Attribute Group Editing for Reliable Few-shot Image Generation

Guanqi Ding[1,2*], Xinzhe Han[1,2*], Shuhui Wang[2,4†], Shuzhe Wu[3], Xin Jin[3], Dandan Tu[3], Qingming Huang[1,2,4]

[1]University of Chinese Academy of Sciences, Beijing, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[3]Huawei Cloud EI Innovation Lab, China   [4]Peng Cheng Laboratory, Shenzhen, China

{dingguanqi19, hanxinzhe17}@mails.ucas.ac.cn, wangshuhui@ict.ac.cn,

{wushuzhe2, jinxin11, tudandan}@huawei.com, qmhuang@ucas.ac.cn

Abstract

Few-shot image generation is a challenging task even using the state-of-the-art Generative Adversarial Networks (GANs). Due to the unstable GAN training process and the limited training data, the generated images are often of low quality and low diversity. In this work, we propose a new "editing-based" method, i.e., Attribute Group Editing (AGE), for few-shot image generation. The basic assumption is that any image is a collection of attributes and the editing direction for a specific attribute is shared across all categories. AGE examines the internal representation learned in GANs and identifies semantically meaningful directions. Specifically, the class embedding, i.e., the mean vector of the latent codes from a specific category, is used to represent the category-relevant attributes, and the category-irrelevant attributes are learned globally by Sparse Dictionary Learning on the difference between the sample embedding and the class embedding. Given a GAN well trained on seen categories, diverse images of unseen categories can be synthesized through editing category-irrelevant attributes while keeping category-relevant attributes unchanged. Without re-training the GAN, AGE is capable of not only producing more realistic and diverse images for downstream visual applications with limited data but achieving controllable image editing with interpretable category-irrelevant directions. Code is available at https://github.com/UniBester/AGE.

1. Introduction

The emergence of GANs [10] has enabled the deep generative model to generate images with higher quality and diversity. However, due to the characteristics of GANs, training a good GAN model often requires a lot of training data
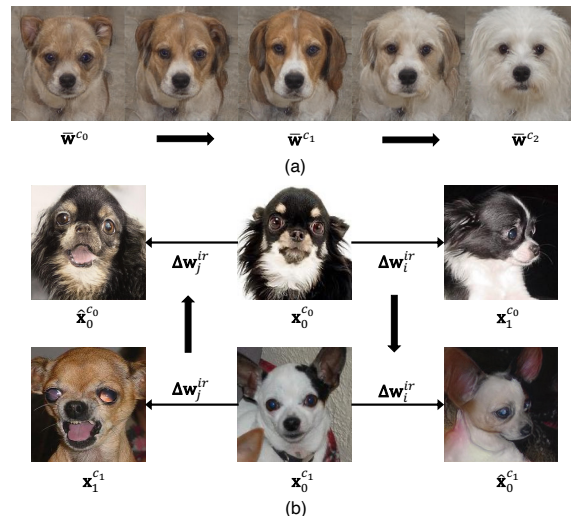
---

*These authors contributed equally to this work.
†Corresponding authors.

Figure 1. **Illustration of attribute editing in the latent space.** (a) The category-relevant attribute editing between the class embedding $\overline{\mathbf{w}}^{c_m}$. (b) The category-irrelevant attribute editing within categories. The same editing in the latent space is universal between different categories

and is difficult to converge [41]. Given a few images from an unseen category, it is hard for GANs to generate new realistic and diverse images for this category. This task is referred to as few-shot image generation, which can benefit a bunch of downstream tasks like low-data detection [8] and few-shot classification [34, 36].

Existing few-shot image generation methods can be roughly divided into three types, *i.e.*, optimization-based [7, 23], fusion-based [4, 11, 14, 15], and transformation-based [3, 13]. Optimization-based methods introduce meta-learning paradigms to generate new images by learning an initialized base model and fine-tuning the model with each unseen category, but images generated by methods of this type are blurry and of low quality. Fusion-based methods fuse several input images in a feature space and decode the fused feature back to a realistic image of the same category. However, these methods need at least two images as

input and can only generate images similar to input images. Transformation-based methods find intra-category transformations and apply these transformations to unseen category samples to generate more images of the same category. However, the end-to-end training of image transformation and generation is very unstable. Worse still, these transformations tends to be complicated, and the generated images are often of low-quality and even crashed.

Drawing on the idea of StyleGAN [20, 21], an image can be regarded as a collection of different attributes. The category of an image is decided by objects' category-relevant attributes, such as the shape of the face and the morphology of the fur. Differences between images of the same category are reflected in category-irrelevant attributes including expressions, postures, *etc*. Moreover, many works [12, 32] have shown that GANs represent these attributes in the latent space. Moving the latent code along an identified direction can accordingly change the semantic in the output images. Theoretically, given a pretrained GAN, an object of an unseen category can be generated with combinations of attributes from seen categories as shown in Figure 1(a). Diverse images of the same category can be generated by editing category-irrelevant attributes, which are shared across all categories as shown in Figure 1(b). If these semantically meaningful directions can be distinguished, we can achieve reliable few-shot image generation needless of re-training a GAN.

In order to identify such directions, image editing methods typically annotate a collection of synthesized samples and train linear classifiers in the latent space. They require a clear definition of the target attributes as well as the corresponding manual annotations. However, it will be unrealistic to obtain such detailed annotations for more complicated multi-category image generation [6, 24, 26]. Therefore, the key challenge of the editing-based method is to factorize the meaningful directions for category-relevant and category-irrelevant attributes without explicit supervision.

To achieve this goal, we propose *Attribute Group Editing (AGE)*, which examines the variation relationship between the image and the internal representation. The core of AGE is to factorize the directions of category-irrelevant attributes and category-relevant attributes without explicit supervision. First, the class embedding for a specific category is obtained from the mean representation of all samples from this category. As shown in Figure 1(a), for a seen category with a large amount of training data, this embedding is likely to disentangle all category-relevant attributes from major category-irrelevant attributes. Afterward, any samples in the dataset can be regarded as category-irrelevant editing from the corresponding class embedding. In order to factorize category-irrelevant directions in latent space, we model this editing process with Sparse Dictionary Learning (SDL) [2, 25]. A number of constraints are used to en-

sure that every direction in the dictionary is semantically meaningful and category-irrelevant. Different linear combinations of directions in the dictionary can facilitate generation of diverse images without changing their category.

Our contributions can be summarized as follows:

- We present a new perspective for few-shot image generation, *i.e.*, diverse images of unseen categories can be produced through category-irrelevant image editing.

- We propose a new method, Attribute Group Editing (AGE), which can identify groups of category-relevant and category-irrelevant editing directions from a pretrained GAN without explicit supervision.

- Extensive experiments suggest that AGE achieves more stable few-shot image generation with high quality and diversity. Besides, since the editing directions discovered by AGE are semantically meaningful, we can also perform controllable image generation based on the learned attribute dictionary.

2. Related Work

Few-shot image generation. Existing few-shot image generation methods can be roughly divided into optimization-based methods, fusion-based methods, and transformation-based methods. Optimization-based methods [7, 23, 29] combine meta-learning and adversarial learning to generate images of unseen category by fine-tuning the model. However, the images generated by such methods have poor authenticity. Fusion-based methods fuse the features by matching the random vector with the conditional images [14] or interpolate high-level features of conditional images by filling in low-level details [11, 15]. Simple content fusion limits the diversity of generated images. Transformation-based methods [3, 13] capture the cross-category or intra-category transformations to generate novel data of unseen categories. These works capture the transformations from the image differences and may corrupt due to the complex transformations between intra- and inter-category pairs. From our new "editing-based" perspective, the intra-category transformation can be alternatively modeled as category-irrelevant image editing based on one sample instead of pairs of samples.

Few-shot image-to-image translation. Few shot image-to-image translation methods map images from one domain to another based on a few images, like category transfer [16, 24, 37], weather transfer [30] and style transfer [22, 27, 39]. These methods also focus on the few-shot setting but mainly handles domain transfer rather than object categories.

Image manipulation. Recent studies have shown that GANs can represent multiple interpretable attributes in the latent space [20, 40]. For image editing, supervised learning methods [9, 32, 40] annotate predefined attributes according to pre-trained classifiers, and then learn the potential directions of the attributes in the latent space. However, they
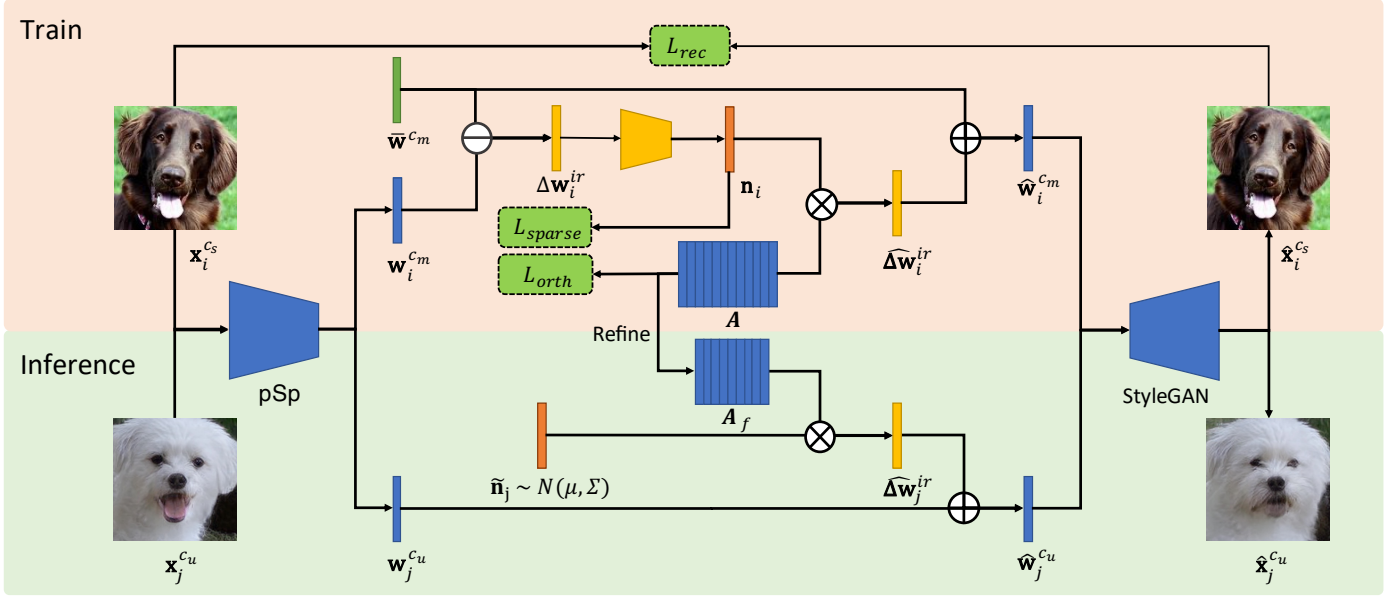
Figure 2. The overview of AGE. AGE learns a dictionary $\boldsymbol{A}$ consisting of category-irrelevant attribute editing directions in the training stage and generates diverse images via category-irrelevant manipulation according to the refined dictionary $\boldsymbol{A}_f$.

heavily rely on the attribute predictors and human annotations. Some concurrent work studies unsupervised semantic discovery in GANs. The meaningful dimensions can be identified by using segmentation-based networks [5], linear subspace models [12], Principal Components Analysis in the activation space [17], or carefully designed disentanglement constraints [28, 33, 38]. Different from traditional image editing, AGE focuses on attribute factorization on more challenging multi-category image generation, which cannot be performed by traditional image editing methods.

3. Method

A training set $\mathcal{D}_{train} = \{x_i^{c_m}\}^{N_m \times M}$ consists of M seen categories and a testing set $\mathcal{D}_{test} = \{x_i^{c_k}\}^{N_k \times K}$ consists of K unseen categories, where the number of images N_k in each category is generally small, *i.e.*, 10 or 15. Few-shot image generation aims to train a multi-category generative network with $\mathcal{D}_{train}$ but generates diverse images of the K unseen categories by the few images in $\mathcal{D}_{test}$.

In this section, we introduce Attribute Group Editing (AGE), a method to generate images of unseen categories without re-training a GAN model. AGE makes use of a large number of images of known categories to identify the semantically meaningful directions of category-relevant and category-irrelevant attributes without explicit supervision.

3.1. Preliminaries

GAN Inversion. The generator $G(\cdot)$ in GANs learns a mapping from the d-dimensional latent space $\mathcal{Z} \in \mathbb{R}^d$ to higher dimensional images $\mathcal{X} \in \mathbb{R}^{H \times W \times C}$. On the contrary, given

an image $x_i \in \mathcal{X}$, it can also be embedded to the latent space with GAN inversion [1, 19, 31, 35, 42]. The process of GAN inversion $I(\cdot)$ and generation can be formulated as follow:

$$\mathbf{z}_i = I(x_i), \quad \hat{x}_i = G(\mathbf{z}_i). \tag{1}$$

Semantic Manipulation. The latent space of GANs has recently been shown to encode rich semantic knowledge [9, 18, 40]. Different directions in the latent space control different attributes. Many works [12, 32, 33, 38] proposed to manipulate the latent vector $\mathbf{z}_i$ in a certain direction $\Delta \mathbf{z}_i \in \mathbb{R}^d$ to edit the corresponding attribute:

$$\texttt{edit}(G(\mathbf{z}_i)) = G(\mathbf{z}'_i) = G(\mathbf{z}_i + \alpha \Delta \mathbf{z}_i), \tag{2}$$

where $\texttt{edit}(\cdot)$ denotes the editing operation on images. α stands for the manipulation intensity.

Given a well-trained GAN for multi-category image generation, editing can be divided into category-relevant editing and category-irrelevant editing. For a sampled latent vector $\mathbf{z}_i^{c_m}$ of the category c_m, category-relevant editing $\texttt{edit}_r(\cdot)$ is:

$$\texttt{edit}_r(G(\mathbf{z}_i^{c_m})) = G(\mathbf{z}_i^c + \alpha \Delta \mathbf{z}^r) = \hat{x}_i^{c_k}, \tag{3}$$

where $\Delta \mathbf{z}^r$ denotes the directions of category-relevant manipulation and $\hat{x}_i^{c_k}$ is an image of a new category c_k.

On the other hand, for category-irrelevant editing $\texttt{edit}_{ir}(\cdot)$, we have:

$$\texttt{edit}_{ir}(G(\mathbf{z}_i^{c_m})) = G(\mathbf{z}_i^{c_m} + \alpha \Delta \mathbf{z}^{ir}) = \hat{x}_i^{c_m}, \tag{4}$$

where $\Delta \mathbf{z}^{ir}$ denotes the directions of category-irrelevant editing and $\hat{x}_i^{c_m}$ is an image of the same category c_m.
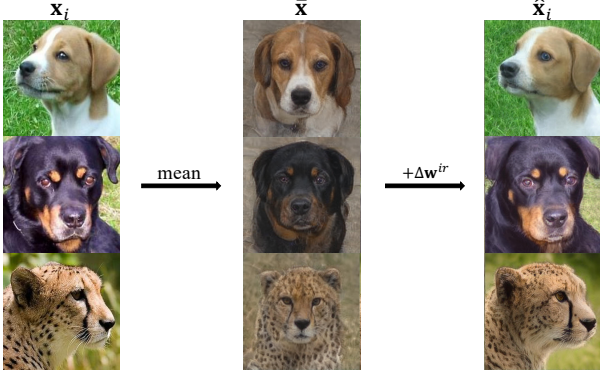
Figure 3. Any image can be regarded as category-irrelevantly edited from the corresponding class embedding. This linear transformation can reconstruct most attributes of the input images.

3.2. Attribute Group Editing

The overall framework of AGE is shown in Figure 2, which consists of two main parts: Image Embedding and Attribute Factorization.

3.2.1 Image Embedding

To achieve image editing, we should first embed the image samples into the latent space. In practice, we employ pSp [31] to encode an image x_i to $\mathcal{W}^+$ [1] space of StyleGAN2 [20,21].

$$\mathbf{w}_i = \text{pSp}(x_i), \tag{5}$$

where $\mathbf{w}_i \in \mathbb{R}^{18 \times 512}$ is the corresponding latent vector of x_i in the $\mathcal{W}^+$ space.

3.2.2 Attribute Factorization

Given latent representations in $\mathcal{W}^+$ space, we aim to distinguish a set of category-relevant attribute directions and category-irrelevant attribute directions according to the given dataset.

Category-relevant Attributes. The combination of category-relevant attributes identifies which category an image belongs to. It encodes the common attributes among all samples from one specific category. Therefore, we use the mean vector $\overline{\mathbf{w}}^{c_m} \in \mathcal{W}^+$ of all samples in a category c_m to represent the class embedding, *i.e.*, the combined category-relevant attributes of c_m:

$$\overline{\mathbf{w}}^{c_m} = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbf{w}_i^{c_m}, \tag{6}$$

where N_m is the number of samples from the category c_m.

The dictionary of category-relevant attributes of M seen categories is defined as $\boldsymbol{B} = [\overline{\mathbf{w}}^{c_1}, \overline{\mathbf{w}}^{c_2}, ..., \overline{\mathbf{w}}^{c_M}]$.

Category-irrelevant Attributes. As shown in Figure 3, given $\overline{\mathbf{w}}^{c_m}$ contains all category-relevant attributes, any image $x_i^{c_m}$ of the category c_m can be obtained by manipulating with category-irrelevant editing $\Delta \mathbf{w}^{ir}$ as Eq. 7:

$$\mathbf{w}_i^{c_m} = \overline{\mathbf{w}}^{c_m} + \Delta \mathbf{w}_i^{ir}. \tag{7}$$

The category-irrelevant directions are common across all known and unknown categories. To learn the category-irrelevant directions, we model this manipulation process as Sparse Dictionary Learning (SDL) [2,25].

Given a sample of manipulation direction $\Delta \mathbf{w}_i^{ir}$, we optimize a global dictionary $\boldsymbol{A} \in \mathbb{R}^{18 \times 512 \times l}$ that contains all directions of category-irrelevant attributes and a sparse representation $\mathbf{n}_i \in \mathbb{R}^{18 \times l}$ with

$$\min_{\mathbf{n}} \|\mathbf{n}_i\|_0 \quad \text{s.t.} \ \Delta \mathbf{w}_i^{ir} = \boldsymbol{A} \mathbf{n}_i, \tag{8}$$

where $\|.\|_0$ is the L_0 constraint. This sparse constraint encourages each element in $\boldsymbol{A}$ to be semantically meaningful.

In practice, it is optimized via an Encoder-Decoder architecture. The sparse representation $\mathbf{n}_i$ is obtained from $\Delta \mathbf{w}_i^{ir}$ with a Multi-layer Perceptron (MLP):

$$\mathbf{n}_i = \text{MLP}(\Delta \mathbf{w}_i^{ir}). \tag{9}$$

Since the L_0 loss is not derivable, we approximate L_0 constraint with L_1 with the sigmoid activation:

$$L_{\text{sparse}} = \|\sigma(\theta_0 \mathbf{n}_i - \theta_1)\|_1, \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid function. θ_0 and θ_1 are hyper-parameters to control the sparsity.

The generator is to generate an image close to the input x_i, which is optimized with the L_2 reconstruction loss:

$$L_{\text{rec}} = \|G(\overline{\mathbf{w}}^{c_m} + A\mathbf{n}_i) - x_i^{c_m}\|_2. \tag{11}$$

Moreover, to further guarantee that $\boldsymbol{A}\mathbf{n}_i$ only edits category-irrelevant attributes, the embedding of an edited images $\hat{\mathbf{w}}_i^{c_m}$ should have the same category-relevant attributes response as the class embedding $\overline{\mathbf{w}}^{c_m}$:

$$\boldsymbol{B}^T \hat{\mathbf{w}}_i^{c_m} = \boldsymbol{B}^T \overline{\mathbf{w}}^{c_m},$$
$$\boldsymbol{B}^T \overline{\mathbf{w}}^{c_m} + \boldsymbol{B}^T \boldsymbol{A} \mathbf{n}_i = \boldsymbol{B}^T \overline{\mathbf{w}}^{c_m}, \tag{12}$$
$$\boldsymbol{B}^T \boldsymbol{A} \mathbf{n}_i = \mathbf{0}.$$

To ensure the satisfaction of Eq. 12, we formulate an orthogonal constraint between $\boldsymbol{A}$ and $\boldsymbol{B}$ with:

$$L_{\text{orth}} = \|\boldsymbol{B}^T \boldsymbol{A}\|_F^2, \tag{13}$$

where $\|.\|_F^2$ denotes the Frobenius Norm.

The overall loss function is

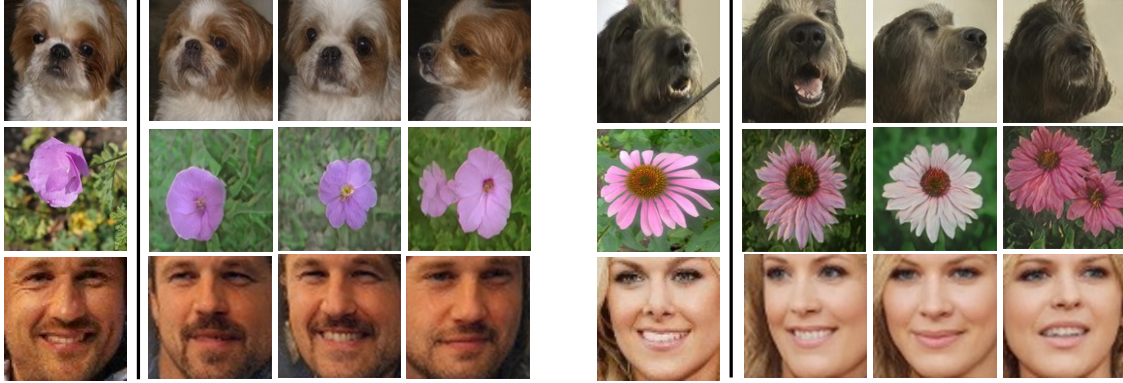$$L = L_{\text{rec}} + \lambda_1 L_{\text{orth}} + \lambda_2 L_{\text{sparse}}. \tag{14}$$

Figure 4. One-shot image generation from AGE on Animal Faces, Flowers and VGGFaces.

Table 1. FID(↓) and LPIPS(↑) of images generated by different methods for unseen categories. VGGFaces is marked with * because most methods report different number of unseen categories on this dataset (*e.g.* 552 in LoFGAN, 96 in DeltaGAN and 497 in L2GAN).

Method	Settings	Flowers		Animal Faces		VGG Faces*	
		FID(↓)	LPIPS(↑)	FID(↓)	LPIPS(↑)	FID(↓)	LPIPS(↑)
FIGR [7]	3-shot	190.12	0.0634	211.54	0.0756	139.83	0.0834
GMN [4]	3-shot	200.11	0.0743	220.45	0.0868	136.21	0.0902
DAWSON [23]	3-shot	188.96	0.0583	208.68	0.0642	137.82	0.0769
DAGAN [3]	1-shot	179.59	0.0496	185.54	0.0687	134.28	0.0608
MatchingGAN [14]	3-shot	143.35	0.1627	148.52	0.1514	118.62	0.1695
F2GAN [15]	3-shot	120.48	0.2172	117.74	0.1831	109.16	0.2125
LoFGAN [11]	3-shot	79.33	0.3862	112.81	0.4964	**20.31**	0.2869
DeltaGAN [13]	1-shot	109.78	0.3912	89.81	0.4418	80.12	0.3146
AGE	1-shot	**45.96**	**0.4305**	**28.04**	**0.5575**	34.86	**0.3294**

In the inference phase, in order to find out the most common category-irrelevant editing directions, we first back-project $\Delta\mathbf{w}^{ir}$ onto the representation $\hat{\mathbf{n}}$:

$$\hat{\mathbf{n}}_i = \boldsymbol{A}^{-1}\Delta\mathbf{w}_i^{ir}, \tag{15}$$

where $\boldsymbol{A}^{-1}$ is the pseudo-inverse matrix of $\boldsymbol{A}$. Afterward, we count the mean of the absolute value of $|\hat{\mathbf{n}}_i|$ across all M seen categories:

$$\overline{|\hat{\mathbf{n}}|} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N_m}\sum_{i=1}^{N_m}|\hat{\mathbf{n}}_i^{c_m}|, \tag{16}$$

where $\overline{|\hat{\mathbf{n}}|}$ can be interpreted as the commonality of the directions across the whole dataset. For each layer of the $\mathcal{W}^+$ space, we select t directions from $\boldsymbol{A}$ that correspond to top-t values in $\overline{|\hat{\boldsymbol{n}}|}$. The final dictionary for category-irrelevant editing is $\boldsymbol{A}_f \in \mathbb{R}^{18\times512\times t}$.

To automatically generate diverse images, we assume that the sparse representation $\mathbf{n}$ obeys a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$, which is obtained by counting the $\hat{\mathbf{n}}_i$ of all seen categories in the training set. We sample an arbitrary $\tilde{\mathbf{n}}_j$ from $\mathcal{N}(\mu,\Sigma)$ and apply editing to unseen category images. The manipulation intensity α is introduced to control

the diversity of generated images. Given a single image $x_i^{c_n}$, a set of images can be generated with

$$x_j^{c_k} = G(\mathbf{w}_i^{c_k} + \alpha\boldsymbol{A}_f\tilde{\mathbf{n}}_j), \tag{17}$$

where $\mathbf{w}_i^{c_k} = \text{pSp}(x_i^{c_k})$.

4. Experiment

4.1. Implementation Details

In the training stage, we first train a StyleGAN2 [21] with seen categories. Given a trained GAN, the sparse representation encoder is a 5-layer multi-layer perceptron with Leaky-ReLU activation function. The length l of dictionary $\boldsymbol{A}$ is set to 100. For more stable and interpretable editing, we group the 18-layers $\mathcal{W}^+$ space of StyleGAN2 into bottom layers, middle layers, and top layers, corresponding to 0-2, 3-6, and 7-17 layers respectively. Layers in each group share the same sparse representation $\mathbf{n}$.

4.2. Datasets

We evaluate our method on Animal Faces [24], Flowers [26], and VGGFaces [6] following the settings in [13]. **Animal Faces**. We select 119 categories as seen categories for training and 30 as unseen categories for testing.
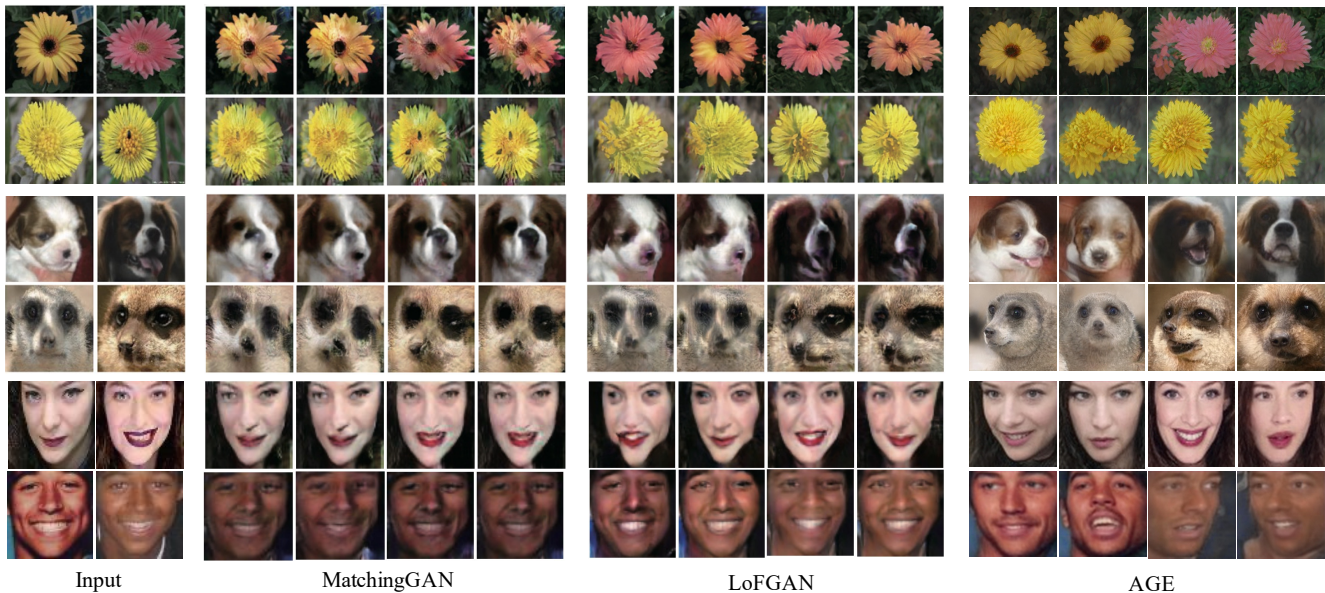
Figure 5. Comparison between images generated by MatchingGAN, LoFGAN, and AGE on Flowers, Animal Faces, and VGGFaces.
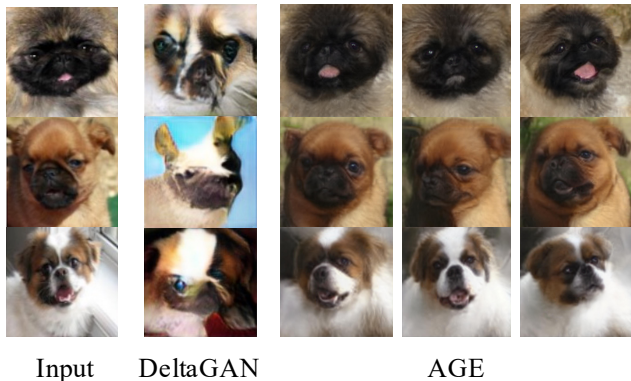

Input   DeltaGAN   AGE

Figure 6. Many failure cases in DeltaGAN can be stably generated by AGE.

Flowers. We split it into 85 seen categories for training and 17 unseen categories for testing.

VGGFaces. For VGGFaces [6], we randomly select 1802 categories for training and 572 for evaluation.

4.3. Ablation Study on Downstream Task

We test data augmentation for image classification on Animal Faces [24]. We randomly select 15, 35, 100 images for each category as train, val, and test, respectively. Following [11], a ResNet-18 backbone is first initialized from the seen categories, then the model is fine-tuned on the unseen categories. 75 images are generated for each unseen category as data augmentation.

"Sample Train" is an ablation that randomly samples $\Delta\mathbf{w}$ of seen categories from the train set and directly used to edit the unseen categories. As shown in Table 3, the directly sampled $\Delta\mathbf{w}$ is unstable for image editing, resulting in crashed images and much higher FID, which proves the

Table 2. Ablations of different manipulation intensity α.

α	Accuracy	FID ($\downarrow$)	LPIPS ($\uparrow$)
Baseline	67.3	–	–
Sample Train	70.8	54.66	0.6103
0.3	69.8	**31.79**	0.5429
0.5	70.6	31.81	0.5482
0.7	70.9	33.43	0.5532
1.0	**71.4**	38.18	0.5609
1.5	69.9	49.70	0.5719
2.0	66.1	63.99	**0.5809**

necessity of the attribute factorization with SDL.

The diversity and quality of generated images are largely controlled by the manipulation intensity α. With the growth of α, AGE generates images of higher diversity but lower quality. According to the accuracy, $\alpha = 1$ achieves the best performance on the classification. This result indicates that generating images that follow the same distribution with the original training set is the best choice for data augmentation. Although lower α can achieve more satisfactory FID and LPIPS, we still decide the best α according to the performance on downstream tasks.

4.4. Quantitative Comparison with State-of-the-art

We evaluate the quality of the generated images based on commonly used FID and LPIPS. Following the former works [11, 13, 14], we generate 128 images based on sampled real images of each unseen category and calculate FID and LPIPS based on the generated images. Following one-shot settings in [3, 13], one real image is used each time to generate adequate images for unseen categories.
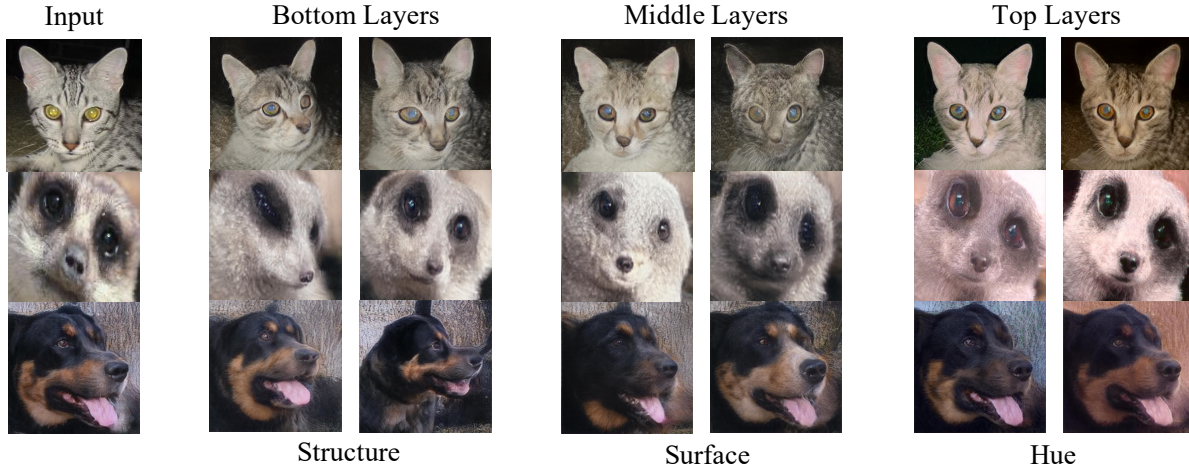
Figure 7. Hierarchical interpretable manipulations discovered in $\boldsymbol{A}_f$.

Figure 8. Manipulate images from different categories with the same sparse representation $\mathbf{n}$.

Figure 9. Manipulate images along disentangled attribute editing directions in different layers in $\boldsymbol{U}$.

The results of different methods are reported in Table 1, our method achieves significant improvements on both FID and LPIPS. Since we do not need to re-train a GAN, AGE is much more stable, achieving impressive FID gain. Compared with fusion-based and transformation-based methods, our generated images are also more diverse. Besides, we can achieve one-shot image generation.

4.5. Qualitative Evaluation

One-shot image generation from AGE on Animal Faces, Flowers and VGGFaces is shown in Figure 4. We qualitatively compare our method with the fusion-based methods MatchingGAN [14] and LoFGAN [11], and transformation-based method DeltaGAN [13]. All images are reported in [11] and [13]. As shown in Figure 5, different from LoF-GAN [11] that can only fuse features from conditional images, AGE can produce images that have new attributes. For example, we can generate images with two flowers and generate dogs of diverse positions and expressions.

Compared with the transformation-based method Delta-GAN that learns intra-category transformations from different image pairs, AGE obtains the latent category-irrelevant
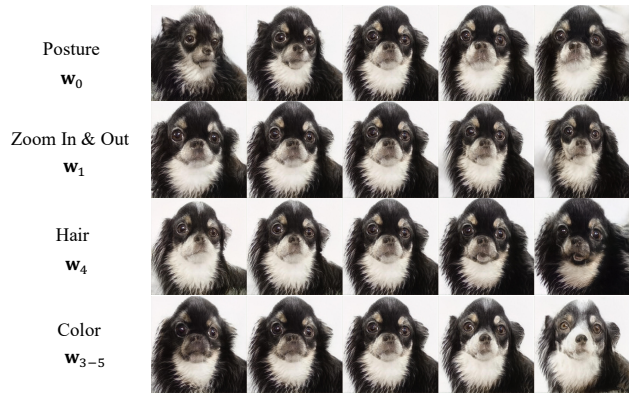
semantics based on more robust class embeddings. Many failure cases in DeltaGAN can be stably generated with AGE as shown in Figure 6. Moreover, since AGE does not retrain a GAN, the generated images from AGE are of much higher quality compared with existing fusion-based and transformation-based methods.

4.6. Semantic Attribute Factorization

Apart from few-shot image generation, an additional advantage for AGE is controllable image editing about the category-irrelevant attributes. In this section, we will experimentally demonstrate the transferability and interpretability of the learned dictionary $\boldsymbol{A}_f$.

Transferability. Since $\boldsymbol{A}_f$ is category-irrelevant, it is transferable across all categories. We edit the images from 4 categories with the same editing direction, the output images are shown in Figure 8. Trans1 is to open the mouth and Trans2 is to turn the head to right. It demonstrates that the dictionary $\boldsymbol{A}_f$ is global and the same $\mathbf{n}$ controls similar attributes for images of different categories.

Figure 10. Image editing without L_{sparse} or L_{otrh}

Interpretability. Although AGE is performed in a completely self-supervised manner, the learned directions in $\boldsymbol{A}_f$ are still interpretable due to the sparse constraint and the meaningful latent space of StyleGAN2.

First, different groups of layers control different category-irrelevant attributes. In particular, we interpret a target model at the levels of bottom layers, middle layers, and top layers, respectively. Figure 7 shows the versatile semantic directions found in Animal Faces. It demonstrates that most directions in $\boldsymbol{A}_f$ is category-irrelevant. Concretely, the bottom layers mainly control the structure of objects, such as the position, zoom in/out, and the shape of the face. The middle layers mainly control the surface features like color and the expressions. The top layers decide the background and the overall hue of the image. We can achieve controllable category-irrelevant editing by sampling in corresponding groups.

To find the disentangled attribute editing directions in different layers/groups, we further conduct Singular Value Decomposition (SVD) on dictionary $\boldsymbol{A}_f$:

$$\boldsymbol{A}_f = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*. \tag{18}$$

Matrix $\boldsymbol{U}$ contains the commonly shared directions of each layer of dictionary $\boldsymbol{A}_f$. Figure 9 is an illustration of image editing by moving along the most salient directions of each layer. Although the disentanglement is rough, moving along a single direction in $\boldsymbol{U}$ can continuously edit one specific category-irrelevant attribute.

4.7. Ablation Study of Loss Components

We conduct an ablation study on L_{orth} and L_{sparse} on Animal Faces [24]. L_{orth} is to encourage category-irrelevance of the learned directions. The edited images are more likely to encounter category change without L_{orth} in Figure 10. As a result, the accuracy of few-shot image classification is obviously lower than AGE without L_{orth} as shown in Table 3. L_{sparse} is for sparsity of the representation $\mathbf{n}$, which is related to the interpretability of the learned directions. Compared with Figure 9, it shows that different semantics are more entangled in the learned directions after SVD without L_{sparse}. Editing image along one direction will cause the change of several attributes.

4.8. Failure Cases and Limitations

There are two major limitations of AGE. The first one comes from the sampling during inference. This sampling

Table 3. Ablation of L_{orth} on few-shot image classification.

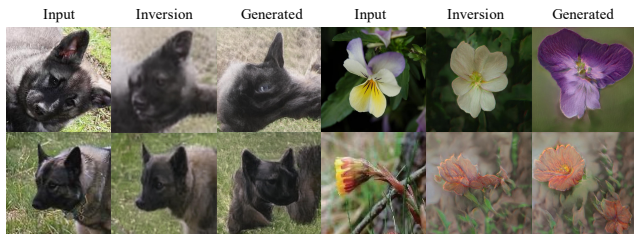| Setting	baseline	w/o L_{orth}	AGE
Accuracy	67.3	69.8	71.4

Figure 11. **Failure Cases**. The failure cases majorly comes from bad GAN inversion, irregular positions of input objects, and wrong category-relevant editing.

is based on the statistics of the training set rather than adaptive to the input images. Although the diversity of images is guaranteed, it will also lead to failure on images of special status. As shown in Figure 11, if the objects are of an irregular posture (*e.g.* dogs in sideways), the generated images are more likely to be crashed. Moreover, some attributes are not category-irrelevant for all categories, but they will be learned in $\boldsymbol{A}_f$. For example, the number of petals is category-irrelevant for most flowers, but it may be identical for some specific categories. Although the generated images are realistic, the category has been changed.

Moreover, the performance of AGE largely relies on the pretrained styleGAN and the inversion method. If the category-relevant attributes of input image can not be well embedded, the editing will also fail. In future, we will try to factorize both category-relevant and irrelevant attributes with better disentanglement, and train the GAN inversion and attribute factorization end to end.

5. Conclusion

In this work, we present a new editing-based perspective for few-shot image generation. In order to achieve category-irrelevant editing on unseen classes, we propose a new method, Attribute Group Editing (AGE), which can identify category-relevant and irrelevant semantics learned by GANs without explicit supervision. Extensive experiments demonstrate the diversity and stability of our method on both few-shot image generation and controllable image manipulation on category-irrelevant attributes.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 3, 4

[2] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 2, 4

[3] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 1, 2, 5, 6

[4] Sergey Bartunov and Dmitry P. Vetrov. Few-shot generative modelling with generative matching networks. In *AISTATS*, 2018. 1, 5

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

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018. 2, 5, 6

[7] Louis Clouâtre and Marc Demers. FIGR: few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019. 1, 2, 5

[8] Kun Fu, Tengfei Zhang, Yue Zhang, Menglong Yan, Zhonghan Chang, Zhengyuan Zhang, and Xian Sun. Meta-ssd: Towards fast adaptation for few-shot object detection with meta-learning. *IEEE Access*, 7:77597–77606, 2019. 1

[9] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. *arXiv preprint arXiv:1906.10112*, 2019. 2, 3

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

[11] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofgan: Fusing local representations for few-shot image generation. In *ICCV*, 2021. 1, 2, 5, 6, 7

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

[13] Y. Hong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Deltagan: Towards diverse few-shot image generation with sample-specific delta. In *CVPR*, 2020. 1, 2, 5, 6, 7

[14] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Matchinggan: Matching-based few-shot image generation. In *ICME*, 2020. 1, 2, 5, 6, 7

[15] Yan Hong, Li Niu, Jianfu Zhang, Weijie Zhao, Chen Fu, and Liqing Zhang. F2gan: Fusing-and-filling gan for few-shot image generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2535–2543, 2020. 1, 2, 5

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

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

[18] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020. 3

[19] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. 2021. 3

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 4

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2, 4, 5

[22] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 2

[23] Weixin Liang, Zixuan Liu, and Can Liu. Dawson: A domain adaptive few shot generation framework. *arXiv preprint arXiv:2001.00576*, 2020. 1, 2, 5

[24] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsueprvised image-to-image translation. In *ICCV*, 2019. 2, 5, 6, 8

[25] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. Supervised dictionary learning. In *NeurIPS*, 2009. 2, 4

[26] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 2, 5

[27] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 2

[28] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*, 2020. 3

[29] Aniwat Phaphuangwittayakul, Yi Guo, and Fangli Ying. Fast adaptive meta-learning for few-shot image generation. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 2

[30] Fabio Pizzati, Jean-François Lalonde, and Raoul de Charette. Manifest: Manifold deformation for few-shot image translation. *arXiv preprint arXiv:2111.13681*, 2021. 2

[31] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 3, 4

[32] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 2020. 2, 3

[33] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 3

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 1

[35] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 3

[36] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016. 1

[37] Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Semi-supervised learning for few-shot image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4453–4462, 2020. 2

[38] Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *ICCV*, 2021. 3

[39] Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. One-shot generative domain adaptation. *arXiv: Computer Vision and Pattern Recognition*, 2021. 2

[40] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *IJCV*, 2020. 2, 3

[41] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 1

[42] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei Efros. Generative visual manipulation on the natural image manifold. 2016. 3