Hierarchical Semantic Regularization of Latent Spaces in StyleGANs

Tejan Karmali[1,2*], Rishubh Parihar[1], Susmit Agrawal[1], Harsh Rangwani[1], Varun Jampani[2], Maneesh Singh[3†], and R. Venkatesh Babu[1]

[1] Indian Institute of Science, Bengaluru, India
[2] Google Research
[3] Motive Technologies, Inc

Abstract. Progress in GANs has enabled the generation of high-resolution photorealistic images of astonishing quality. StyleGANs allow for compelling attribute modification on such images via mathematical operations on the latent style vectors in the $\mathcal{W}/\mathcal{W}+$ space that effectively modulate the rich hierarchical representations of the generator. Such operations have recently been generalized beyond mere attribute swapping in the original StyleGAN paper to include interpolations. In spite of many significant improvements in StyleGANs, they are still seen to generate unnatural images. The quality of the generated images is predicated on two assumptions; (a) The richness of the hierarchical representations learnt by the generator, and, (b) The linearity and smoothness of the style spaces. In this work, we propose a Hierarchical Semantic Regularizer (HSR)[1] which aligns the hierarchical representations learnt by the generator to corresponding powerful features learnt by pretrained networks on large amounts of data. HSR is shown to not only improve generator representations but also the linearity and smoothness of the latent style spaces, leading to the generation of more natural-looking style-edited images. To demonstrate improved linearity, we propose a novel metric - Attribute Linearity Score (ALS). A significant reduction in the generation of unnatural images is corroborated by improvement in the Perceptual Path Length (PPL) metric by 16.19%% averaged across different standard datasets while simultaneously improving the linearity of attribute-change in the attribute editing tasks.

1 Introduction

Recent years have seen tremendous advances in Generative Adversarial Network (GAN) [16] architectures and their training methods to produce highly photo-realistic images [8, 30]. Progress in the StyleGAN family of GAN architectures has shown promise by improving both the image quality, as well as the quality

* Work done while at Indian Institute of Science
† Work done while at Verisk Analytics
[1] Project Page and code: https://sites.google.com/view/hsr-eccv22

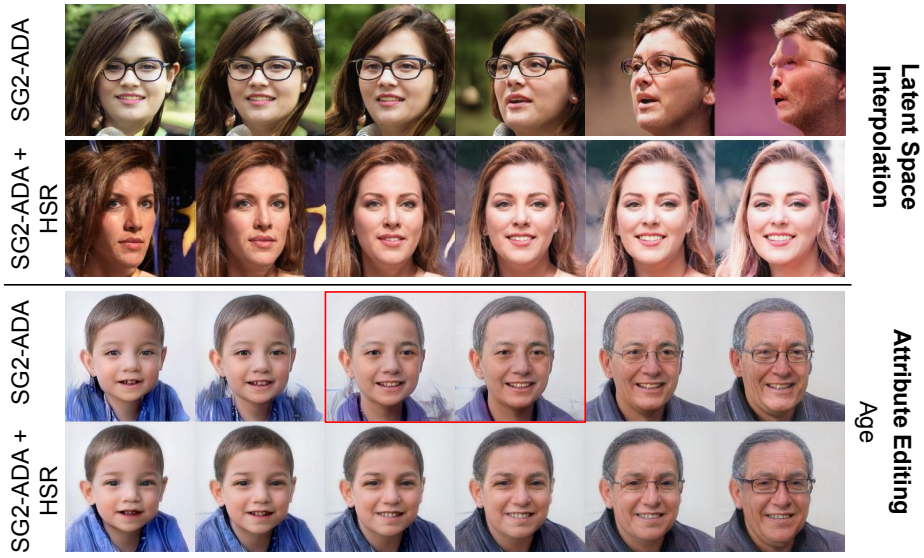


Fig. 1. Hierarchical Semantic Regularizer (HSR) improves the latent space to semantic image mapping to produce more natural-looking images. **Top:** We show latent interpolation for images from bottom 10%-ile image pairs ranked by PPL, a metric to measure smoothness of latent space. **Bottom:** Latent space using HSR mitigates artefacts in images during attribute edit transition and can transition smoothly (young to old (SG2-ADA) vs. young to middle-age to old (SG2-ADA+HSR), in continuous attributes like "Age"). Zoom in to observe the effects.

of latent space representations which enables controlled image generation. This is achieved by transforming an input noise space $\mathcal{Z}$ to a latent style space $\mathcal{W}$ which modulates a synthesis network at various levels of representation hierarchies to generate an image with that style. This enables generation of compelling synthetic images with novel styles as well as practically useful applications such as GAN-based image attribute editing, style mixing, etc. [3, 23, 43, 44, 47, 48]. Nonetheless, such networks still often produce unrealistic images (ref. Fig. 1).

These quality issues in StyleGANs can have the following sources: (a) the hierarchical representation spaces in the synthesis network, (b) the latent style space, in particular the linearity and smoothness of such spaces, and (c) the functions used to transform the representation spaces in (a) using the corresponding hierarchical style vectors in (b). Our work seeks to address some of these issues.

We take inspiration from the recent advances in self-supervised and supervised learning [9, 11, 20, 51] which have allowed for the learning of semantically rich image representations translating into significant performance improves on image classification and other vision tasks [15, 33, 51]. Training on large datasets of natural images, like ImageNet [13], allows these techniques to learn hierarchically organized feature spaces capturing richer statistical patterns in natural images: shallower layer capturing low-level image features and the deeper lay-

ers abstract features highly correlated with visual semantics. Such pre-trained representations can be harnessed to enhance the representational power of StyleGANs.

In fact, we demonstrate that transferring rich pretrained representations mentioned above allow us to mitigate simultaneously the challenges associated with both the representation spaces in the synthesis network as well as the latent style spaces modulating these representations. To allow for such a transfer, we propose to use a regularization mechanism, called the Hierarchical Semantic Regularizer (HSR) which aligns the generator's features to those from an appropriate, state of the art, pretrained feature extractor at several corresponding scales (levels) of the generator network. The architecture is shown in Fig. 3.

Karras *et al.* [29] introduced the Perceptual Path Length (PPL) metric to measure the smoothness of mapping from a latent space to the output image and showed its correlation with the generated image quality. We demonstrate that HSR regularization in StyleGAN training leads to 16.19% relative improvement in PPL over StyleGAN2, leading to more realistic interpolations (refer Fig. 1).

A standard approach for controlled synthesis of novel images is via linear (convex) interpolation between attributes[4] corresponding to real images. Applications such as image editing utilize such capabilities under the presumption that style spaces are both linear as well as decorrelated allowing for desired controlled edits. Since, PPL does not measure linearity, we propose a novel metric, Attribute Linearity Score (ALS), to measure linearity in the attibute space. We demonstrate that HSR simultaneously improves linearity leading to smoother edits with significantly reduced editing artifacts (Fig. 1). A mean relative improvement of 15.5% over StyleGAN2-ADA is achieved on the ALS metric.

Our contributions are: (a) A novel Hierarchical Semantic Regularizer (HSR) improving the generation of natural-looking synthetic images from StyleGANs. HSR is presented in (Sec. 3 with an analysis of design choices 3.3). (b) Extensive bench-marking of improvements by HSR regularization on popular datasets, especially when utilizing linear interpolations for attribute editing (Sec. 4). (c) Since linearity of the latent attribute space is very important for performing controlled edits, we propose a new metric, Attribute Linearity Score (ALS), in (Sec. 4.3) and demonstrate improved linearity over the baselines.

2 Related Works

Generative Adversarial Networks. GAN proposed by Goodfellow *et al.* [16] a combination of two neural networks, *i.e.* generator G and discriminator D. For image synthesis the goal of D is to differentiate between real and generated images, whereas the G tries to fool the discriminator into classifying generated images as real. In the recent years several improvements in architecture [18,28,31, 39,46], optimization objectives [5,7,36,37] and regularization [19,38] have made GANs an ubiquitous choice for image synthesis. It has been observed that GANs

[4] We use style and attributes interchangeably.

developed for large scale datasets, suffer mode collapse when trained on limited data. Augmentation methods like DiffAugment [63], ADA [27], ContraD [24] , APA [25] etc. mitigate the collapse by reducing the discriminator's overfitting.

Hierarchical Representations. In classical vision, methods which decompose image into a hierarchy have been exploited for the tasks of image stitching, manipulation and fusion [2, 12]. Building on this motivation Shocher *et al.* [49] develop an image translation and manipulation method, which exploits hierarchical consistency of features of generator and a classifier. However this method is restricted to single image translation and manipulation. In contrast our work we aim to train a smooth and generalizable GAN which can simultaneously generate diverse images, by using semantic hierarchical consistency of features.

Knowledge Transfer Using Pre-Trained Features. Using pre-trained features trained on large scale datasets (*e.g.* ImageNet etc.) [21,52] have been useful for various downstream tasks across applications [14,22,54,58]. The recent development of the self-supervised approaches for representation learning [10, 17, 46] have further immensely improved the quality of features learnt. These features are being used in various applications like part segmentation, localization etc. without being explicitly trained on such tasks [9], which motivates our work which aims to transfer these semantic properties to G's feature space. Currently much work for transfer learning for GANs has focused on the fine-tuning large GANs using a few images for adapting it to a different domain [35, 40–42]. Recently a concurrent work [34] also aims to use pre-trained features to improve GANs. However their goal is to improve discriminator. On contrary we aim to enrich GAN feature space by imparting it with semantic properties, leading to a disentangled and smooth latent space.

Image Editing Using Latent Space Interpolations. Latent space of pre-trained StyleGAN models is highly structured [47] and is popularly used to perform realistic image edits in the generated images [1, 4, 23, 47, 48, 56, 60]. The primary idea in most of these approaches is to find a direction in the the extended latent space $\mathcal{W}+$ for editing attributes and transforming a latent code by moving in that direction to perform edits. StyleCLIP [44] learns the directions for attribute editing by getting the guidance from pretrained CLIP [45]. On the contrary, our work imposes constraints so that latent space has more naturally interpretable directions when used by the GAN-based image editing methods.

3 Approach

In this section, we first describe the objective of GAN framework, properties of StyleGAN, and its evaluation in Sec. 3.1. Then, we describe Hierarchical Semantic Regularizer (HSR) (Sec. 3.2) and discuss its design in Sec. 3.3.

3.1 Preliminaries

Generative Adversarial Networks. GAN involves two competing networks, namely a Generator G and a Discriminator D. Taking a noise $\mathbf{z}$ sampled from

a distribution P_z as input, G generates an image $G(\mathbf{z}) \in \mathbb{R}^{3\times H\times W}$. Whereas, D takes an input image $x \in \mathbb{R}^{3\times H\times W}$, and tries to classify it as real or generated. The objective of G is to fool D into making it classify the generated image as a real one. Formally, the learning objective can be written as:

$$\begin{aligned} \max_D \mathcal{L}_D &= \mathop{\mathbb{E}}_{x\sim P_r}[log(D(x))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[log(1 - D(G(\mathbf{z})))] \\ \min_G \mathcal{L}_G &= \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[log(1 - D(G(\mathbf{z})))] \end{aligned} \tag{1}$$

StyleGAN. In StyleGAN, an architectural modification is introduced where $\mathbf{z}$ is transformed into a semantic latent space through a sequence of linear layers called Mapping Network G_m, before generating the image I through a Synthesis Network G_s as $I = G_s(\mathbf{w})$. Hence, $G = G_s \circ G_m$. The space learnt by G_m is known as $\mathcal{W}+$-space. It is observed that $\mathcal{W}+$ is more meaningful in terms of attributes learned from the training data as compared to noise space $\mathcal{Z}$. Several methods [23,47,48] propose ways to find attribute-specific directions in $\mathcal{W}+$ latent space.

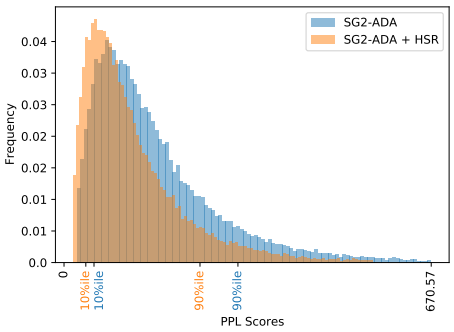


Fig. 2. Distribution of PPL over 50k images from SG2-ADA and SG2-ADA+HSR. HSR improves the perceptual quality of top and bottom 10%-ile images, thus leading to more natural-looking images.

Perceptual Path Length. To measure the smoothness of the mapping from a latent space to the output image, Karras *et al.* [30] proposed Perceptual Path Length (PPL). The requirement for this metric arises due to generation of unnatural images by GAN despite having low FID [30]. PPL aims to quantify the smoothness of latent space to output space mapping by measuring average of LPIPS [61] distances between two generated images under small perturbations in the latent space. A smoother latent space should have lesser PPL when compared to an uneven latent space. It is shown [30] that PPL correlates well with image quality, *i.e.* good quality images pairs will have less PPL, while if any one of the image is of bad quality, the PPL would be high. The images are sampled randomly without any truncation trick [8,32] to compute PPL. As observed in Fig. 6, the bottom 10%-ile by PPL (sorted in increasing order) among the generated images appear as out-of-distribution images. Hence, the mean PPL score can be used to quantify the extent of non-smooth regions of latent space which produce unnatural images. Hence, we will be using this metric as a primary metric for comparison of the smoothness of latent space learnt by the models.

3.2 Hierarchical Semantic Regularizer

Feature extractors of networks pretrained on large datasets (*e.g.* ImageNet etc.) of natural images using classification or self-supervised losses store strong priors

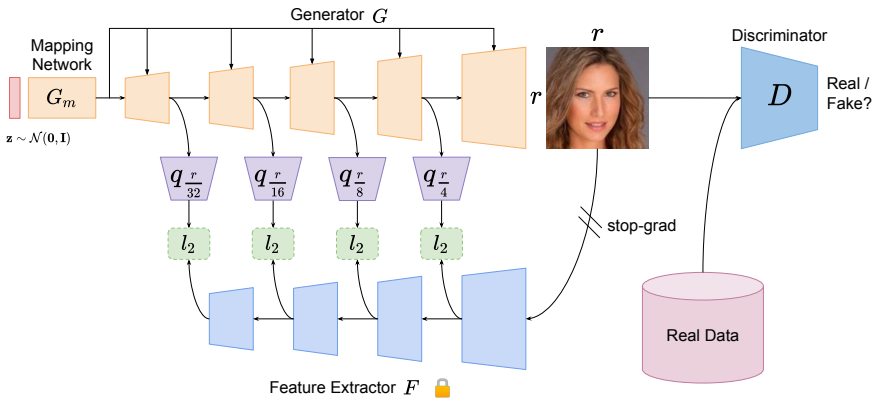


Fig. 3. Hierarchical Semantic Regularizer: We use a pre-trained network to extract features at various resolution hierarchically. We then train linear predictors over generator features to predict the pre-trained features hierarchically. This transfers the semantic knowledge to generator feature space, making it's latent space meaningful, disentangled and editable.

about the data, that are organized hierarchically. Each level of hierarchy captures a different semantic feature of data. The statistics of wide variety of natural images are captured by these networks [6,21,53]. Due to the inherent differences in the nature of tasks, discriminative models capture different kinds of features compared to generative ones. Therefore, we seek to enrich the G's intermediate feature space with guidance from a pretrained feature extractor.

We first give a general idea of the proposed regularizer and then dive into various design choices made in it's formulation. Given an image $\mathbf{x}$ as input, the feature extractor F returns semantically meaningful features from it. We attempt to make the generator aware of this explicit semantic feature space. To this end, we freeze the feature extractor and treat it as a fixed function that maps from image space to a semantically meaningful feature space.

Given such a mapping of the generated image, we try to align the Generator's features of this image through a set of feature predictors. This alignment is inspired by BYOL [17]. As illustrated in Fig. 3, we attach a predictor branch q to the Generator G. The objective of q is to learn a mapping from generator's intermediate feature map $G^{\pi^i_G}(\mathbf{z})$ to pretrained feature extractor's intermediate feature $F^{\pi^i_F}(G(\mathbf{z}))$, where π_G in π_F denote the ordered set of layer numbers in the G and F at which we attach the predictors (ref. Eq. 2). We attach multiple such predictor networks q_i at different scales of generator.

$$\mathcal{L}_G = \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[log(1 - D(G(\mathbf{z})))] + \sum_{i=0}^{|\pi^i_G|} \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z} \|q(G^{\pi^i_G}(\mathbf{z})) - F^{\pi^i_F}(G(\mathbf{z}))\|_2^2 \quad (2)$$

3.3 Design Choices

We analyse the effect of our Hierarchical Semantic Regularizer (HSR) against different design choices. For this purpose, we choose AnimalFace-Dog dataset which consists of 389 images. Since this is a low-shot dataset, we use StyleGAN2-ADA as our baseline. We perform all our experiments on 256×256 resolution.

What should be the choice of feature extractor? For this analysis, we choose 5 different feature extractors. We take combinations of CNN or transformer based networks trained using either self-supervised or supervised classification objective. We take ResNet-50 as the CNN backbone for both self-supervised (DINO) and supervised networks. For transformer-based networks, we use ViT-DINO and DeiT. Apart from trained networks, we also consider a randomly initialized ViT for baseline comparison.

We find that all pretrained feature extractors when used through HSR loss lead to introduction of meaningful semantic features in the intermediate latent spaces of the Generator. This is evidenced by reduction of PPL Score in Table 1, which signifies reduction in non-meaningful generations from the GAN. The reduction in PPL also implies improved disentanglement [30] and linearity in the $\mathcal{W}$ space of the Generator, which is a desired property for many applications. We get $\geq 6.2\%$ improvement in the PPL score when guided by these networks. ViT DINO's features stand apart, by improving the PPL score by 19% over the baseline. This is also supported by recent findings of Amir *et al.* [6], where they show several inherent properties of features from ViT-DINO, that are useful for computer vision tasks. With these results, we fix ViT DINO as the choice of the feature extractor for the rest of the experiments.

Which layers of Generator are more important? The StyleGAN generator G generates images using 7 synthesis blocks: starting from 4×4, up to full resolution of 256×256. Of these, we consider synthesis blocks having features of resolution 8, 16, 32, 64. This corresponds to scaling down of resolution r to $\frac{r}{32}$, $\frac{r}{16}$, $\frac{r}{8}$, and $\frac{r}{4}$. We choose these scales as it largely corresponds to the scales of downsampling by each block in SoTA CNN architectures [20,51]. The first block of G (which have low resolution, but are responsible for high-level semantics) are supervised by the last block of the feature extractor (as they also are responsible for high-level semantics). Similarly, other blocks of G are supervised by the blocks of the feature extractor that bring out similar level of semantics.

To decide which layers contribute the most to the improvement in PPL, we divide the 4 blocks into 3 groups. The 3 groups specialize in high ($\frac{r}{32}$, $\frac{r}{16}$), mid ($\frac{r}{16}$, $\frac{r}{8}$), and low ($\frac{r}{8}$, $\frac{r}{4}$) level of semantics. We observe, in Table 2, that it is the supervision at low-level semantics which is most useful for the G. We observe a gradation in the improvement over the baseline, as high-level semantic supervision is least useful, followed by middle, and low. Overall, supervision at all levels turns out to cause the highest improvement.

Does Path Length Regularizer (PLR) complement HSR? Path Length Regularizer (PLR) was introduced in StyleGAN2 [30]. The intuition behind PLR is to promote fixed magnitude non-zero change in the resulting image when moving by a fixed step size in the $\mathcal{W}+$-space. As reported in Table 4, we find that

Table 1. Feature space ablation: Ablating over different feature extractors for usage in HSR. HSR with ViT-DINO's features gives best results.

	FID ↓	PPL↓
StyleGAN2-ADA	53.28	59.27
+ ViT (RandInit)	53.65	56.97
+ ResNet50 DINO [9]	54.33	55.6
+ DeiT [53]	53.22	54.71
+ ResNet50 [20]	52.88	52.23
+ ViT DINO [9]	**51.58**	**48.02**

Table 2. Level of semantics: A gradation in the improvement over the baseline is observed as we supervise from high-level semantics to low-level semantics. Best results are obtained when all the levels are supervised.

	FID↓	PPL↓
StyleGAN2-ADA	53.28	59.27
+ High-level ($\frac{r}{32}$, $\frac{r}{16}$)	53.15	57.73
+ Mid-level ($\frac{r}{16}$, $\frac{r}{8}$)	52.91	54.18
+ Low-level ($\frac{r}{8}$, $\frac{r}{4}$)	53.66	51.77
+ All levels	**51.58**	**48.02**

Table 3. FFHQ-140k Results: We report FID, Precision, Recall and PPL for different methods. With large data our method (SG2+HSR) produces better results even compared to SG2-ADA.

FFHQ-140k	FID↓	Precision↑	Recall↑	PPL↓
SG2	3.92	**0.68**	0.45	175.09
+ HSR	**3.74**	**0.68**	**0.48**	**144.59**
SG2-ADA	**4.30**	0.69	**0.40**	163.11
+ HSR	5.26	**0.70**	0.38	**131.41**

Table 4. Performance wrt PLR. PLR and HSR complement each other, while being equally effective individually.

PLR	HSR	FID↓	PPL↓
✗	✗	57.97	75.63
✓	✗	53.28	59.27
✗	✓	52.98	58.60
✓	✓	**51.58**	**48.02**

HSR itself gives slightly better improvement than the PLR over the baseline. While the best effect is noted when both, PLR and HSR, are applied together. **Insight.** PLR's objective is to improve latent space smoothness, which leads to better PPL. Since PPL and image quality (natural-ness of image) are correlated, applying PLR improves the image quality. Whereas in HSR, we enforce the generator to predict in a feature space learnt from natural images using a pretrained feature extractor as prior. We observe that this objective, which targets bringing feature space of generator closer to a "natural" feature space also leads to improvement in the smoothness of latent space, as measured by PPL. This shows that image quality and latent space smoothness are complementary and related concepts. Therefore, optimizing for both gives better PPL score.

4 Experiments

In this section, we demonstrate the effectiveness of HSR experimentally. We first describe the experimental setup for all our experiments. Then, we evaluate the quantitative performance on several real-world datasets of varying sizes. Finally, we show improved linearity of latent space through attribute editing.

Table 5. Results on Limited Data We present results on different limited data cases for FFHQ (left) dataset and on real-world datasets (right). We apply our regularizer on the strong baseline of StyleGAN2+ADA which is designed for limited data. We observe a significant decrease in PPL over baselines which implies a smooth, disentangled and meaningful latent space, while preserving photorealism (comparable FID).

Dataset	Method	FID↓	PPL↓	Dataset	Method	FID↓	PPL↓
FFHQ-1k	StyleGAN2-ADA	**19.14**	98.79	AnimalFace Dog	StyleGAN2-ADA	53.28	59.27
	+ HSR	21.76	**90.39**		+ HSR	**51.58**	**48.02**
FFHQ-2k	StyleGAN2-ADA	**14.74**	136.14	AnimalFace Cat	StyleGAN2-ADA	**39.50**	50.76
	+ HSR	15.53	**115.38**		+ HSR	40.25	**40.75**
FFHQ-10k	StyleGAN2-ADA	**7.16**	164.21	CUB	StyleGAN2-ADA	**5.78**	265.46
	+ HSR	8.08	**126.35**		+ HSR	6.15	**237.81**

4.1 Experimental Setup

Datasets. We run our experiments on FFHQ [26] (70k images), AnimalFace-Dog (389 images), AnimalFace-Cat (160 images) [50], CUB200 (12k images) [55], and LSUN-church [59] (126k images) (ref. supl. mat.) datasets. We augment the datasets by taking the horizontal flip of every image, doubling the number of images in the original dataset. **Implementation Details.** We use StyleGAN2-ADA (SG2-ADA) as the baseline GAN, with its architecture for 256×256 images, with batch size of 16. Predictors q contain `Conv1x1-LeakyReLU-Conv1x1`, with hidden dimension of 4096. We make use of 2 A6000 GPUs for training our models.

4.2 Results

On standard full dataset of FFHQ, we compare over both StyleGAN2 [30] (SG2) and StyleGAN2-ADA [27]. This is because StyleGAN2 shows slightly better performance against the ADA variant on large datasets. We also evaluate our method for limited data sizes. Traditionally, GANs have shown to perform poorly on smaller datasets, until recently several approaches [25, 27, 57] have been proposed which enables GANs to learn well on limited data. We observe that irrespective of dataset size, asking the generator to be predictive of semantic features of rich feature extractors via HSR improves the smoothness of the latent space, as it is evident by an average relative improvement in PPL scores of about 14.2% on average in Table 5, while that of 17.42% in case of full FFHQ. This is also evident qualitatively in Fig. 4 and 6, where we observe an improved latent-to-image mapping even in bottom 10%-ile images, when ranked by PPL scores. We also present images sampled randomly in Fig. 5, where we observe the mitigation of the unnatural faces and artefacts (highlighted images) upon application of HSR. Thus, HSR raises the lower bound for the natural-ness of the images produced by a generator (also ref. Fig. 2).

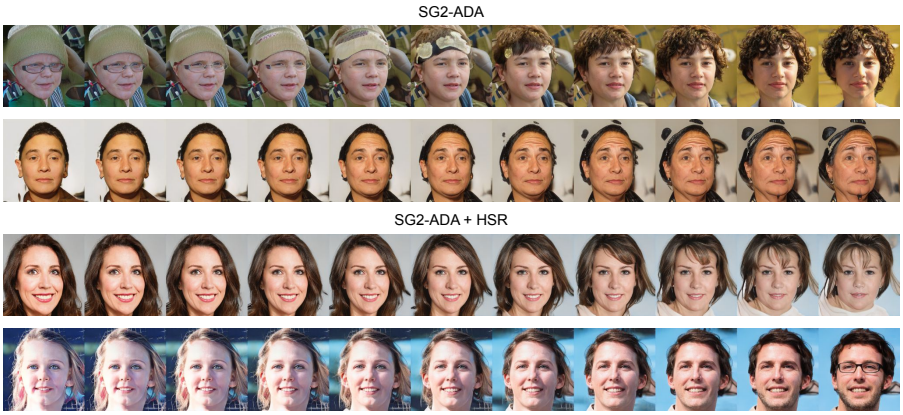


Fig. 4. Latent space interpolation of top 10-%ile images, ranked by PPL score. SG2-ADA images show traces of artifacts which are absent after applying HSR.

4.3 Analysis of Linearity of Latent Space

Motivation. Latent space of a pre-trained StyleGAN has meaningful directions embedded in it. Shen *et al.* [47] shows that $\mathcal{W}+$ latent space is disentangled with respect to image semantics and there exist linear directions $\mathbf{d}$ in this space that control specific semantic attributes in the generated images. This is an important property of the latent space which is commonly used in controlled image synthesis [44] and image editing [1], as it leads to smooth interpolation between any two generated images. Furthermore, it is observed that the magnitude of latent transformations linearly correlates with the magnitude of the attribute changes in generated images [62]. Although, multiple works [1,23,44,47,56] are built upon this property to generate desired image transformations, there is no established metric to evaluate the extent of this linear correlation in the latent space. To this end, we propose a new metric called *Attribute Linearity Score (ALS)* for quantifying this linear correlation between the extent of latent transformations and the attribute changes.

Attribute Linearity Score (ALS). Let the attribute strength be given by attribute score (logit value) from a pretrained attribute classifier C [29]. Consider two latent codes $\mathbf{w_0}$ and $\mathbf{w_1} \in \mathcal{W}+$ and their corresponding generated images $G(\mathbf{w_0})$ and $G(\mathbf{w_1})$ (using the generator G). Convex combinations of $\mathbf{w_0}$ and $\mathbf{w_1}$ generate interpolated latent codes $\mathbf{w_t}$ (Eq. 3) on the line segment joining the two latent codes $\mathbf{w_0}$ and $\mathbf{w_1}$. Let the corresponding generated images be denoted by $G(\mathbf{w_t})$. Linearity of the latent space [62] with respect to the attribute strength C implies that the attribute score for the image $G(\mathbf{w_t})$ should be the same convex combination of the attribute strengths of $G(\mathbf{w_0})$ and $G(\mathbf{w_1})$ (Eq. 4).

$$\mathbf{w_t} = \mathbf{w_0} + t * (\mathbf{w_1} - \mathbf{w_0}),\ t \in (0, 1) \tag{3}$$

$$C(G(\mathbf{w_t})) \approx C(G(\mathbf{w_0})) + t * (C(G(\mathbf{w_1})) - C(G(\mathbf{w_0}))),\ t \in (0, 1) \tag{4}$$

Fig. 5. Comparison over uniformly random sampled images from StyeGAN2 and StyleGAN2+HSR. StyleGAN2 produces unnatural faces and artefacts over it such as peculiar eyeglasses (as shown in the highlighted images).

Consider the example shown in Fig. 7a, where we depict the transformation of the smile attribute. On the left, we show the plot of attribute scores with the interpolation parameter t using smile classifier C_s and on the right we show the image samples $G(\mathbf{w_t})$ for $t \in (0, 1)$. A model with a linear latent space structure should have this plot close to the "ideal" (shown in dotted) straight line between the two end points. Similar plots are shown for the "smile" and "male" attributes in Fig. 7b. In both cases, we observe a significant departure from linearity.

The ALS score quantifies the deviation from the line segment defined in Eq. 4 using the mean squared error metric. To compute this, we first define a set of equally-spaced interpolation points $t \in \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$. For each attribute $j \in \{1, \ldots, M\}$), the squared difference (Δ_{tj}) is computed using Eq. 5. The ALS score (Δ_T) is defined as the mean of Δ_{tj} over all M attributes and N interpolation points (*i.e.* $\Delta_T = \frac{1}{NM} \sum_{t=1}^{N} \sum_{j=1}^{M} \Delta_{tj}$).

$$\Delta_{tj} = ||C_j(G(\mathbf{w_t})) - C_j(G(\mathbf{w_0})) - t * (C_j(G(\mathbf{w_1})) - C_j(G(\mathbf{w_0})))||^2 \quad (5)$$

In the following sections, we first evaluate effect of linearity on appplying HSR, by measuring ALS. Then we show it's application in measuring edits in images. We use StyleGAN2-ADA model as the baseline trained on FFHQ-10k for results in the rest of this section.

ALS Evaluation. Our proposed HSR is able to provide a smooth structure to the latent space which is evident by the lower ALS scores of our model. To further analyse the structure of the latent space we perform latent space interpolations and generate a sequence of images. To quantitatively evaluate the interpolation results, we used the proposed ALS scores for the interpolations. The lower ALS score represent the latent space is well structured and the magnitude of the attributes are linearly correlated with the latent transformation. The ALS scores for our model and baseline model in Table. 8 for following set

Fig. 6. Latent space interpolation of bottom 10-%ile images, ranked by PPL score. SG2-ADA latent space accommodates more unnatural images, leading to increase in PPL score. The latent space maps to more natural face-like images in SG2-ADA-HSR.

of popular attributes {gender, smile, age, hair, bangs, beard} [23, 47, 48]. Additionally, Fig. 8 (right) shows the variation of the mean attribute delta ($\Delta_{t,\cdot}$) with the interpolation parameter t. We can observe that in the middle region $t \in [0.4, 0.8]$ the baseline model has high deviation from linear behaviour, which is significantly less in our HSR regularized model. This is also seen quantitatively through proposed ALS-attribute score, in which our model outperforms baseline by **15%** of relative improvement. We can observe that the interpolations generated using the HSR results in smooth transitions and has high visual quality throughout the interpolation. The StyleGAN2+ADA model without HSR has sudden transitions in between and has some artifacts present (ref. Fig. 1).

Editability. The semantically rich structure of the latent space is widely used for performing semantic edits on the generated images [1, 4, 44, 47, 56, 60]. For instance, if we have to add the attribute smile to the generated face image, one can edit the latent code as $\mathbf{w_{edit}} = \mathbf{w} + \alpha\mathbf{d}$ where α is edit strength and $\mathbf{d}$ is the direction for the smile attribute edit operation. However, often, the attribute scores of the edits performed by such methods does not change linearly with the edit strength parameter α as observed in Fig. 9. To this end, we perform the following experiment: Given an input source image I_0, we first perform attribute edit on it using latent space transformation to obtain I_1 using an existing approach [47]. Then, we use the latent code optimization to find the corresponding latent codes $\mathbf{w_0}$ and $\mathbf{w_1}$ in the latent space. Finally, we followed the same approach explained in Sec. 4.3 to generate intermediate images I_t using w_t for $t\{\in 0, \frac{1}{N}, \frac{2}{N}, ...1\}$. The results of the interpolation for edits are shown in Fig. 9. We compared StyleGAN2-ADA with and without HSR in this experiment. One can observe that in all the cases, adding HSR resulted in added linearity in the attribute scores plots. This property is highly desired in editing methods as it provides a fine-grained control over the attributes in the generated im-

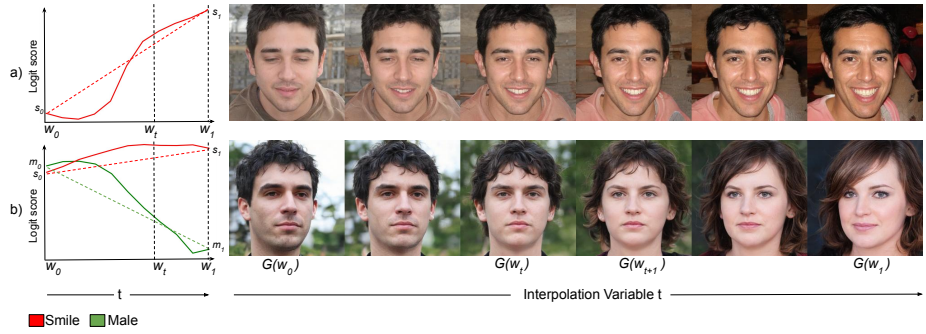
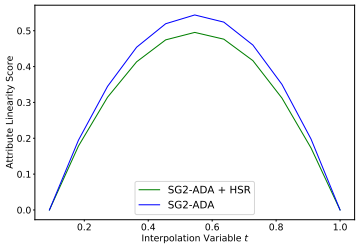


Fig. 7. Linearity of the latent space: Here we show the transition images generated by the intermediate latent code $\mathbf{w_t}$ in the right and the corresponding attribute scores s_t for for smile (row 1 and 2) and m_t for male attribute (row 2). For brevity we have written $s_t = C_s(G(\mathbf{w_t}))$ and $m_t = C_m(G(\mathbf{w_t}))$.

Fig. 8. ALS score comparison upon adding HSR. *(Right)*: Mean ALS computed for each value of the interpolation variable t. HSR is able to achieve a lower value of ALS supporting the linearity induced by ALS. *(Bottom)*: ALS score computed for all the face attributes separately.



	Gender	Smile	Age	Hair	Bangs	Beard	Mean
SG2-ADA	1.38	1.48	1.18	1.96	1.95	1.60	1.59
+HSR	**1.12**	**0.99**	**1.15**	**1.87**	**1.62**	**1.16**	**1.32**

ages. Also, observe that both the models evaluated are following the linear line closely in the first two examples. This suggests that the transitions along the age attribute is much more interpretable as it follows linearity. In all the three examples the model with HSR is able to approximate the linear line the better than the baseline without HSR. From the images, we can visually observe that the interpolations produced smooth transitions and their is no sudden jump in the attribute when using HSR. Also note that, the first and last images from both the models do not match "pixel perfectly", as they is generated by optimization of latent code by different models (with and without HSR). Note that HSR improves the reconstruction quality of real images when embedded in the latent space using projection (ref. supl. mat.).

5 Conclusion

We proposed a novel, hierarchical semantic regularizer called HSR which allows us to regularize the latent representations in StyleGANs by aligning them to semantically rich ones learnt by state-of-the-art classifiers trained on large datasets.

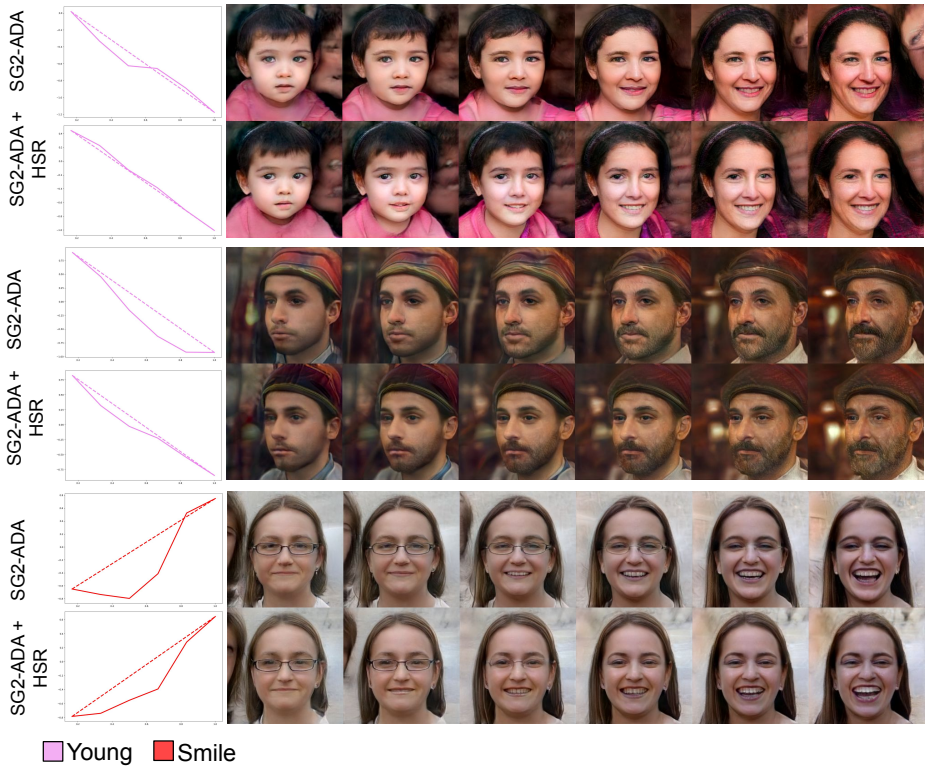


Fig. 9. Applying HSR improves the linearity of change in attributes. Here we show improved linearity for "Young" and "Smile" attributes. Plots show attribute score on Y-axis, interpolation variable t on X-axis.

HSR is shown to significantly improve the quality of the generated images, especially those created via linear interpolation between attributes corresponding to real images. It further has a desirable property that the latent attribute space becomes more linear. To measure linearity, a novel metric *Attribute Linearity Score (ALS)* was introduced. Copious experiments on standard benchmarks validate the benefits of HSR and demonstrate statistically significant improvement in the quality of synthesized images. This leads us to interesting avenues for the future work: Enforcing structural priors (*e.g.* linear) in the latent space while training a GAN, which can lead to easier and fine-grained attribute editing.

Acknowledgements. This work was supported by MeitY (Ministry of Electronics and Information Technology) project (No. 4(16)2019-ITEA), Govt. of India and Uchhatar Avishkar Yojana (UAY) project (IISC_010), MHRD, India.

References

1. Abdal, R., Zhu, P., Mitra, N.J., 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) 4, 10, 12
2. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA engineer **29**(6), 33–41 (1984) 4
3. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. ACM Trans. Graph. **40**(4) (2021) 2
4. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. ACM Transactions on Graphics (TOG) **40**(4), 1–12 (2021) 4, 12
5. Albuquerque, I., Monteiro, J., Doan, T., Considine, B., Falk, T., Mitliagkas, I.: Multi-objective training of generative adversarial networks with multiple discriminators. In: ICML (2019) 3
6. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. CoRR **abs/2112.05814** (2021), https://arxiv.org/abs/2112.05814 6, 7
7. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein Generative Adversarial Networks. In: ICML (2017) 3
8. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019) 1, 5
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2, 4, 8
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 4
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 2
12. Choudhary, B.K., Sinha, N.K., Shanker, P.: Pyramid method in image processing. Journal of Information Systems and Communication **3**(1), 269 (2012) 4
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 2
14. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML (2014) 4
15. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 27 (2014) 1, 3
17. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems (2020) 4, 6
18. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017) 3

19. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: NeurIPS (2017) 3
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 2, 7, 8
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 4, 6
22. Huh, M., Agrawal, P., Efros, A.A.: What makes imagenet good for transfer learning? arXiv preprint arXiv:1608.08614 (2016) 4
23. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. In: Proc. NeurIPS (2020) 2, 4, 5, 10, 12
24. Jeong, J., Shin, J.: Training GANs with stronger augmentations via contrastive discriminator. In: International Conference on Learning Representations (2021) 4
25. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Deceive D: Adaptive Pseudo Augmentation for GAN training with limited data. In: NeurIPS (2021) 4, 9
26. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018) 9
27. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020) 4, 9
28. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS (2021) 3
29. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 3, 10
30. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 1, 5, 7, 9
31. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020) 3
32. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: Advances in Neural Information Processing Systems (2018) 5
33. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: CVPR (2019) 2
34. Kumari, N., Zhang, R., Shechtman, E., Zhu, J.Y.: Ensembling off-the-shelf models for gan training. arXiv preprint arXiv:2112.09130 (2021) 4
35. Liu, B., Zhu, Y., Song, K., Elgammal, A.: Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In: International Conference on Learning Representations (2020) 4
36. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: CVPR (2017) 3
37. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: ICML (2018) 3
38. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018) 3
39. Miyato, T., Koyama, M.: cgans with projection discriminator. In: ICLR (2018) 3
40. Mo, S., Cho, M., Shin, J.: Freeze the discriminator: a simple baseline for fine-tuning gans. In: CVPRW (2020) 4
41. Noguchi, A., Harada, T.: Image generation from small datasets via batch statistics adaptation. In: ICCV (2019) 4
42. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: CVPR (2021) 4

43. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
44. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: ICCV (2021) 2, 4, 10, 12
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4
46. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016) 3, 4
47. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE TPAMI (2020) 2, 4, 5, 10, 12
48. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: CVPR (2021) 2, 4, 5, 12
49. Shocher, A., Gandelsman, Y., Mosseri, I., Yarom, M., Irani, M., Freeman, W.T., Dekel, T.: Semantic pyramid for image generation. In: CVPR (2020) 4
50. Si, Z., Zhu, S.C.: Learning hybrid image templates (hit) by information projection. PAMI **34**(7), 1354–1367 (2011) 9
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 2, 7
52. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019) 4
53. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (2021) 6, 8
54. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017) 4
55. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds-200-2011 (cub-200-2011). Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 9
56. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 4, 10, 12
57. Yang, C., Shen, Y., Xu, Y., Zhou, B.: Data-efficient instance generation from instance discrimination. In: Advances in Neural Information Processing Systems (2021) 9
58. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NeurIPS (2014) 4
59. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 9
60. Yüksel, O.K., Simsar, E., Er, E.G., Yanardag, P.: Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 4, 12
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 5
62. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: ICLR (2021) 10
63. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient GAN training. In: NeurIPS (2020) 4