Discovering Density-Preserving Latent Space Walks in GANs for Semantic Image Transformations

Guanyue Li
South China University of Technology,
Guangzhou, P. R. China.
csliguanyue007@mail.scut.edu.cn

Yi Liu, Xiwen Wei
South China University of Technology,
Guangzhou, P. R. China.
{csyiliu,xwwei123}@gmail.com,

Yang Zhang
City University of Hong Kong,
Kowloon, Hong Kong.
yzhang3233-c@my.cityu.edu.hk

Si Wu*
South China University of Technology,
Guangzhou, P. R. China
cswusi@scut.edu.cn

Yong Xu
South China University of Technology,
Guangzhou, P. R. China
yxu@scut.edu.cn

Hau-San Wong*
City University of Hong Kong,
Kowloon, Hong Kong.
cshswong@cityu.edu.hk

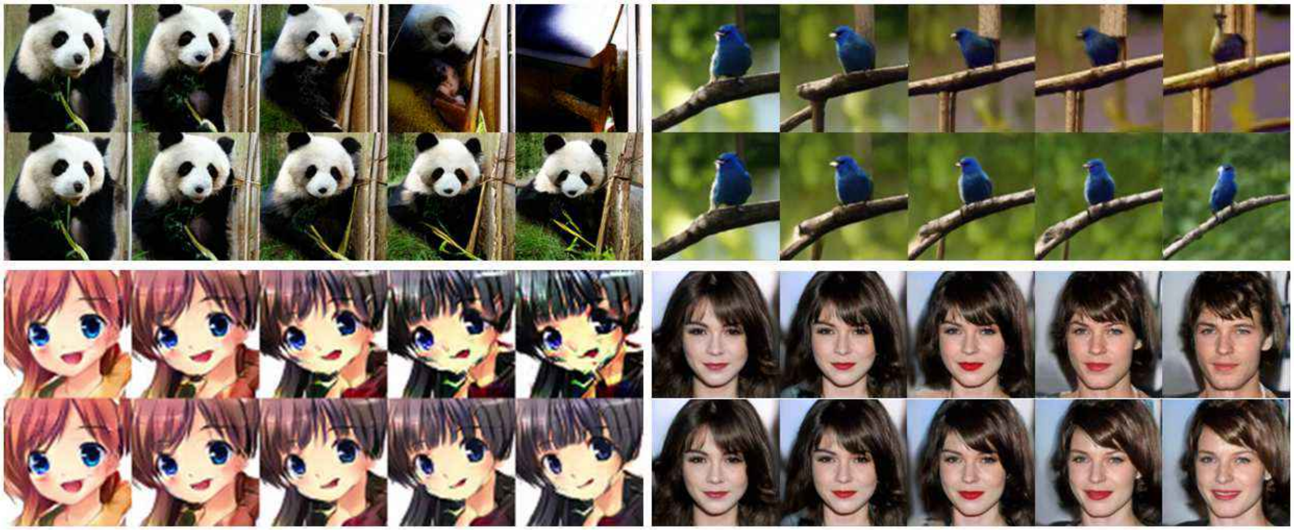


Figure 1: Examples to illustrate the effectiveness of the proposed Density-Preserving Regularization (DP Reg.) approach in stabilizing the transformation that is performed by latent space walks. In each example, the images are synthesized (*upper row*) w/o DP Reg. and (*bottom row*) w/ DP Reg..

ABSTRACT

Generative adversarial network (GAN)-based models possess superior capability of high-fidelity image synthesis. There are a wide range of semantically meaningful directions in the latent representation space of well-trained GANs, and the corresponding latent space walks are meaningful for semantic controllability in the synthesized images. To explore the underlying organization of a latent space, we propose an unsupervised Density-Preserving Latent Semantics Exploration model (DP-LaSE). The important latent directions are determined by maximizing the variations in intermediate features, while the correlation between the directions is minimized. Considering that latent codes are sampled from a prior distribution, we adopt a density-preserving regularization approach to ensure latent space walks are maintained in iso-density regions, since moving to a higher/lower density region tends to cause unexpected transformations. To further refine semantics-specific transformations, we perform subspace learning over intermediate feature channels, such that the transformations are limited to the most relevant subspaces. Extensive experiments on a variety of benchmark datasets demonstrate that DP-LaSE is able to discover interpretable latent space walks, and specific properties of synthesized images can thus be precisely controlled.

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475293

CCS CONCEPTS

• **Computing methodologies → Learning paradigms**.

KEYWORDS

generative adversarial networks; latent space walks; density-preserving; semantic controllability

ACM Reference Format:
Guanyue Li, Yi Liu, Xiwen Wei, Yang Zhang, Si Wu, Yong Xu, and Hau-San Wong. 2021. Discovering Density-Preserving Latent Space Walks in GANs for Semantic Image Transformations. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475293

1 INTRODUCTION

Generative Adversarial Networks (GANs) [14] have become a dominant generative modeling paradigm, and GAN-based generative models have made significant progress in modeling complex data distributions over the past years. The state-of-the-art GAN-based models have shown the capability of synthesizing high-fidelity and high-resolution images, like BigGAN [5], PGGAN [19] and StyleGAN [20, 21]. In these generative models, a generator aims to learn a mapping from a latent space to a higher dimensional data space. Image semantics is typically controlled by the latent codes that encode pre-defined class labels or attributes. Semantic controllability has gained lots of attention, since there is great potential in downstream applications, such as image editing.

To understand the generation process, researchers have begun to pay attention to the interpretability of the latent space in GANs [4, 18, 32, 39]. The previous works provide evidence that there are a wide range of semantically meaningful directions in the latent space of well-trained GANs [12, 18, 32, 36, 43]. The identified latent directions can be utilized for semantic manipulation in synthesized images. This in turn reveals how GANs encode semantics. Due to the diversity and entanglement of image semantics, discovering meaningful directions in a latent space is challenging. Some recent works study this problem in a supervised manner. In this case, attribute labels of training data or a pre-trained attribute predictor are required, and the identified latent directions are limited to the pre-defined range. On the other hand, unsupervised learning techniques are also applied to explore the underlying organization of GAN feature space. In [15, 37], important latent directions are determined by performing principal component analysis and matrix factorization on the features and weights of an intermediate layer, respectively. However, there is an issue arising from the above exploration process: Only latent directions are determined, such that the precision of semantic controllability is sensitive to the strides of latent space walks, e.g., identity of a person changed in face images. Due to the fact that latent codes are drawn from a prior distribution, like Gaussian, the ones located in low-density regions may lead to unrealistic images, while complex semantics become interwoven in high-density regions. We consider that one possible reason of the above issue is the significant density differences between the two end points of a latent space walk. To address this issue, we regularize the process of latent semantic discovery via a density-preserving constraint to perform more precise semantic manipulation as shown in Figure 1.

In this work, we aim to discover steady latent space walks to precisely control semantics in synthesized images via pre-trained GANs. Toward this end, we propose an unsupervised Density-Preserving Latent Semantics Exploration model (DP-LaSE). The learning process is enforced to search for the latent space walks associated with the independent factors of variation in the synthesized images. More specifically, the important latent directions are determined by maximizing the variations in intermediate features. To capture diverse factors of variations, we also minimize the correlation between the directions and between the feature changes caused by different directions. On the other hand, we jointly train a density estimation module, such that we can perform density-preserving regularization on latent space walks. Benefiting from this regularization, the resulting latent space walks are able to be located in iso-density regions. Considering that the intermediate feature channels also have influence on semantics, we further perform subspace learning over the channels to identify the main subspace, which is most relevant to each identified latent direction. As a result, the transformation through a latent space walk is limited to the corresponding subspace. The structure of DP-LaSE is illustrated in Figure 2. Extensive experiments are performed to verify the effectiveness of the adopted techniques in improving semantic controllability. Moreover, DP-LaSE is able to outperform the previous state-of-the-art methods in terms of synthesis quality and interesting image manipulation.

The main contributions of this work are summarized as follows: (1) We explore steady latent space walks for precise semantic controllability in synthesized images via pre-trained GANs in an unsupervised manner. (2) We find that moving between different density regions with respect to a prior distribution tends to cause unexpected changes in image contents. To address this issue, we impose a density-preserving regularization on latent space walks. (3) By utilizing the association between intermediate feature channels and underlying semantics, we perform subspace learning over the channels for each important latent direction, such that the semantic transformation can be limited to the most relevant subspace.

2 RELATED WORK

2.1 Generative Adversarial Networks

GAN-based generative models have demonstrated the superior capability of modeling real image distributions by synthesizing diverse and high-fidelity images from scratch [2, 5, 14, 19–21, 27, 33]. To improve the stability of the training process of GANs, a variety of regularization methods were incorporated, such as Wasserstein distance [2, 13, 41], Lipschitz continuity constraint [26, 44] and other useful techniques [3, 9, 13, 31, 35]. Furthermore, a number of works focus on improving visual quality of synthesized images. Brock et al. [5] explored how to increase the capacity of the constituent networks in GANs and load throughput via a larger batch size, and the resulting model is referred to as BigGAN. BigGAN has demonstrated the capability of generating a wide range of realistic images over the ImageNet benchmark [7]. To synthesize high-resolution images, Karras et al. [19] adopted a progressive enhancement strategy to increase image resolution, such that the finer-scale details can be gradually incorporated. To enhance semantic controllability

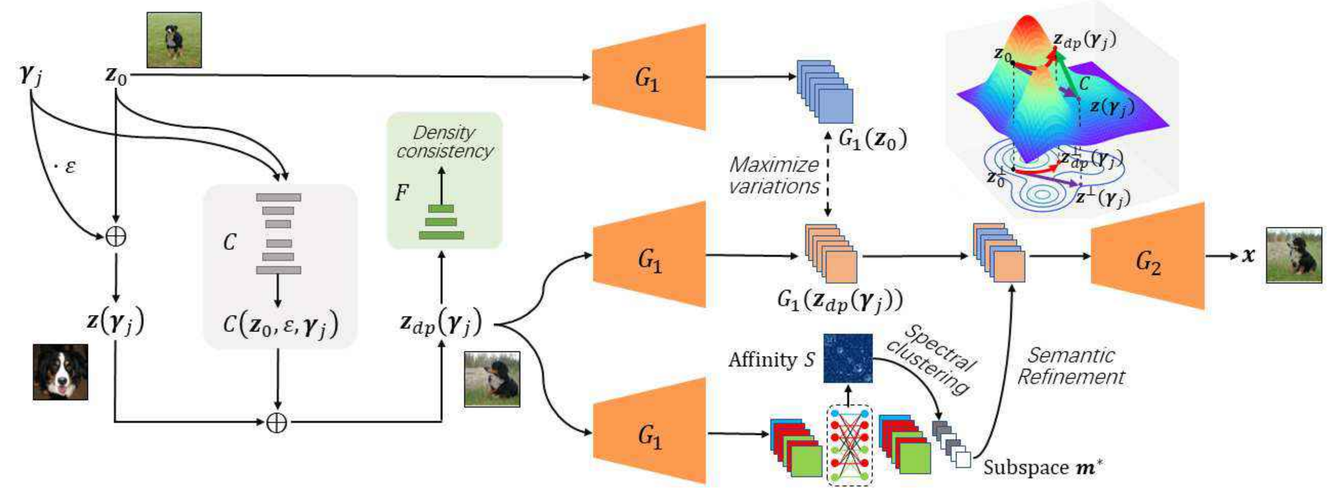


Figure 2: An overview of the proposed DP-LaSE model, which mainly consists of a pre-trained generator $\{G_1, G_2\}$, a density estimation module F and a correction module C. To discover semantically meaningful latent space walks, we identify a set of important latent directions $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ by maximizing the variations in the intermediate features $G_1(\cdot)$. Benefiting from F, we impose a density-preserving regularization on latent space walks, and C learns a correction vector to maintain the resulting latent codes z_{dp} in iso-density regions. For each γ_j, we perform subspace learning over the intermediate feature channels, and the transformations can be further controlled by limiting the effect of latent space walks to the most relevant subspace.

in the synthesis process, Karras et al. [20, 21] proposed a StyleGAN model, in which latent codes and attribute information were mapped into different intermediate feature spaces of a style-based generator with adaptive instance normalization [17]. There are also a number of works that explore how to reduce the dependence of GANs on labeled training data [10, 11, 24, 25, 30, 38, 42].

2.2 Interpretation of Prior Latent Space

It is an important research direction to investigate how GAN-based generative models learn the factors of variation from real data. In recent works [4, 12, 18, 36, 43], researchers found that the latent space or intermediate feature space of GANs typically encode a variety of semantics. To discover latent semantics, a supervised strategy was typically used to learn meaningful directions in the latent space, which corresponds to changes in labeled attributes. Goetschalckx et al. [12] used a memorability predictor [22] to guide the process of learning latent directions along which produced images have increasing or decreasing memorability. Shen et al. [36] employed a set of attribute classifiers to partition the latent space in GANs, and the normal vectors with respect to the obtained separating hyperplanes were associated with the corresponding attributes.

To avoid labeling training data or pre-training an attribute predictor, a variety of techniques were developed to learn the underlying factors of variations in an unsupervised fashion. Chen et al. [6] proposed a regularization method to explicitly learn a factorized representation by maximizing the mutual information between latent codes and synthesized images. Mollenhoff and Cremers [28] proposed the FlatGAN model, in which training data was modeled as an oriented k-dimensional manifold, and moving along tangent planes leads to interpretable manipulations. Jahanian et al [18] adopted a self-supervised training strategy to construct training image pairs via simple transformations, which have been commonly used for data augmentation. A set of latent directions were thus learnt to associate with the transformation in a supervised manner. Similarly, Plumerault et al. [32] also utilized pre-defined transformations to construct training pairs and searched for latent directions encoding the transformations. Voynov and Babenko [40] incorporated a reconstructor to reproduce the latent shift, conditioned on the transformed images, and trained a classifier to identify semantic latent directions. Harkonen et al. [15] applied principal component analysis over the intermediate features of random samples to determine the principal latent directions. Shen and Zhou [37] performed factorization over the weights of a pre-trained generator, and a closed-form algorithm was developed to determine latent semantic directions that lead to maximum image variations.

3 METHOD

3.1 Overview

GAN-based generative models learn a mapping from a latent space to a data space via a generation network $G : \boldsymbol{z} \longrightarrow \boldsymbol{x}$, where $\boldsymbol{z}$ and $\boldsymbol{x}$ denote a latent code and an image, respectively. $\boldsymbol{z}$ is typically sampled from a prior probability distribution $\boldsymbol{p}_0$. To explore the structure of the latent space, we investigate the factors of variation in synthesized images in terms of semantically meaningful latent space walks, which correspond to changes in well-defined attributes and are easy to distinguish from each other. Toward this end, we search for a set of latent directions $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$, where k denotes the number of directions. Moving $\boldsymbol{z}_0$ in the directions can lead to variations in G's intermediate features to as large an extent as possible. Due to the fact that latent codes are drawn from $\boldsymbol{p}_0$ in the training process of GANs, the ones located in low-density regions may lead to unrealistic images, while complex semantics become interwoven in high-density regions. The resulting images $G(z_0 + \varepsilon\boldsymbol{\gamma}_j)$ may be displaced from the underlying manifold, where ε denotes the manipulation intensity. To address this issue, we propose a density-preserving regularization approach to learn a correction vector $C(z_0, \varepsilon, \boldsymbol{\gamma}_j)$, such that $z_0 + \varepsilon\boldsymbol{\gamma}_j + C(z_0, \varepsilon, \boldsymbol{\gamma}_j)$ can be located in a position where the probability density is the same as that in the original position. We consider that this density-preserving regularization benefits image semantic controllability. Further, we perform subspace learning over G's intermediate feature channels to find the channels which are most relevant to the identified latent directions. As a result, the transformation can be precisely controlled by limiting the effect of a latent space walk to the corresponding subspace.

3.2 Exploration of Latent Space Walks

A latent space walk is represented by a direction vector $\boldsymbol{\gamma}_j$ multiplied with a continuous parameter ε, and the resulting latent vector is defined as follows:

$$z(\boldsymbol{\gamma}_j) = z_0 + \varepsilon\boldsymbol{\gamma}_j. \tag{1}$$

Increasing the value of ε leads to a greater degree of transformation. In the learning process, the value of ε is randomly sampled in the range of $[-\alpha, \alpha]$. We provide details on the determination of $\boldsymbol{\gamma}_j$ below.

Density-Preserving Regularization. To preserve the probability density after a latent space walk, we consider the following two cases: (1) When the prior distribution of latent codes is known, like Gaussian, we can simply modify the norm of $z(\boldsymbol{\gamma}_j)$ as follows:

$$z_{dp}(\boldsymbol{\gamma}_j) = \frac{\|z_0\|}{\|z(\boldsymbol{\gamma}_j)\|} z(\boldsymbol{\gamma}_j). \tag{2}$$

As a result, the resulting latent vector is located in the same density region as the original latent vector, since they are equidistant from the origin. (2) The prior distribution of latent codes is unknown. For instance, initial latent codes are mapped into a higher dimensional space in [5, 19–21, 33]. There can be richer semantic knowledge encoded in the new latent space, but the distribution of latent codes can be complex. To impose density-preserving regularization on latent space walks, we adopt a density estimation module F to capture the distribution. Instead of directly modeling the probability density function, we propose to learn the corresponding cumulative density function based on a well-defined set of properties. For simplicity, we assume that the elements of each latent code $z = \{z^{(1)}, z^{(2)}, \ldots, z^{(d)}\}$ are independent and identically distributed, where d denotes the number of dimensions. By definition, the probability density is the derivative of the cumulative distribution function, and we thus have $F(z^{(i)}) = Pr(\tau < z^{(i)})$ and $p_0(z^{(i)}) = F'(z^{(i)})$, which can be approximated as follows:

$$F'(z^{(i)}) \approx \frac{F(z^{(i)} + \varsigma) - F(z^{(i)})}{\varsigma}, \tag{3}$$

where we use a very small positive value of ς. To ensure that F is monotonically increasing, we define a training loss function ℓ_{mon} as follows:

$$\ell_{mon} = \sum_{i=1}^{d} \max(0, \nu - F'(z^{(i)})), \tag{4}$$

where ν denotes the minimum slope that must be maintained. To approximate the underlying distribution of latent codes, we define a likelihood-based loss function ℓ_{lik} as follows:

$$\ell_{lik} = \sum_{i=1}^{d} -\log F'(z^{(i)}). \tag{5}$$

Benefiting from the density estimation, we formulate density-preserving latent space walks as follows:

$$z_{dp}(\boldsymbol{\gamma}_j) = z_0 + \varepsilon\boldsymbol{\gamma}_j + C(z_0, \varepsilon, \boldsymbol{\gamma}_j), \tag{6}$$

where C denotes a correction module that aims to learn a correction vector. To enforce the density consistency at the start and end points, a loss ℓ_{den} is defined as follows:

$$\ell_{den} = \sum_{j=1}^{k} \left\| \sum_{i=1}^{d} \log F'(z_0^{(i)}) - \sum_{i=1}^{d} \log F'(z_{dp}^{(i)}(\boldsymbol{\gamma}_j)) \right\|^2. \tag{7}$$

As will be illustrated in the experiments, C plays an important role when performing a large ε-step transformation.

Latent Direction Exploration. Next, we determine the latent directions that give rise to variations in synthesized images in an unsupervised manner. To measure the difference between images in a feature space is typically more effective than that in the image space. We decompose the generator G into two components at an intermediate layer, and the resulting sub-networks are denoted by G_1 and G_2, respectively. G_1 takes a latent vector z as input and produce a set of feature maps $G_1(z)$, which are fed to G_2 to synthesize an image. Let $\Gamma = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_k]$ denote a matrix with its columns corresponding to a set of latent directions. We require that moving a latent code in the directions lead to large variations in the intermediate feature space, and define a corresponding loss function ℓ_{var} as follows:

$$\begin{aligned} \ell_{var} = & \sum_{j=1}^{k} -\|\Delta(z_0, \boldsymbol{\gamma}_j)\|_F^2 \\ & + \lambda \sum_{j,h=1}^{k} 1_{j \neq h} \cdot \|\Delta(z_0, \boldsymbol{\gamma}_j)^T \Delta(z_0, \boldsymbol{\gamma}_h)\|_F^2, \end{aligned} \tag{8}$$

where the function $1_{j \neq h}$ outputs 1 if $j \neq h$ and 0 otherwise,

$$\Delta(z_0, \boldsymbol{\gamma}_j) = G_1(z_0) - G_1(z_{dp}(\boldsymbol{\gamma}_j)), \tag{9}$$

λ is a weighting factor, and $\|\cdot\|_F^2$ denotes the squared Frobenius norm. To encourage Γ to associate with different factors of variations, the second term in Eq.(8) serves as a penalty for the similarity of feature changes caused by different latent directions. On the other hand, we also require the latent directions to be orthogonal with each other, and define a regularization loss ℓ_{reg} as follows:

$$\ell_{reg} = \|\Gamma^T\Gamma - I\|_F^2, \tag{10}$$

where I denotes the identity matrix.

After integrating the above two aspects: density-preserving regularization and latent direction determination, the corresponding optimization problem can be formulated as follows:

$$\min_{F,C,\Gamma} \mathsf{E}_{z_0 \sim \boldsymbol{p}_0}[\ell_{mon} + \ell_{lik} + \ell_{den} + \ell_{var}] + \zeta\ell_{reg}, \tag{11}$$

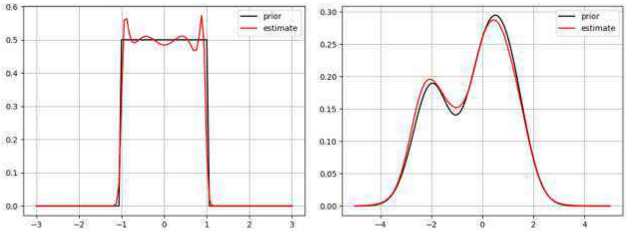


Figure 3: Estimation of unknown prior distributions: (*left*) uniform and (*right*) mixture of Gaussian.

where ζ is a weighting factor. The density estimation module F, correction module C and latent directions Γ are jointly optimized in the proposed framework.

3.3 Subspace-based Refinement

There are only a few works [4, 36] that focus on investigating the intermediate feature space of the generator in GANs. It is found that the feature channels of an intermediate layer are highly correlated and interact with each other to exert significant influence on semantics in synthesized images. We demonstrate that the main semantic factors can be disentangled from other factors by finding the most relevant subspaces over intermediate feature channels of GANs.

For each latent direction $\boldsymbol{\gamma}_j$, subspace learning is performed over the output channels of G_1. Specifically, we feed a latent vector $z_{dp}(\boldsymbol{\gamma}_j)$ to G_1, and represent $G_1(z_{dp}(\boldsymbol{\gamma}_j))$ in the form of a matrix $M(z_{dp}(\boldsymbol{\gamma}_j))$ with its columns corresponding to the flattened feature maps. To model the self-expressiveness property of the feature channels, a reconstruction coefficient matrix R is learnt to minimize ℓ_{rec} defined as follows:

$$\ell_{rec} = \|M(z_{dp}(\boldsymbol{\gamma}_j)) - M(z_{dp}(\boldsymbol{\gamma}_j))R\|_F^2. \tag{12}$$

To avoid a trivial solution ($R = I$), it is necessary to set $diag(R) = 0$. On the other hand, a sparsity regularization criterion is imposed on R to ensure subspace discovery, and the optimization formulation is summarized as follows:

$$\begin{aligned} &\min_R \ \mathsf{E}_{z_0 \sim \boldsymbol{p}_0}[\ell_{rec}] + \eta\|R\|_1, \\ &s.t. \quad diag(R) = 0, \end{aligned} \tag{13}$$

where η is a weighting factor. Based on R, we compute an affinity matrix S among the channels as follows:

$$S = \frac{|R| + |R^T|}{2}, \tag{14}$$

and then apply a spectral clustering method [29] to partition the feature channels into a specified number of subsets. We find that the largest subset is most relevant to the semantics associated with the latent direction $\boldsymbol{\gamma}_j$, and limit the transformation to the corresponding subspace as follows:

$$\boldsymbol{x} = G_2(\boldsymbol{m}^* \odot G_1(z_{dp}(\boldsymbol{\gamma}_j)) + (1 - \boldsymbol{m}^*) \odot G_1(z_0)), \tag{15}$$

where the feature channels in the largest subset are indicated by a binary mask $\boldsymbol{m}^*$, and $\odot$ denotes channel-wise multiplication. The advantage of subspace-based refinement is to allow the modification of intermediate features within a range of channels, while leaving the other channels unchanged.

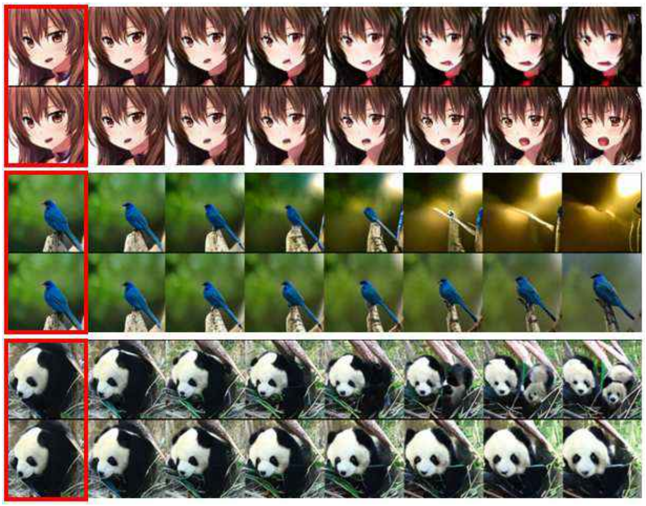


Figure 4: Synthesized images of latent space walks (*upper row*) w/o DP Reg. and (*bottom row*) w/ DP Reg.. The initial images are annotated with red bounding boxes.

4 EXPERIMENTS

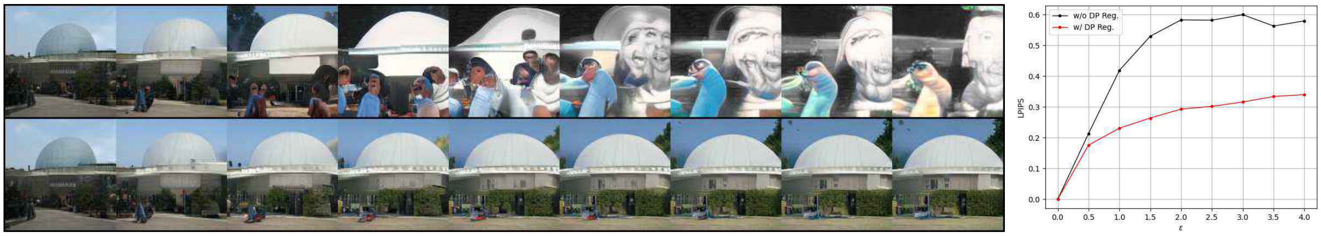
In this section, we perform extensive experiments to evaluate the proposed DP-LaSE model on a variety of standard image synthesis benchmarks, including CelebA-HQ [19], FF-HQ [20], Anime Faces [1] and ImageNet [8]. We first provide the implementation details of DP-LaSE and experiment settings. Next, we verify the effectiveness of our density-preserving regularization approach in stabilizing the transformation that is performed by latent space walks. We also conduct experiments to highlight the association between semantics and intermediate feature channels. Further, we demonstrate the advantage of DP-LaSE over the state-of-the-art competing methods in semantic controllability in synthesized images.

4.1 Implementation and Settings

All the experiments are based on well-trained GANs, including SNGAN [26], BigGAN [5], PGGAN [19] and StyleGAN [20, 21]. In addition to the generator G in a pre-trained GAN, the proposed DP-LaSE model mainly consists of a density estimation module F and a correction module C, which are composed of 3 and 6 fully connected layers, respectively. For important latent direction discovery, we divide G into G_1 and G_2 by the 3rd intermediate layer (1st-4th layers for StyleGAN), and set the total number of directions to 20. The moving step ε randomly takes values in the range of $[-10, 10]$. In Eq.(4), Eq.(8), Eq.(11) and Eq.(13), the parameters ν, λ, ζ and η are set to 0.01, 0.001, 10 and 0.5, respectively. G's parameters are frozen when jointly training F and C together to determine Γ. We adopt the Adam optimizer [23] with 50,000 training iterations, learning rate of 0.0001 and momentum parameters of $(0.9, 0.999)$. For subspace learning, we incorporate a fully connected layer on top of G_1, and the setting of the Adam optimizer is the same as the above task. We

Table 1: IS and FID scores of synthesized images by performing random latent space walks without and with density-preserving regularization on ImageNet.

	$\varepsilon = -10$		$\varepsilon = -5$		$\varepsilon = 0$		$\varepsilon = 5$		$\varepsilon = 10$	
Method	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓
w/o DP Reg.	3.44±0.04	199.23	5.47±0.07	143.68	96.45±1.94	43.84	5.54±0.08	143.74	3.39±0.05	203.10
w/ DP Reg.	**97.10±2.09**	**44.05**	**96.22±2.91**	**44.01**	**97.31±2.07**	**43.68**	**99.63±2.10**	**43.70**	**99.78±1.86**	**43.69**



(a)

(b)

Figure 5: (a) Visual comparison of synthesized images by performing a random latent space walk (*upper row*) w/o DP Reg. and (*bottom row*) w/ DP Reg., and (b) the perceptual distance between the initial and transformed images. The stride ε increases from 0 to 4 with an interval of 0.5.

set the total number of subspaces to 4 when performing spectral clustering on the intermediate channels.

4.2 Probability Density Estimation

For the case that the prior distribution of latent codes is unknown, we adopt the module F to capture the underlying distribution. In this example, we highlight the effectiveness of F on two toy datasets: latent codes are sampled from uniform and mixture of Gaussian distributions. In Figure 3, we can observe that the estimated probability density curves can effectively approximate the ground-truth ones.

4.3 Density-Preserving Latent Walks

Latent codes are typically sampled from a prior distribution in the training process of GANs. The latent codes in the low-density regions have much less chance of being used for image synthesis, while the ones in the high-density regions encode complex semantics. When performing latent space walks, we consider that maintaining the start and end points in the iso-density regions benefits semantic controllability. To verify this point, We perform an experiment based on SNGAN and BigGAN, which are trained on Anime and ImageNet. Figure 4 shows the synthesized images for the cases without and with density-preserving regularization.

In addition, we randomly sample latent codes and move each one in a random direction, and quantitatively evaluate the quality of synthesized images on ImageNet in terms of Inception Score (IS) [34] and Fréchet Inception Distance (FID) [16]. As shown in Table 1, when increasing the stride ε of latent space walks, we find that the IS/FID score of synthesized images without regularization falls/rises rapidly. This suggests that the quality of synthesized images becomes significantly poor, when the degree of transformation becomes greater. In contrast, the quality of synthesized images with regularization is stable. In Figure 5, we also visualize examples of latent space walks without and with density-preserving regularization. Continuing to increase the amplitude of the transformations typically leads to unrealistic images or undesirable deviation from the initial ones. We consider that the resulting images become displaced from the underlying manifold, and the Learned Perceptual Image Patch Similarity metric (LPIPS) [45] between the initial and transformed images thus increases significantly.

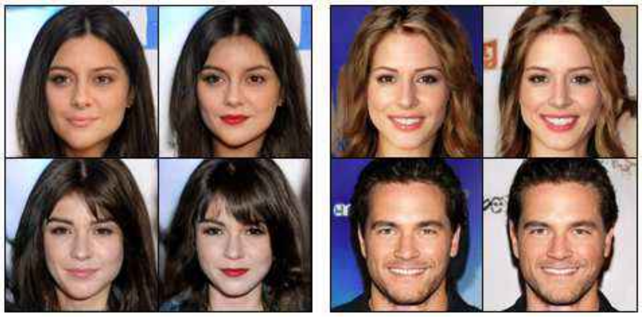


Figure 6: Examples of exchanging feature maps in the subspaces associated with the attributes of (*left*) lip color and (*right*) background. In each example, the reference images are in the diagonal positions.

4.4 Subspace-based Image Manipulation

We exploit the self-expressiveness property of intermediate feature channels in PGGAN trained on CelebA-HQ [19] to perform subspace learning over them. We first consider the case without

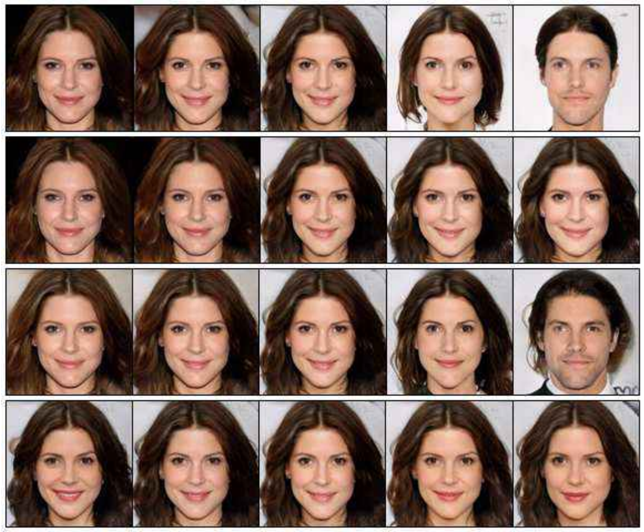


Figure 7: Given a random latent space walk (*first row*), subspace learning is performed over intermediate feature channels to disentangle the attributes of (*second row*) background, (*third row*) gender and (*fourth row*) expression.

performing any latent space walks. To demonstrate the characteristics of a specific subspace, we exchange the corresponding intermediate feature maps of two reference images, and feed the resulting features in G_2 to synthesize two new images. As shown in Figure 6, we find that the obtained subspaces have influence on meaningful attributes: lip color and background. Next, we perform latent space walks with a random direction, and apply the subspace learning approach to further disentangle the associated semantics. In Figure 7, we can observe that the transformation involves multiple semantics. We select three largest subspaces, allow the changes occur in the one of them, and fix the remaining subspaces. The results demonstrate that the subspaces disentangle the attributes of background, gender and expression.

4.5 Interpretable Latent Walks

To demonstrate the capability of DP-LaSE in capturing the interpretable factors in synthesized images, we perform a set of experiments based on BigGAN and StyleGAN, which are trained on ImageNet and FF-HQ, respectively. Although different network architectures are used in BigGAN and StyleGAN, DP-LaSE is able to identify diverse and meaningful latent directions, and the corresponding latent space walks lead to interpretable transformations in the synthesized images. In Figure 8, each row demonstrates how a latent space walk affects the reference (first) image. The variations with the latent space walks can be easily interpreted, and the corresponding transformations are in the same manner for each case. For instance, we can successfully zoom in on the dog by constructing a latent space walk. In addition, the thickness of beard and the length of hair can be well controlled without changing identity and expression. The results suggest that DP-LaSE is able to semantically change the reference images, while remaining consistent with the image context.

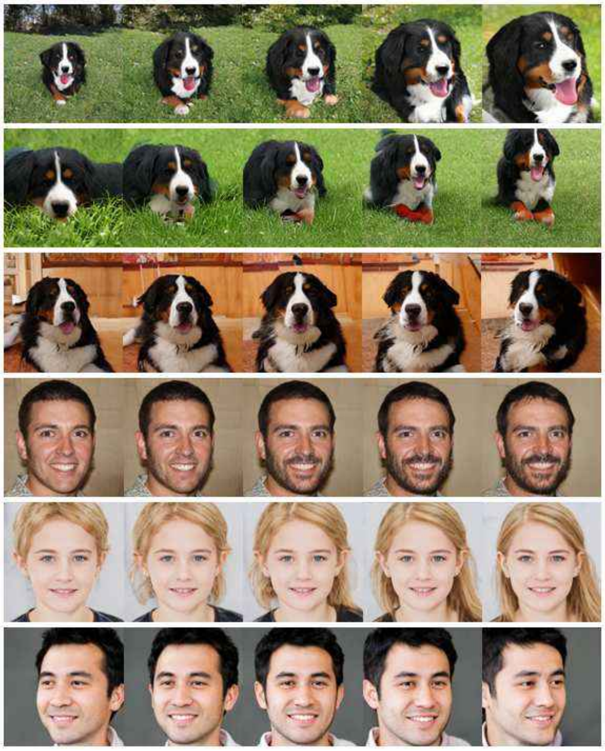


Figure 8: A number of synthesized images with the semantically meaningful transformations we explore.

4.6 Comparison to State-of-the-arts

We further perform an experiment to conduct a comparison between the proposed DP-LaSE and state-of-the-art methods, UDID [40] and SeFa [37], in the extent of similar transformations. The experiment is based on StyleGAN and PGGAN, which are trained on FF-HQ and CelebA-HQ, respectively. We adopt the same experimental configuration for the methods, and the transformation results are shown in Figure 9. There are four attributes involved in the transformations: glasses, hat, hair color and age. Separating the semantics is not trivial. UDID fails to disentangle the attributes of sunglasses-hat and gender-age, and SeFa changes hairstyle/identity when performs the transformation on glasses/hair color. In contrast, the transformations determined by DP-LaSE are desirable, and control large-scale variations. Based on the above results, we consider that the competing methods fail to reach our transformation level. In Figure 10, we investigate the capacity of semantic transformations by increasing the stride ε. A quantitative comparison between DP-LaSE and two competing methods is performed in terms of LPIPS. The result suggests that DP-LaSE performs better than UDID and SeFa in maintaining the perceptual distances between initial and transformed images.

5 CONCLUSION

In this work, we present an unsupervised semantically meaningful latent space walk exploration model, which is useful for understanding the underlying structure of the latent space in well-trained

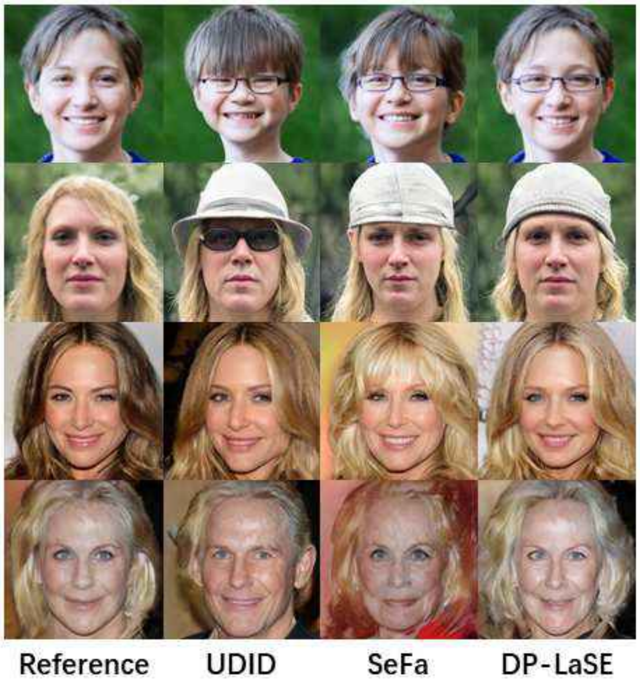


Figure 9: Visual comparison of DP-LaSE and state-of-the-art methods in semantic controllability.

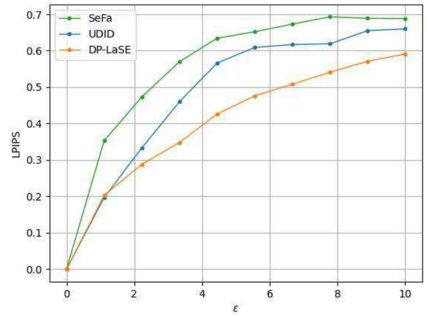


Figure 10: Comparison between DP-LaSE and state-of-the-art methods in terms of the LPIPS distance between initial and transformed images.

GANs. In our model, a set of orthogonal latent directions are determined by maximizing the changes in the generator’s intermediate features. To encourage the discovery of the latent space walks that lead to meaningful transformations, we incorporate a density-preserving regularization criterion in the learning process. On the other hand, we further limit the transformation to the most relevant intermediate feature subspace. On the standard benchmarks, we find that exploitation of the discovered latent space walks would make image manipulation more straightforward. Our exploration facilitates the understanding of GANs and enriches the techniques of semantic controllability.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Project No. 62072188, 62072189), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11201220), and in part by the Natural Science Foundation of Guangdong Province (Project No. 2019A050510010, 2020A1515010484).

REFERENCES

[1] Anonymous, Danbooru Community, and Gwern Branwen. 2020. A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2019

[2] Martin Arjovsky, Soumith Chintala, and Leon Bottou. 2017. Wasserstein generative adversarial networks. In *Proc. International Conference on Machine Learning.*

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: fine-grained image generation through asymmetric training. In *Proc. International Conference on Computer Vision.*

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, Willian T. Freeman, and Antonio Torralba. 2019. GAN Dissection: visualizing and understanding generative adversarial networks. In *Proc. International Conference on Learning Representation.*

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. In *Proc. International Conference on Learning Representation.*

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In *Proc. Neural Information Processing Systems.*

[7] Jia Deng, Alexander C. Berg, Sanjeev Satheesh, Hao Su, Aditya Khosla, and Li Fei-Fei. 2012. ImageNet Large Scale Visual Recognition Competition 2012. http://www.image-net.org/challenges/LSVRC/2012/

[8] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[9] Zhijie Deng, Hao Zhang, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, and Eric P. Xing. 2017. Strctured generative adversarial networks. In *Proc. Neural Information Processing Systems.*

[10] Jinhao Dong and Tong Lin. 2019. MarginGAN: adversarial training in semi-supervised learning. In *Proc. Neural Information Processing Systems.*

[11] Zhe Gan, Liqun Chen, Weiyao Wang, Yunchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. 2017. Triangle generative adversarial networks. In *Proc. Neural Information Processing Systems.*

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. GANalyze: toward visual definitions of cognitive image properties. In *Proc. International Conf. on Computer Vision.*

[13] Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. 2019. Twin auxiliary classifiers GAN. In *Proc. Neural Information Processing Systems.*

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. Neural Information Processing Systems.*

[15] Erik Harkonen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: discovering interpretable GAN controls. In *Proc. Neural Information Processing Systems.*

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, and Bernhard Nessler. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Neural Information Processing Systems.*

[17] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. International Conference on Computer Vision.*

[18] Ali Jahanian, Lucy Chai, and Phillip Isola. 2020. On the steerability of generative adversarial networks. In *Proc. International Conference on Learning Representation.*

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. International Conference on Learning Representation.*

[20] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[22] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proc. International Conference on Computer Vision.*

[23] Diederik Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations.*

[24] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang. 2017. Triple generative adversarial nets. In *Proc. Neural Information Processing Systems.*

[25] Yi Liu, Guangchang Deng, Xiangping Zeng, Si Wu, Zhiwen Yu, and Hau-San Wong. 2020. Regularizing discriminative capability of CGANs for semi-supervised generative learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. In *Proc. International Conference on Learning Representation.*

[27] Takeru Miyato and Masanori Koyama. 2018. cGANs with projection discriminator. In *Proc. International Conference on Learning Representation.*

[28] Thomas Mollenhoff and Daniel Cremers. 2019. Flat metric minimization with applications in generative modeling. In *arXiv:1905.04730.*

[29] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. On spectral clustering: analysis and an algorithm. In *Proc. Neural Information Processing Systems.*

[30] Augustus Odena. 2016. Semi-supervised learning with generative adversarial networks. In *Proc. International Conference on Learning Representation.*

[31] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier GANs. In *Proc. International Conference on Machine Learning.*

[32] Antoine Plumerault, Herve Le Borgne, and Celine Hudelot. 2020. Controlling generative models with continuous factors of variations. In *Proc. International Conf. on Learning Representation.*

[33] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. International Conference on Learning Representation.*

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Proc. Neural Information Processing Systems.*

[35] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. 2020. A U-Net based discriminator for generative adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[36] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of GANs for semantic face editing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[37] Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in GANs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[38] Jost Tobias Springenberg. 2016. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *Proc. International Conference on Learning Representation.*

[39] Andrey Voynov and Artem Babenko. 2019. RPGAN: GANs interpretability via random routing. In *arXiv:1912.10920.*

[40] Andrey Voynov and Artem Babenko. 2020. Unsupervised discovery of interpretable directions in the GAN latent space. In *Proc. International Conference on Machine Learning.*

[41] Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. 2018. Improving the improved training of Wasserstein GANs: a consistency term and its dual effect. In *Proc. International Conference on Learning Representation.*

[42] Si Wu, Guangchang Deng, Jichang Li, Rui Li, Zhiwen Yu, and Hau-San Wong. 2019. Enhancing TripleGAN for semi-supervised conditional instance synthesis and classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*

[43] Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2020. Semantic hierarchy emerges in deep generative representations for scene synthesis. In *arXiv:1911.09267v3.*

[44] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-attention generative adversarial networks. In *arXiv:1705.05512.*

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*