k\)-th axis (Eq 4).

We ran the Off-manifold experiments on \(W\)-space of StyleGAN2. Figure 2b shows the final objective \(||w_{ptb} - f(z_{opt})||^2\) after the optimization for each LB \(v_{k, init}^{w_{init}}\) with...
3.2. Comparison to Previous Rank Estimation

**Sparsity Constraint** LowRankGAN (Zhu et al., 2021) introduced a convex optimization problem called Principal Component Pursuit (PCP) (Candes et al., 2011) to find a low-rank factorization of Jacobian $(\nabla_z f)(z)$ (Eq 14):

$$
\min_{L,S} \|L\|_* + \gamma \cdot \|S\|_1 \\
\text{s.t. } L + S = (\nabla_z f)^T(z) \cdot (\nabla_z f)(z),
$$

where $\|L\|_*$ is the nuclear norm, i.e., the sum of all singular values, $\|S\|_1 = \sum_{i,j} |S_{i,j}|$, and $\gamma > 0$ is a positive regularization parameter. PCP encourages the sparsity on corruption $S$ through $\ell_1$ regularizer.

However, we believe that this sparsity assumption is not adequate for finding the intrinsic rank of Jacobian. To test the validity of the sparsity assumption, we monitored how the low-rank representation $L$ changes as we vary the regularization parameter $n = 1/\gamma$ as in (Zhu et al., 2021) (Fig 2c). The estimated rank decreases unceasingly without saturation as we increase $n$, i.e., refine the Jacobian stronger. We consider that the rank saturation should occur if this assumption is adequate for finding an intrinsic rank because it implies regularization robustness. But the low-rank factorization through PCP does not show any saturation until the Frobenius norm of corruption $\|S\|_F$ reaches over 50% of the initial matrix $\|(\nabla_z f)^T(z) \cdot (\nabla_z f)(z)\|_F$.

**Interpretation as Frobenious Norm** Our algorithm can be interpreted as a Nuclear-Norm Penalization (NNP) problem (Eq 15) for matrix denoising (Donoho & Gavish, 2014). This NNP framework is similar to PCP in LowRankGAN except for the regularization $\|E\|_F$. While PCP requires an iterative optimization of Alternating Directions Method of Multipliers (ADMM) (Boyd et al., 2011; Lin et al., 2010), NNP provides an explicit closed-form solution $L^*$ through SVD, i.e., $(\nabla_z f)(z) = U \Sigma V^T$:

$$
\min_{L,S} \|L\|_* + \gamma \|E\|_F \\
\text{s.t. } L + E = (\nabla_z f)(z). \quad (15)
$$

for $(M_{i,j})_{i,j} = \max (M_{i,j} \cdot 0)$. Therefore, the intrinsic rank by NNP is determined by a threshold $1/(2\gamma)$ for the singular values $(\sigma_i^*)$ of Jacobian. Our algorithm can be interpreted as selecting this threshold via a series of hypothesis tests.

3.3. Latent Space Analysis of StyleGAN

We analyzed the intermediate layers of the mapping network in StyleGAN2 trained on FFHQ using our local dimension estimation (See Fig 10 for StyleGAN architectures). First, Figure 3 shows the distribution of estimated local dimensions for 1k samples of each intermediate layer for each $\theta_{pre}$ (See the appendix for the rank statistics under all $\theta_{pre}$). Note that the algorithm provides an unstable rank estimate on the most unsparselayer 1st layer (Fig 2a) under the small $\theta_{pre} \in \{0.0005, 0.001\}$. However, this phenomenon was not observed in the other layers. Hence, we focused our analysis on the layers with reasonable depth, i.e., from 3 to 8. Even though changing $\theta_{pre}$ results in an overall shift of the estimation, the trend and relative ordering between layers are the same. In accordance with Fig 2a, the intrinsic dimension monotonically decrease as the layer goes deeper.

Second, we evaluated the estimated rank on image space (Fig 4). Figure 4a and 4b show the image traversal along the first two axes and the two axes $(d - 1, d)$ around the estimated rank $d$ with $\theta_{pre} = 0.01$ (See Fig 22 for comparison between axis $i < d$ and $i \geq d$.) Fig 4c presents the size of the directional derivative $\|\nabla_{\psi^i} g(w)\|_F$ along the $i$-th LB $\psi^i$ at $w$, estimated by the finite difference scheme. The
Analyzing the Latent Space of GAN through Local Dimension Estimation

These two questions are closely related because the ideal disentanglement includes a global basis representation where each element corresponds to the attribute-coordinate. To answer these questions, we propose an unsupervised global disentanglement metric, called Distortion. We evaluate the global-basis-compatibility by the image fidelity (Q1) and the disentanglement by semantic factorization (Q2). Our experimental results show that our proposed metric achieves a high correlation with the global-basis-compatibility (Q1) and the supervised disentanglement score (Q2) on various StyleGANs. (See Fig 9 in appendix for robustness of Distortion to $\theta_{pre}$.)

Distortion Score Intuitively, our global disentanglement score assesses the inconsistency of intrinsic tangent space for each latent manifold. The framework of analyzing the semantic property of a latent space via its tangent space was first introduced in Choi et al. (2022b). This framework was inspired by the observation that each basis vector (LB) of a tangent space corresponds to a local semantic perturbation. In this work, we develop this idea and propose a layer-wise score for global disentanglement property. As in Choi et al. (2022b), we employ the Grassmannian (Boothby, 1986) metric to measure a distance between two tangent spaces.

For two $k$-dimensional subspaces $W, W'$ of $\mathbb{R}^n$, let $M_W, M_{W'} \in \mathbb{R}^{n \times k}$ be the column-wise concatenation of orthonormal basis for $W, W'$, respectively. Then, the dimension-normalized Geodesic Metric is defined as

$$d_{geo}(W, W') = \left( \frac{1}{k} \sum_{i=1}^{k} \theta_i^2 \right)^{1/2}$$

where

$$\theta_i = \cos^{-1}(\sigma_i(M_W^T M_{W'}^T))$$

denotes the $i$-th principal angle between $W$ and $W'$ for $i$-th singular value $\sigma_i$. Then, Distortion score $D_M$ for the latent manifold $M$ is evaluated as follows:

1. To assess the overall inconsistency of $M$, measure the Grassmannian distance between two intrinsic tangent spaces $T_w, M_w^k$ (Eq 6) at two random $w \in M$ for $k = \min(k_1, k_2)$.

$$I_{rand} = \mathbb{E}_{z_i \sim p(z)} \left[ d_{geo}^2 \left( T_{w_1, w_1}^k, T_{w_2, w_2}^k \right) \right]$$

2. To normalize the overall inconsistency, measure the same Grassmannian distance between two close $w \in M$ for $\epsilon = 0.1$.
Figure 6: **Correlation between Distortion metric (\( \downarrow \)) and DCI (\( \uparrow \)) when \( \theta_{pre} = 0.005 \).** DCI (Eastwood & Williams, 2018) is a supervised disentanglement metric that requires attribute annotations.

\[
I_{local} = \mathbb{E}_{z_1 \sim p(z)} \left[ d^h_{geo} \left( T_{w_1} \mathcal{M}_{\theta_1}, T_{w_2} \mathcal{M}_{\theta_2} \right) \right].
\]

3. **Distortion and Global Disentanglement** In this paragraph, we clarify why the globally disentangled latent space shows a low Distortion score. Assume a latent space \( \mathcal{M} \) is globally disentangled. Then, there exists an optimal global basis of \( \mathcal{M} \), where each basis vector corresponds to an image attribute on the entire \( \mathcal{M} \). By definition, this optimal global basis is the local basis at all latent variables. Assuming that LB finds the local basis (Choi et al., 2022b), each global basis vector would correspond to one LB vector at each latent variable. In this regard, our local dimension estimation finds a principal subset of LB, which includes these corresponding basis vectors. In conclusion, if the latent space is globally disentangled, this principal set of LB at each latent variable would contain the common global basis vectors. Hence, the intrinsic tangent spaces would contain the common subspace generated by this common basis, which leads to a small Grassmannian metric between them. Therefore, the global disentanglement of the latent space leads to a low Distortion score.

**Global-Basis-Compatibility** We tested whether Distortion \( D_M \) is meaningful in estimating the global-basis-compatibility. We chose GANSpace (Häärkönen et al., 2020) as a reference global-basis because of its broad applicability. As a measure of global-basis-compatibility, we adopted FID (Heusel et al., 2017) gap between LB and GANSpace. Here, we interpret FID score under LB traversal as the optimal image fidelity that the latent space can achieve because LB is the basis of tangent space. Thus, FID gap between LB and GANSpace represents the difference between the optimal image fidelity and the fidelity achieved by a global basis. Therefore, we selected this FID gap for assessing the global-basis-compatibility. FID is measured for 50k samples of perturbed images along the 1st component of LB and GANSpace, respectively. Distortion metric is tested on StyleGAN2 on LSUN Cat, StyleGAN2 with configs E and F (Karras et al., 2020b) on FFHQ to test the generalizability of correlation to the global-basis-compatibility. StyleGAN2 in Fig 5 denotes StyleGAN2 with config F because config F is the usual StyleGAN2 model. The perturbation intensity is set to 5 in LSUN Cat and 3 in FFHQ. Distortion metric shows a strong positive correlation of 0.98, 0.81 and 0.70 to FID gap in Fig 5. (See Sec G for correlations on other \( \theta_{pre} \).) This result demonstrates that Distortion metric can be an unsupervised criterion for selecting a latent space with...
Analyzing the Latent Space of GAN through Local Dimension Estimation

![Image](image.png)

Figure 7: **Subspace Traversal on the intermediate layers** along the global basis. The upper-right corner of max-distorted layer 3 and layer 8 show visual artifacts. However, the min-distorted layer 7 does not show such a failure. The initial image (center) is traversed along the 1st (horizontal) and 2nd (vertical) components of GANSpace with a perturbation intensity 4.

This suggests that, before finding a global basis, we can use Distortion metric as a prior investigation for selecting a proper target latent space.

**Disentanglement Score** We assessed a correlation between the unsupervised Distortion metric and a supervised disentanglement score. Following Wu et al. (2020), we adopted DCI score (Eastwood & Williams, 2018) as the supervised disentanglement score for evaluation, and employed 40 binary attribute classifiers pre-trained on CelebA (Liu et al., 2015) to label generated images. Each DCI score is assessed on 10k samples of latent variables with the corresponding attribute labels. In Fig 6, StyleGAN1, StyleGAN2-e, and StyleGAN2 refer to StyleGAN1 and StyleGAN2s with config E and F trained on FFHQ. Note that DCI experiments are all performed on FFHQ because the DCI score requires attribute annotations. DCI and Distortion metrics show a strong negative correlation on StyleGAN1 and StyleGAN2-e. The correlation is relatively moderate on StyleGAN2. This moderate correlation is because Distortion metric is based on the Grassmannian metric. The Grassmannian metric measures the distance between tangent spaces, while DCI is based on their specific basis. Even if the tangent space is identical so that Distortion becomes zero, DCI can have a relatively low value depending on the choice of basis. Hence, in StyleGAN2, the high-distorted layers showed low DCI scores, but the low-distorted layers showed relatively high variance in DCI score. Nevertheless, the strong correlation observed in the other two experiments suggests that, in practice, the basis vector corresponding to a specific attribute has a limited variance in a given latent space. Therefore, Distortion metric can be an unsupervised indicator for the supervised disentanglement score.

**Traversal Comparison** In all StyleGAN models, layer 7 and layer 8 ($\mathcal{W}$-space) achieved the smallest Distortion metric (Fig 5 and 6). These results explain the superior disentanglement of renowned $\mathcal{W}$-space. Moreover, our Distortion suggests that layer 7 can be an alternative to $\mathcal{W}$-space. Our global-basis-compatibility results imply that the min-distorted layer can present better **image fidelity** under the global basis traversal. For a visual comparison, we observed the subspace traversal (Choi et al., 2022b) along the global basis on the max-distorted layer 3, min-distorted layer 7, and $\mathcal{W}$-space of StyleGAN2 on FFHQ. The subspace traversal visualizes the two-dimensional latent perturbation. Hence, the subspace traversal can reveal the deviation from the tangent space more clearly than the linear traversal, because the deviation is assessed in two dimensions. Note that the tangent space is the optimal perturbation in image fidelity. In Fig 7, the global basis shows visual artifacts at the corners in the subspace traversal on the max-distorted layer 3 and $\mathcal{W}$-space. Nevertheless, the min-distorted layer 7 shows the stable traversal without any failure. This result proves that comparing Distortion scores can be a criterion for selecting a better latent space with higher global-basis-compatibility. (See Sec J for additional image fidelity comparisons and Sec I for **semantic factorization** comparisons.)

5. Conclusion

In this paper, we proposed a local intrinsic dimension estimation algorithm for the intermediate latent space in a pre-trained GAN. Using this algorithm, we analyzed the intermediate layers in the mapping network of StyleGANs on various datasets. Moreover, we suggested an unsupervised global disentanglement metric called Distortion. The analysis of the mapping network demonstrates that Distortion metric shows a high correlation between the global-basis-compatibility and disentanglement score. Although finding an optimal preprocessing hyperparameter $\theta_{pre}$ was beyond the scope of this work, the proposed metric showed robustness to the hyperparameter. Moreover, our local dimension estimation has the potential to be applied to feature spaces of various models. This kind of research would be an inter-
esting future work.

References

Abdal, R., Qin, Y., and Wonka, P. Image2stylegan: How to embed images into the stylegan latent space? In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4432–4441, 2019.

Abdal, R., Zhu, P., Mitra, N. J., and Wonka, P. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (TOG), 40(3):1–21, 2021.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

Boothby, W. M. An introduction to differentiable manifolds and Riemannian geometry. Academic press, 1986.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning, 3(1):1–122, 2011.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2018.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? Journal of the ACM (JACM), 58(3), 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395.

Choi, J., Hwang, G., Cho, H., and Kang, M. Finding the global semantic representation in gan through frechet mean. arXiv preprint arXiv:2210.05509, 2022a.

Choi, J., Lee, J., Yoon, C., Park, J. H., Hwang, G., and Kang, M. Do not escape from the manifold: Discovering the local coordinates on the latent space of GANs. In International Conference on Learning Representations, 2022b.

Donoho, D. and Gavish, M. Minimax risk of matrix denoising by singular value thresholding. The Annals of Statistics, 42(6):2413–2440, 2014.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In International Conference on Learning Representations, 2018.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. Psychometrika, 1(3):211–218, 1936.

Goetschalckx, L., Andonian, A., Oliva, A., and Isola, P. Ganalyze: Toward visual definitions of cognitive image properties. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5744–5753, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In NIPS, 2014.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. Advances in Neural Information Processing Systems, 33, 2020.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

Jahanian, A., Chai, L., and Isola, P. On the” steerability” of generative adversarial networks. In International Conference on Learning Representations, 2019.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. The Annals of statistics, 29(2):295–327, 2001.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In International Conference on Learning Representations, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410, 2019.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676, 2020a.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119, 2020b.

Karrasch, D. An introduction to grassmann manifolds and their matrix representation. 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Kritchman, S. and Nadler, B. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1): 19–32, 2008.

Lin, Z., Chen, M., and Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.

Plumerault, A., Borgne, H. L., and Hudelot, C. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*, 2020.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Ramesh, A., Choi, Y., and LeCun, Y. A spectral regularizer for unsupervised disentanglement. *arXiv preprint arXiv:1812.01161*, 2018.

Sauer, A., Schwarz, K., and Geiger, A. Stylegan-xl: Scaling stylegan to large diverse datasets. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pp. 1–10, 2022.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

Voynov, A. and Babenko, A. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pp. 9786–9796. PMLR, 2020.

Wu, Z., Lischinski, D., and Shechtman, E. Stylespace analysis: Disentangled controls for stylegan image generation. *arXiv preprint arXiv:2011.12799*, 2020.
A. Definition of disentangled latent space

Disentangled perturbation In the GAN disentanglement literature, several studies investigated the disentanglement property of the latent space by finding disentangled perturbations that make a disentangled transformation of an image in one generative factor, such as GANSpace (Härkönen et al., 2020), SeFa (Shen & Zhou, 2021), and Local Basis (Choi et al., 2022b). To be more specific, for a latent variable \( z \in \mathcal{Z} \subset \mathbb{R}^d \), let \( f = (f_1, f_2, \ldots, f_d) \) be a generative factor of \( G(z) \) where \( G \) denotes the generator. \( T_j(x) \) denotes a transformation of an image \( x \) in the \( j \)-th generative factor. The disentangled perturbation \( v_j(z) \) for the base latent variable \( z \) on the \( j \)-th generative factor is defined as follows (The perturbation intensity \( \|v_j(z)\| \) and the corresponding change in \( j \)-th generative factor \( \beta f_j \) is omitted for brevity.):

\[
G(z + v_j(z)) = T_j(G(z)).
\]

(17)

In this paper, the global basis refers to the sample-independent disentangled perturbations on a latent space:

\[
v_j(z) = v_j \quad \text{for all} \quad z \in \mathcal{Z}. \quad (18)
\]

For example, consider a pre-trained GAN model that generates face images. Then, the disentangled perturbation in this model is the latent perturbation direction that make the generated face change only in the wrinkles or hair color as presented in Härkönen et al. (2020). This disentangled perturbation is the global basis if all generated images show the same semantic variation when latent perturbed along it.

Disentangled space The (globally) disentangled latent space is defined in terms of disentangled perturbations. The latent space is globally disentangled if there exists the global basis for the generative factors of data. In other words, for each generative factor \( f_j \) for \( 1 \leq j \leq d \), there exists a corresponding latent perturbation direction \( v_j \) such that all latent variables show the semantic variation in \( f_j \) when perturbed along \( v_j \). Then, we can interpret the vector component of this global basis \( v_j \) as having a correspondence with the \( j \)-th generative factor \( f_j \).

\[
f_j \leftrightarrow c_j \quad \text{when} \quad z = \sum_{1 \leq j \leq d} c_j \cdot v_j \quad \text{and} \quad f_j \text{ denotes the } j \text{-th generative factor of } G(z). \quad (19)
\]

In this paper, we described the above correspondence as the representation of globally disentangled latent space in the attribute-coordinate in Sec 4. This is consistent with the definition of disentanglement, introduced in Bengio et al. (2013). For example, consider the dSprites dataset. The dSprites is a synthetic dataset consisting of two-dimensional shape images, which is widely used for disentanglement evaluation. The generative factors of dSprites are shape, scale, orientation, position on the x-axis, and position on the y-axis. Then, the (globally) disentangled latent space for the dSprites dataset is a five-dimensional vector space \( Z = \mathbb{R}^5 \) where \( c_1 \) represents the shape, \( c_2 \) represents the scale, and so on.

B. Noise Estimation of Pseudorank Algorithm

For completeness, we include the convergence theorem for the largest eigenvalue of the empirical covariance matrix for Gaussian noise in Johnstone (2001) and the noise estimation algorithm provided in Kritchman & Nadler (2008).

Theorem B.1 ((Johnstone, 2001)). The distribution of the largest eigenvalue \( \lambda_1 \) of the empirical covariance matrix for \( n \)-samples of \( \mathcal{N}(0, I_p) \) converges to a Tracy-Widom distribution:

\[
P \left( \lambda_1 < \sigma^2 (\mu_{n,p} + s \cdot \sigma_{n,p}) \right) \rightarrow F_\beta(s) \quad \text{as} \quad n, p \rightarrow \infty \quad \text{with} \quad c = p/n \text{ fixed.} \quad (20)
\]

where

\[
\mu_{n,p} = \frac{1}{n} \left( \sqrt{n - \frac{1}{2}} + \sqrt{p - \frac{1}{2}} \right)^2, \quad (21)
\]

\[
\sigma_{n,p} = \frac{1}{n} \left( \sqrt{n - \frac{1}{2}} + \sqrt{p - \frac{1}{2}} \right) \left( \frac{1}{\sqrt{n - 1/2}} + \frac{1}{\sqrt{p - 1/2}} \right)^{1/3}, \quad (22)
\]

where \( F_\beta \) denotes the Tracy-Widom distribution of order \( \beta = 1 \) for real-valued observations.
Algorithm. Solve the following non-linear system of $K + 1$ equations involving the $K + 1$ unknowns $\hat{\rho}^1, \ldots, \hat{\rho}^K$ and $\sigma^2_{\text{est}}$:

\[
\sigma^2_{KN} - \frac{1}{p-K} \left[ \sum_{j=K+1}^{p} \lambda_j + \sum_{j=1}^{K} (\lambda_j - \hat{\rho}_j) \right] = 0, \\
\hat{\rho}_j^2 - \hat{\rho}_j \left( \lambda_j + \sigma^2_{\text{est}} - \frac{p-K}{n} \right) + \lambda_j \sigma^2_{\text{est}} = 0.
\]

This system of equations can be solved iteratively. Check Kritchman & Nadler (2008) for detail.

C. Relation Between Rank Estimation Algorithm and Optimization

Theorem C.1. The following optimization problem, called Nuclear-Norm Penalization (NNP),

\[
\begin{array}{ll}
\min_{L,S} & ||L||_* + \gamma \cdot ||E||_F \\
\text{s.t.} & L + E = (\nabla_x f)(z),
\end{array}
\]

has a solution

\[
L^* = U \left( \Sigma - \frac{1}{2\gamma} I \right)_+ \top V,
\]

where $(\nabla_x f)(z) = U \Sigma V^\top$ (SVD) and $(M_+)_{i,j} = \max (M_{i,j}, 0)$.

Proof. Denote $Y := \nabla_x f$ and $h(L) := ||L||_* + \gamma ||Y - L||_F$. We want to show that $L_*$ minimizes $h(L)$. Then, the necessary and sufficient condition for this is:

\[
0 \in \partial h(L_*) = \{2\gamma (L_* - Y) + z : z \in \partial ||L_*||_* \} \iff 2\gamma (Y - L_*) \in \partial ||L_*||_*.
\]

Note that $Y = U \Sigma V^\top$ and $\hat{L} = U (\Sigma - \frac{1}{2\gamma} I)_+ \top V$. We can write

\[
Y = U_1 \Sigma_1 V_1^\top + U_2 \Sigma_2 V_2^\top,
\]

with $\text{diag}(\Sigma_1) > \frac{1}{2\gamma}$ and $\text{diag}(\Sigma_2) \leq \frac{1}{2\gamma}$. Then,

\[
\hat{L}_* = U_1 \left( \Sigma - \frac{1}{2\gamma} I \right) V_1^\top,
\]

\[
Y - \hat{L}_* = U_2 \Sigma_2 V_2^\top + \frac{1}{2\gamma} U_1 V_1^\top = \frac{1}{2\gamma} (U_1 V_1^\top + 2\gamma U_2 \Sigma_2 V_2^\top).
\]

By doing tedious calculation, we can verify that $U_1 V_1^\top + 2\gamma U_2 \Sigma_2 V_2^\top$ meet the condition of Lemma C.2, so that $U_1 V_1^\top + 2\gamma U_2 \Sigma_2 V_2^\top \in \partial ||\hat{L}_*||_*$. Therefore, $0 \in \partial h(\hat{L}_*)$ and it completes the proof.

Lemma C.2. Let $X \in \mathbb{R}^{m \times n}$ and $f(x) = ||X||_*$. Then,

\[
\partial f(X) = \partial ||X||_* = \{ Z \in \mathbb{R}^{m \times n} : ||Z||_2 \leq 1 \text{ and } \langle Z, X \rangle = ||X||_* \}.
\]

Proof. If $Z \in \partial f(X)$, then

\[
f(Y) \geq f(X) + \langle Z, Y - X \rangle, \quad \forall Y \in \mathbb{R}^{m \times n},
\]

\[
\iff \langle Z, X \rangle - ||X||_* \geq \langle Z, Y \rangle - ||Y||_*, \quad \forall Y \in \mathbb{R}^{m \times n},
\]

\[
\iff \langle Z, X \rangle - ||X||_* \geq \sup_{Y \in \mathbb{R}^{m \times n}} \{ \langle Z, Y \rangle - ||Y||_* \} = \begin{cases} 0, & \text{if } ||Z||_2 \leq 1, \\ \infty, & \text{otherwise}. \end{cases}
\]

And $0 \leq \langle Z, X \rangle - ||X||_* = \langle Z, X \rangle - \sup_{||M||_2 \leq 1} \langle M, X \rangle \leq 0$, thus $\langle Z, X \rangle = ||X||_*$. 

\[\square\]
D. Grassmannian Metric for Distortion - Geodesic vs. Projection

Our proposed Distortion metric $D = \frac{I_{\text{rand}}}{I_{\text{local}}}$ is defined as the relative inconsistency of intrinsic tangent spaces on a latent manifold (Sec 4). The inconsistency $(I_{\text{rand}}$, $I_{\text{local}}$) is measured by the Grassmannian (Boothby, 1986) distance between tangent spaces, particularly by Geodesic Metric (Ye & Lim, 2016). In this section, we present why we choose the Geodesic Metric instead of the Projection Metric (Karrasch, 2017) among Grassmannian distances. Informally, the Geodesic Metric provides a better discriminability compared to the Projection Metric. For completeness, we begin with the definitions of the Grassmannian manifold and two distances defined on it.

Definitions  Let $V$ be the $n$-dimensional vector space. The Grassmannian manifold $\text{Gr}(k, V)$ (Boothby, 1986) is defined as the set of all $k$-dimensional linear subspaces of $V$. Then, for two $k$-dimensional subspaces $W, W' \in \text{Gr}(k, V)$, two Grassmannian metrics are defined as follows:

$$d_{\text{proj}}(W, W') = \|P_W - P_{W'}\|, \quad d_{\text{geo}}(W, W') = \left(\sum_{i=1}^{k} \theta_i^2\right)^{1/2}.$$  \hspace{1cm} (35)

For the Projection Metric $d_{\text{proj}}(W, W')$, $P_W$ and $P_{W'}$ denote the projection into each subspace and $\|\cdot\|$ represents the operator norm. For the Geodesic Metric $d_{\text{geo}}(W, W')$, $\theta_i$ denotes the $i$-th principal angle between $W$ and $W'$. To be more specific, $\theta_i = \cos^{-1}(\sigma_i(M_W^\top M_{W'}))$ where $M_W, M_{W'} \in \mathbb{R}^{n \times k}$ are the column-wise concatenation of orthonormal basis for $W, W'$ and $\sigma_i$ represents the $i$-th singular value.

Experiments  To test the discriminability of these two metrics, we designed a simple experiment. Let $W, W'$ be the two 50-dimensional subspaces of $\mathbb{R}^{512}$ because the dimension of intermediate layers in the mapping network is 512. We measure the Grassmannian distance between two subspaces as we vary $\dim(W \cap W') = k_0$.

$$W = \langle e_1, e_2, \cdots, e_k \rangle, \quad W' = \langle \{e_1, e_2, \cdots, e_{k_0}\} \cup \{e_{k+1}, \cdots, e_{2k-k_0}\} \rangle$$  \hspace{1cm} (36)

where $\{e_i\}_{1 \leq i \leq n}$ denotes the standard basis of $\mathbb{R}^n$. Fig 8 reports the results. The Geodesic Metric reflects the degree of intersection between two subspaces. As we increase the dimension of intersection, the Geodesic Metric decreases. However, the Projection Metric cannot discriminate the intersected dimension until it reaches the entire space.

Figure 8: Grassmannian metric between two 50-dimensional subspaces $W, W' \in \text{Gr}(50, \mathbb{R}^{512})$ for each intersected dimension $k_0 = \dim(W \cap W')$. While the Geodesic Metric monotonically decreases as more dimensions intersect, the Projection Metric cannot discriminate $0 \leq k_0 \leq 49$. 
E. Robustness to preprocessing

In this section, we assessed the robustness of Distortion $D_M$ to preprocessing hyperparameter $\theta_{pre}$. Figure 9 presents the distribution of 1k samples of distortion before taking an expectation, i.e., $(d_{\text{geo}}^{k}(T_{w_1}, M_{w_1}^{k}, T_{w_2}, M_{w_2}^{k})/I_{local})$, for each intermediate layer. In Fig 9, increasing $\theta_{pre}$ makes an overall translation of Distortion. However, the relative ordering between the layers remains the same. The low Distortion score of layer 8 provides an explanation for the superior disentanglement of $\mathcal{W}$-space observed in many literatures (Karras et al., 2019; Härkönen et al., 2020). Moreover, the results suggest that the min-distorted layer 7 can serve as a similar-or-better alternative.

![Figure 9: Robustness of Distortion metric $D$ to $\theta_{pre}$ of StyleGAN2 on FFHQ.](image)

F. Architecture diagram of StyleGANs

![Figure 10: Architecture of StyleGANs. Our analysis in Sec 4 is performed in the intermediate layers of the mapping network.](image)
G. Robustness of Correlations to preprocessing

Figure 11: Correlation between Distortion metric and FID gap of StyleGAN2 on FFHQ. FID gap represents the difference between FID (Heusel et al., 2017) score of Local Basis (Choi et al., 2022b) and the global basis (Härkönen et al., 2020). Each FID score is measure for 50k samples of latent-perturbed images along the first component of Local Basis and GANSpace. The perturbation intensity is fixed to 3. Each point represents a $i$-th intermediate layer in the mapping network, and the red-line illustrates the linear regression of these points. $\rho$ denotes the Pearson correlation coefficient of Distortion and FID. The positive correlation between Distortion and FID remains robust regardless of $\theta_{pre}$. 

(a) $\theta_{pre} = 0.0005$, $\rho = 0.706$
(b) $\theta_{pre} = 0.001$, $\rho = 0.706$
(c) $\theta_{pre} = 0.005$, $\rho = 0.701$
(d) $\theta_{pre} = 0.01$, $\rho = 0.715$
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 12: Correlation between Distortion metric and FID gap of StyleGAN2 with config E on FFHQ.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 13: Correlation between Distortion metric and FID gap of StyleGAN2 on LSUN Cat (Yu et al., 2015).
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 14: **Correlation between Distortion metric and FID gap** of StyleGAN2 on LSUN Horse.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 15: Correlation between Distortion metric and FID gap of StyleGAN2 on LSUN Church.
Figure 16: **Correlation between Distortion metric and DCI** of StyleGAN2 on FFHQ. Each DCI (Eastwood & Williams, 2018) score is evaluated for 10k samples of generated images, while the attribute label is generated by 40 attribute classifiers pre-trained on CelebA (Liu et al., 2015). As in Fig 11, each point and red-line represents the intermediate layers and linear regression, respectively. The Distortion and DCI score show a negative correlation.
Figure 17: Correlation between Distortion metric and DCI of StyleGAN2 with config E on FFHQ.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 18: Correlation between Distortion metric and DCI of StyleGAN1 on FFHQ.
H. Additional Experimental results

Figure 19: Off-manifold Results in $\mathcal{W}$-space of StyleGAN2 on FFHQ. The red vertical lines denote the estimated local dimension for various $\theta_{pre}$. The small final loss implies that the linear perturbation along that Local Basis component stays inside the learned latent manifold. For every perturbation intensity, there is a transition point from a slow increase to a sharp increase, which is interpreted as an escape from the manifold. The results demonstrate that the proposed dimension estimation algorithm finds a reasonable point without crossing the transition point. In our case, these results are interpreted as choosing the principal part of the manifold without overestimating its dimension.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 20: **Local Dimension Distribution** of the intermediate layers in the mapping network of StyleGAN2 on FFHQ. Each figure presents the distribution of estimated local dimension at each intermediate layer as we vary $\theta_{pre}$. The distributions are illustrated for 1k samples, respectively. The algorithm gives a rather unstable dimension estimate on the most un sparse first layer (Fig 2a) due to its isotropic gaussian assumption. However, this phenomenon is not observed in the layers with moderate depth, i.e., from 3 to 8. As we introduce the higher preprocessing ratio $\theta_{pre}$, the algorithm gives more strict, i.e., smaller, dimension estimates. Nevertheless, the relative trend between layers is the same. The deeper the latent manifold, the smaller its dimension.
Figure 21: **Local Dimension Evaluation in Image Space** of StyleGAN2 on FFHQ. Figure 21b-21e show image traversals along the \((d - 1)\)-th and \(d\)-th axis where \(d\) denotes the estimated local dimension for each \(\theta_{pre}\). Fig 21f presents the image variation intensity \(\|\nabla_{\mathbf{v}_i} g(\mathbf{w})\|_F\) along each Local Basis \(\mathbf{v}_i\). \(\|\nabla_{\mathbf{v}_i} g(\mathbf{w})\|_F\) is evaluated by the finite difference with stepsize=0.01. Note that while Local Basis is discovered by analyzing only the subnetwork from the input to target latent space, the corresponding image variation monotonically decreases and saturates. Moreover, the estimated local dimension includes the major variations in the image space. This can be observed in the image traversals. The image traversals around the estimated dimension (Fig 21b-21e) show much smaller image variations compared to Axis 0 and 1 (Fig 21a).
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 22: **Linear Traversals along the various LB axis** on StyleGAN2-FFHQ. At this latent variable, the estimated dimension with $\theta_{pre} = 0.01$ is $d = 44$. When we traverse along the axis $i \gg d$, the images present minor variations compared to the axis $i \leq d$. Traversals along the $1 \rightarrow 44$ axes display various semantic variations such as gender, skin color, age, etc.
I. Comparison of Semantic Disentanglement

Using our Distortion metric, we chose the Max-distorted layer 3 and Min-distorted layer 7 from the mapping network of StyleGAN2-FFHQ (Karras et al., 2020b). These two layers are compared with the renowned $W$-space (layer 8). Each image traversal is obtained by perturbing a latent variable along the global basis from GANSpace (Härkönen et al., 2020).

![Image](image1.png)

(a) Max-distorted layer 3

(b) Min-distorted layer 7

(c) Layer 8 ($W$-space)

Figure 23: Comparison of Semantic Disentanglement between Layers 3, 7, 8 for the semantic "Open Mouth". Layer 7 exhibits a disentangled semantic variation of "Open mouth" without changing the face, such as a longer nose in Layer 8. Layer 3 shows only minor facial expression changes.

![Image](image2.png)

(a) Max-distorted layer 3

(b) Min-distorted layer 7

(c) Layer 8 ($W$-space)

Figure 24: Comparison of Semantic Disentanglement between Layers 3, 7, 8 for the semantic "Face Bright". Layer 8 shows an entangled semantic variation of "Lipstick" and "Face Bright" on the right end. On Layer 3, the traversal toward the right end does not show "Face Dark" semantics.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 25: **Comparison of Semantic Disentanglement between Layers 3, 7, 8** for the semantic "Gender". On Layer 3, the image traversal shows an entangled semantic variation of "Gender" and "Big face" on the right end. Also, Layer 8 exhibits an entangled semantic variation of "Gender" and "Age".

Figure 26: **Comparison of Semantic Disentanglement between Layers 3, 7, 8** for the semantic "Light Up". Layer 8 exhibits an entangled semantic variation of "Light Down" and "Brown hair" on the right end. On the left end of Layer 3, the traversal shows the "Bright right" transformation. However, the traversal at the right end of Layer 3 does not present the "Bright left" transformation, while showing no shadow on the face. But, the traversal on Layer 7 shows a natural brightness change on the entire face.
J. Additional Subspace Traversal Comparisons

Figure 27: Subspace Traversal \cite{choi2022local} on the min-distorted (7th) and max-distorted (3rd) intermediate layers along the global basis \cite{haarkonen2020understanding} and Local Basis \cite{choi2022local} of StyleGAN2 on FFHQ. The initial image (center) is traversed along the 1st (horizontal) and 2nd (vertical) components of the chosen traversal directions with the perturbation intensity 9. The global basis shows a decent image quality on the min-distorted layer, similar to Local Basis. However, on the max-distorted layer, the subspace traversal along global basis exhibits significant failures at corners, such as image collapse (lower-left), visual artifacts (lower-right), and unnatural transformations (top-left).
Figure 28: **Subspace Traversal** on StyleGAN2-FFHQ. The upper-left corner of layer 3 is severely deteriorated.

Figure 29: **Subspace Traversal** on StyleGAN2-LSUN Cat. The lower-left corner of layer 3 are severely deteriorated.

Figure 30: **Subspace Traversal** on StyleGAN2-LSUN Church. The upper-left corner of layer 3 are severely deteriorated.
Analyzing the Latent Space of GAN through Local Dimension Estimation

Figure 31: **Subspace Traversal** on StyleGAN2-LSUN Horse. The right sides of layer 3 are severely deteriorated.

Figure 32: **Linear Traversal** on StyleGAN2-LSUN Cars. The left sides of layer 3 are severely deteriorated.