Semi-Supervised Single-Stage Controllable GANs for Conditional Fine-Grained Image Generation

Tianyi Chen[1], Yi Liu[1], Yunfei Zhang[1], Si Wu[1,3*], Yong Xu[1], Feng Liangbing[2], and Hau San Wong[3]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Cosmos Vision Technology Co., Ltd
[3]Department of Computer Science, City University of Hong Kong
{csttychen, csly, cszhangyunfei}@mail.scut.edu.cn, {cswusi, yxu}@scut.edu.cn
lb.feng@cosmosvisiontech.com, cshswong@cityu.edu.hk

Abstract

Previous state-of-the-art deep generative models improve fine-grained image generation quality by designing hierarchical model structures and synthesizing images across multiple stages. The learning process is typically performed without any supervision in object categories. To address this issue, while at the same time to alleviate the level of complexity of both model design and training, we propose a Single-Stage Controllable GAN (SSC-GAN) for conditional fine-grained image synthesis in a semi-supervised setting. Considering the fact that fine-grained object categories may have subtle distinctions and shared attributes, we take into account three factors of variation for generative modeling: class-independent content, cross-class attributes and class semantics, and associate them with different variables. To ensure disentanglement among the variables, we maximize mutual information between the class-independent variable and synthesized images, map real data to the latent space of a generator to perform consistency regularization of cross-class attributes, and incorporate class semantic-based regularization into a discriminator's feature space. We show that the proposed approach delivers a single-stage controllable generator and high-fidelity synthesized images of fine-grained categories. SSC-GAN establishes state-of-the-art semi-supervised image synthesis results across multiple fine-grained datasets.

1. Introduction

Deep generative learning [6, 21, 22, 23, 26, 7] has gained a wide range of research interests due to the high capacity of the generative models in learning complex data distributions. Most of them are based on generative adversarial networks (GANs) [17] and variational autoencoders (VAEs) [27]. Unsupervised or supervised training patterns are typically adopted to achieve remarkable success in image synthesis [16, 43, 29, 44, 11, 12, 1, 2]. However, the resulting generators are either unable to control class semantics or require massive labeled samples. To address this issue, semi-supervised generative learning has been studied [13, 25, 31, 33, 40]. Generic semi-supervised generative modeling is based on the assumption that the amount of unlabeled data is adequate. This does not hold when learning on fine-grained data, due to the reason that both data acquisition and annotation may be expensive and require extensive expertise.

Figure 1. The representative images are synthesized by SSC-GAN in semi-supervised (*top row*) and fully supervised (*middle row*) settings on the CUB dataset [42]. Although only half of training data are labeled in the semi-supervised case, the synthesis quality of SSC-GAN is comparable to that of the model with full supervision, and can be close to the quality of real data (*bottom row*).

Training high-fidelity generators for fine-grained object categories is inherently challenging [48, 47, 49], due to the difficulties in the following aspects: On the one hand, both training samples and labels are insufficient; on the other hand, the distinctions among different categories can be

*Corresponding author.

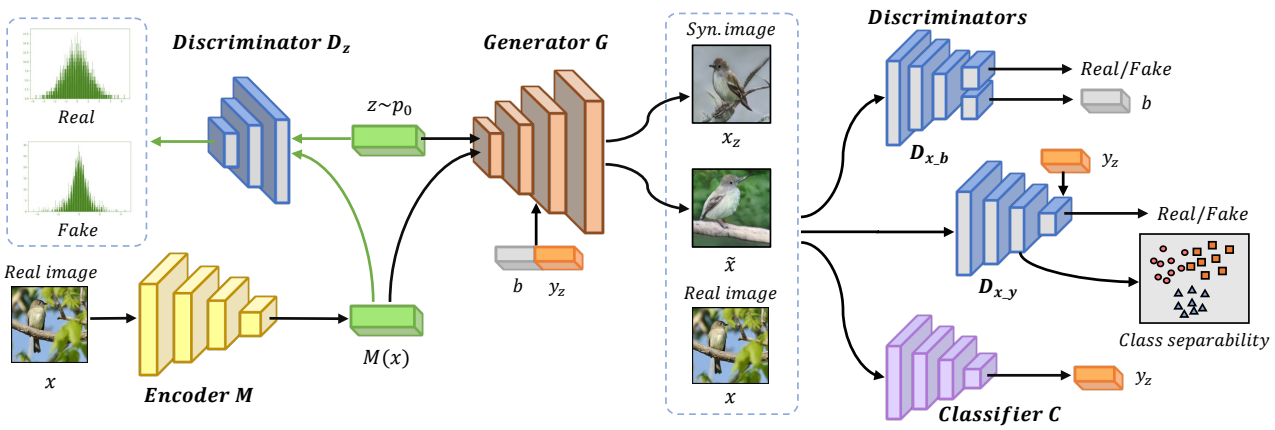


Figure 2. The model structure of SSC-GAN for fine-grained image generation. Generative modeling is performed based on a class-independent variable b, a cross-class variable z and a class variable y_z. An encoder M is incorporated to map images into the latent space of a generator G via adversarial training with a discriminator D_z. On the other hand, the code $M(x)$ is used to synthesize a new image $\tilde{x}$, and z is associated with cross-class attributes by requiring x and $\tilde{x}$ to have similar content independent of b and y_z. In order to make the most of the unlabeled data, an additional discriminator D_{x_b} is incorporated to distinguish real images from fake ones without the condition of class labels, while at the same time to maximize the mutual information between b and the synthesized images. As a result, b is associated with class-independent content. Further, we impose regularization on the feature space of a class-conditional discriminator D_{x_y} to enhance class separability, which is beneficial for class-conditional distribution matching between real and synthesized data.

subtle. To induce a generator to capture the underlying factors which give rise to fine-grained data, previous works [5, 32, 37, 18, 30, 20, 10] adopt hierarchical model structures, and the image generation process consists of multiple stages. To make the generation controllable, different variables are incorporated in different stages to associate with the discovered factors. The model and training complexity can be extremely high. In addition, object-level annotations are usually needed for background and mask generation. More importantly, there is no attempt so far to explicitly control class semantics in fine-grained image synthesis. In this work, we explore an effective way to model the factors of variation without any object-level annotations, while performing class-conditional image generation with limited supervision as shown in Figure 1.

More specifically, we aim to perform semi-supervised class-conditional generative modeling for fine-grained object categories, while at the same time the factors of variations are encoded for generation controllability. We propose a Single-Stage Controllable GAN (SSC-GAN), which learns to synthesize high-fidelity fine-grained images in semi-supervised scenarios. To achieve this goal, fine-grained images are synthesized conditioned on a class-independent variable, a cross-class variable and a class variable. Considering the inadequate amount of training data and labels, the disentanglement of these variables is important for a generator to capture class semantics, and we thus improve a generic semi-supervised GAN-based model in the following three aspects. First, we incorporate an additional discriminator to impose marginal distribution matching between real and synthesized data, while at the same time to maximize the mutual information between the class-independent variable and synthesized images. Second, we leverage an encoder to map images into the latent space of a generator, and generate new images by changing the values of class-independent and class variables. The generator is induced to learn cross-class attributes by minimizing the differences of the latent codes of the original and resulting images. Third, a class label-embedded discriminator is often used for class-conditional distribution alignment. However, the discriminator's features are not necessarily effective for reflecting the distinctions between fine-grained categories. To guide the generator to capture what the class variable essentially represents, we further regularize the discriminator's feature space. The model structure of SSC-GAN is illustrated in Figure 2.

We summarize the main contributions of this work as follows: (1) We propose a semi-supervised GAN-based generative model, SSC-GAN, which is single-stage and controllable for conditional fine-grained image generation. (2) Generative modeling is performed based on a class-independent variable, a cross-class variable and a class variable, which are disentangled by incorporating effective regularizers accordingly without requiring any object-level annotations. (3) An effective solution can be obtained for the optimization problem without heavy tuning.

2. Related Work

Deep generative learning has led to remarkable success in the field of image synthesis. VAE-based [9, 25, 39, 52] models learn data distribution by performing maximum

likelihood estimation, while GAN-based models [6, 21, 22, 23, 51, 50] adopt adversarial learning. In this section, we briefly review the works related to semi-supervised generative learning (SSGL) and fine-grained image synthesis.

2.1. Semi-Supervised Generative Learning

SSGL [13, 25] aims to synthesize high-fidelity conditional images while reducing the dependence of models on labeled data. A common strategy is to make a discriminator play two roles: identifying real and fake images, and inferring class labels of real ones. Springenberg [40] proposed a categorical GAN (CatGAN), in which a discriminator was trained to provide high-confidence class predictions on real instances, while the predicted class probability distributions of fake ones were constrained to be uniform. In [36], Salimans et al. explored a variety of training techniques to improve both training stability and synthesis quality of CatGAN. Further, Wei et al [45] applied Wasserstein GANs [3] to SSGL, and found that the generation performance can benefit from applying Lipschitz continuity regularization on the discriminator's parameters. Another widely used strategy is to incorporate a classifier into the minimax game. Li et al. [31] designed a Triple-GAN model, in which a classifier together with a generator compete with a discriminator by synthesizing label-instance pairs to as realistic an extent as possible. Wu et al. [46] enhanced Triple-GAN by imposing regularization of feature-semantics matching on the generator. On the other hand, Dong and Lin [14] modified Triple-GAN by allowing the generator to compete with both discriminator and classifier. In their model, the classifier was trained to maximize the class margin of real instances while minimizing that of fake ones. To better utilize readily available unlabeled data, Gan et al [15] proposed a Triangle-GAN model, in which an additional discriminator was incorporated to identify the two types of fake label-instance pairs: real unlabeled images with predicted labels, and synthesized image with specified labels. In [33], the generation performance of Triangle-GAN was improved by applying a random regional replacement-based data augmentation strategy to regularize the classifier and discriminator.

2.2. Fine-Grained Image Synthesis

While image synthesis has been widely studied through GAN-based models, conditional generative learning for fine-grained object categories has not been extensively explored. To capture object categories with subtle distinctions, Bao et al. [4] adopted a CVAE-GAN-based fine-grained image generation model, which has the advantages of both conditional VAE (CVAE) [38] and GAN in model training. In [47], Yang et al. presented a layered recursive GAN, in which image background and foreground were generated separately and then stitched to produce a complete fine-grained image. To semantically control synthesized images, a commonly used strategy is to perform inherent disentanglement of a generator's latent space. Chen et al. [8] proposed an InfoGAN model to discover attributes on unlabeled data by imposing mutual information regularization on a GAN's training process. Along this direction, Singh et al [37] developed a hierarchical disentanglement method, which is referred to as FineGAN. Different variables were incorporated into different generation stages to associate with the discovered attributes. Furthermore, Benny and Wolf [5] and Li et al [32] extended FineGAN by modeling more factors of variation and enhancing the generation capability.

The key differences between SSC-GAN and the above GAN-based methods are in terms of task setting and modeling techniques: (1) We focus on semi-supervised fine-grained image synthesis, while the existing fine-grained generative models [4, 37, 5, 32] are based on supervised or unsupervised training strategies. (2) The generation process of FineGAN and variants typically consists of multiple stages, and bounding boxes of objects are needed for background and mask generation. In contrast, SSC-GAN is a single-stage controllable generation model without requiring any object-level annotations. (3) A number of effective regularizers are applied to disentangle the factors of variation, such that we can manipulate the semantics of the synthesized images. However, generic semi-supervised generative models [31, 15, 14, 33] do not possess this capability.

3. Proposed Approach

In a semi-supervised setting, a large amount of unlabeled data $\mathbb{U}$ are observed. In addition, there are a small amount of labeled data $\mathbb{L}$ with $|\mathbb{L}| \ll |\mathbb{U}|$, where the class label of each instance is available. Considering that the fine-grained image synthesis process is typically determined by a number of factors beyond the object category, it is necessary to consider other factors which are not associated with class semantics. In the proposed model, we consider class-independent content, cross-class attribute and class semantics as the factors of variation. We aim at conditional generative modeling, in which an image generator learns to associate the factors with the different variables. In addition to explicitly controlling image synthesis, the synthesized data can better match the statistics of real ones.

3.1. Overview

In SSC-GAN, image generation is controlled by a class-independent variable b, a cross-class variable z and a class variable y_z. To associate the variables to specific semantics of interest, we define the constituent networks of our model as follows: a generator $G : (z, b, y_z) \rightarrow x_z$ synthesizes images conditioned on the variables; an encoder $M : x \rightarrow z$ maps images into G's latent space; a classifier $C : x_z \rightarrow y_z$ infers the class labels of images; a discriminator D_z iden-

tifies z sampled from a prior distribution and generated by M; and two other discriminators D_{x_b} and D_{x_y} distinguish real instances from fake ones without and with class labels, respectively. We perform adversarial training between M and D_z and between G and $\{D_{x\_b}, D_{x\_y}\}$ to match real and fake data distributions. SSC-GAN aims to learn from $\mathbb{L} \bigcup \mathbb{U}$ what semantics the variables essentially represent.

3.2. Conditional Image Synthesis

The generator G has an input variable z and two side input variables b and y_z . Let x_z denote an image synthesized by G from a variable triplet (z, b, y_z), the formulation is expressed as follows:

$$x_z \triangleq G(z, b, y_z), \tag{1}$$

where z is sampled from a prior distribution p_0. For simplicity, the class-independent and class-label codes b and y_z are randomly specified in the form of one-hot vectors. We can also synthesize another type of images as follows:

$$\tilde{x} \triangleq G(M(x), b, y_z), \tag{2}$$

where a real image x is mapped into the G's latent space, and the resulting latent code $M(x)$ is used to synthesize a new image $\tilde{x}$ together with the randomly specified b and y_z. To ensure that the quality of $\tilde{x}$ is as good as x_z, the distribution of $M(x)$ is required to match with p_0. Toward this end, we adopt an adversarial training strategy, in which D_z is trained to identify z sampled from p_0 and generated by M, while M is trained to deceive D_z. The adversarial training loss $\mathcal{L}_{D_z}^{adv}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_{D_z}^{adv} = &\mathbb{E}_{x\sim p_{data}}[\log(1 - D_z(M(x)))] \\ &+ \mathbb{E}_{z\sim p_0}[\log D_z(z)],\end{aligned} \tag{3}$$

where p_{data} denotes the distribution of real images, and $D_z(\cdot)$ represents the predicted probability of a latent code sampled from p_0.

3.3. Regularization for Controlling the Factors

To encourage M to capture cross-class attributes, we explicitly impose constraints between the latent codes of x and $\tilde{x}$, and the corresponding consistency loss $\mathcal{L}_z^{cons}$ is formulated as follows:

$$\mathcal{L}_z^{cons} = \mathbb{E}_{x\sim p_{data}}[\|M(x) - M(\tilde{x})\|_2^2]. \tag{4}$$

Minimizing the consistency between $M(x)$ and $M(\tilde{x})$ is beneficial for disentangling $\{b, y_z\}$ from z, since their latent codes are from the original image, regardless of the variations of other variables.

Considering the issues that the labeled data is limited and the distinction between fine-grained classes can be small, we incorporate two discriminators D_{x_b} and D_{x_y} into SSC-GAN. Both of them are different from the ones used in generic class-conditional image synthesis. More specifically, matching the marginal distributions of real and synthesized data is useful for addressing the issues, since we can use the whole set of real training data to learn class-independent content and cross-class attributes. For instance, different species of birds share similar living environments, shapes and poses. For this purpose, D_{x_b} is introduced to judge whether an image is from real data or synthesized by G without the condition of object category, and the adversarial loss $\mathcal{L}_{D_{x\_b}}^{adv}$ is formulated as follows:

$$\begin{aligned}\mathcal{L}_{D_{x\_b}}^{adv} = &\mathbb{E}_{z\sim p_0}[\log(1 - D_{x\_b}(x_z))] \\ &+ \mathbb{E}_{x\sim p_{data}}[\log(1 - D_{x\_b}(\tilde{x}))] \\ &+ \mathbb{E}_{x\sim p_{data}}[\log D_{x\_b}(x)],\end{aligned} \tag{5}$$

where $D_{x\_b}(\cdot)$ denotes the estimated probability of an image being from real data. On the other hand, a prediction head H_b is built on top of D_{x_b}, and learns to predict the code b of synthesized data, given the features of D_{x_b}. We define the evaluation loss over H_b's predictions as follows:

$$\begin{aligned}\mathcal{L}_b^{sem} = &\mathbb{E}_{z\sim p_0}[\ell(b, H_b(f_{D_{x\_b}}(x_z))] \\ &+ \mathbb{E}_{x\sim p_{data}}[\ell(b, H_b(f_{D_{x\_b}}(\tilde{x}))],\end{aligned} \tag{6}$$

where $f_{D_{x\_b}}(\cdot)$ denotes the features associated with the last hidden layer of D_{x_b}, $H_b(\cdot)$ represents the estimated probability distribution over all possible b values, and $\ell(\cdot, \cdot)$ is the cross entropy function. Minimizing $\mathcal{L}_b^{sem}$ leads to the maximization of the mutual information between b and synthesized images, such that b is enforced to correlate with class-independent content in an unsupervised way.

To match the statistics of real data of each class, G also competes with D_{x_y}, which is used to distinguish the real images from the fake ones, conditioned on the given class labels. In our setting, only a small portion of real images are labeled. When feeding them into D_{x_y}, their labels are determined as follows:

$$y_x = \begin{cases} \texttt{label}(x), & \text{if } x \text{ is labeled,} \\ \texttt{one-hot}(C(x)), & \text{otherwise,} \end{cases} \tag{7}$$

where $C(\cdot)$ denotes the predicted class probability distribution of an unlabeled image by the classifier C. We formulate another adversarial training loss $\mathcal{L}_{D_{x\_y}}^{adv}$ as follows:

$$\begin{aligned}\mathcal{L}_{D_{x\_y}}^{adv} = &\mathbb{E}_{z\sim p_0}[\log(1 - D_{x\_y}(y_z, x_z))] \\ &+ \mathbb{E}_{x\sim p_{data}}[\log(1 - D_{x\_y}(y_z, \tilde{x}))] \\ &+ \mathbb{E}_{x\sim p_{data}}[\log D_{x\_y}(y_x, x)].\end{aligned} \tag{8}$$

Minimizing $\mathcal{L}_{D_{x\_y}}^{adv}$ enforces G to synthesize diverse images that are indistinguishable from real ones on each class. Due to the embedding of class label in D_{x_y}, it focuses on identifying real and synthesized images of the specified class, and

the learnt features are not necessarily effective for reflecting the distinctions between classes. To improve class separability, we incorporate a contrastive constraint to regularize the feature space of D_{x_y} as follows:

$$\mathcal{L}_y^{cntr} = \mathbb{E}_{\substack{x,x'\sim p_{data} \\ y_x=y_{x'}}}[\max(\phi(x,x') - \phi(x,\tilde{x}) + m, 0)], \quad (9)$$

where

$$\phi(x,x') = \|f_{D_{x\_y}}(x) - f_{D_{x\_y}}(x')\|_2^2, \quad (10)$$

$f_{D_{x\_y}}(\cdot)$ represents the features associated with the hidden layer of D_{x_y} before class label embedding, and m denotes a margin that separates the positive pairs (x, x') from the negative ones $(x, \tilde{x})$. Compared to D_{x_y}, the classifier C has a different view to verify the class semantics of synthesized images. To ensure that the synthesized images hold precise class semantics, we further require that they can be correctly recognized by C, and the corresponding evaluation loss $\mathcal{L}_y^{sem}$ is formulated as follows:

$$\mathcal{L}_y^{sem} = \mathbb{E}_{z\sim p_0}[\ell(y_z, C(x_z)] + \mathbb{E}_{x\sim p_{data}}[\ell(y_z, C(\tilde{x})]. \quad (11)$$

Inclusion of $\mathcal{L}_y^{cntr}$ and $\mathcal{L}_y^{sem}$ promotes class separability, which in turn facilitates class-conditional distribution matching between real and synthesized data.

3.4. Model Training

All the constituent networks of SSC-GAN are jointly optimized via adversarial training. For M and G, the overall loss function consists of three adversarial training loss terms associated with three discriminators $\{D_z, D_{x\_b}, D_{x\_y}\}$ and three regularization terms associated with the three variables $\{z, b, y\}$, and the corresponding formulation is presented as follows:

$$\min_{M,G} \mathcal{L}_{D_z}^{adv} + \mathcal{L}_z^{cons} + \mathcal{L}_{D_{x\_b}}^{adv} + \mathcal{L}_{D_{x\_y}}^{adv} + \mathcal{L}_b^{sem} + \mathcal{L}_y^{sem}. \quad (12)$$

To compete with $\{M, G\}$, the optimization formulation of the discriminators is presented as follows:

$$\max_{D_z, D_{x\_b}, D_{x\_y}} \mathcal{L}_{D_z}^{adv} + \mathcal{L}_{D_{x\_b}}^{adv} + \mathcal{L}_{D_{x\_y}}^{adv} - \mathcal{L}_b^{sem} - \mathcal{L}_y^{cntr}. \quad (13)$$

To improve the performance of C, both types of synthesized data $\{x_z, \tilde{x}\}$ are also used to optimize C as well as labeled data, and we formulate the optimization problem as follows:

$$\begin{aligned} \min_C \; & \mathbb{E}_{\substack{x\sim p_{data} \\ x \text{ is labeled}}}[\ell(y_x, C(x)) + \ell(y_z, C(\tilde{x})] \\ & + \mathbb{E}_{x\sim p_{data}}[\mathtt{KL}(\overline{C}(x)||C(x))] \\ & + \mathbb{E}_{z\sim p_0}[\ell(y_z, C(x_z)], \end{aligned} \quad (14)$$

where $\mathtt{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence to measure the difference between the predictions of C and its own aggregated network $\overline{C}$. Since $\overline{C}$ typically provides more reliable predictions than C, it can be used to regularize the network on both labeled and unlabeled data.

4. Experiments

We perform extensive experiments to evaluate the performance of SSC-GAN in disentangling the factors of variation, capturing class semantics, and reducing the dependence on labeled data, by comparing with state-of-the-art generative models in terms of the extent to which the quality of the generated images is enhanced.

4.1. Experimental Setup

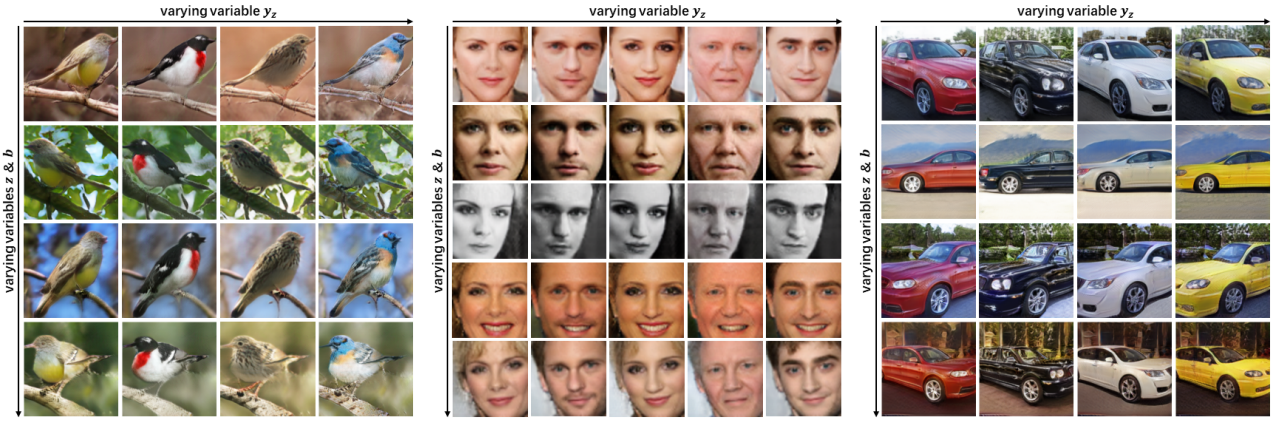
Datasets. The experiments are conducted on diverse fine-grained image datasets: CUB [42], FS-100 [46] and Stanford-Cars [28]. CUB contains about 6K/6K training/test images of resolution 128×128 from 200 bird classes. FaceScrub is a human face dataset. FS-100 consists of about 13K/2K training/test images of resolution 64×64 from the 100 largest classes of FaceScrub [35]. In Stanford-Cars, there are 196 car classes and about 8K/8K images of size 128×128 for training/testing.

Semi-supervised settings. Unless otherwise indicated, we randomly sample 2.8K, 2K and 4K training images to be used as labeled data, and the remaining images are unlabeled for semi-supervised learning on CUB, FS-100 and Stanford-Cars, respectively.

Implementation details. We implement SSC-GAN using PyTorch, and the hardware includes an Intel Core-i7 CPU and a NVIDIA Titan RTX GPU. All the constituent networks are jointly optimized from scratch. The number of training epochs is set to 500, and there are 16/16/16 labeled/unlabeled/synthesized images in each batch. We adopt the Adam optimizer [24] with a learning rate of $\varsigma = 0.0002$ and momentum parameters of $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The hyper-parameter m in Eq.(9) is set to 0.5. We find that the model performance is relatively stable when $m < 1$. We also adopt equal weighting factors of the loss terms in Eqs.(12-14) without heavy tuning.

Baseline. To verify the effectiveness of the adopted improvement techniques, we build a baseline model, which is based on Triple-GAN [31], and performs generic class-conditional image generation without variable disentanglement and related regularization. For fair comparison, we adopt the same backbone architecture as SSC-GAN.

Evaluation protocol. We assess synthesis quality in terms of Inception Score (**IS**) [36] and Fréchet Inception Distance (**FID**) [19]. We also measure the extent to which generated images match with the statistics of real ones on each class, and report the average score of class-wise FIDs (**cFID**). To further verify the class semantics of generated images, we adopt an independent classifier, which is pre-trained with full supervision, to infer their class labels. The class labels specified in the generation process are used as ground truth to calculate the recognition accuracy (**RA**).



(a) CUB (b) FS-100 (c) Stanford-Cars

Figure 3. The representative images are synthesized by SSC-GAN with varying variables z, b and y_z.

4.2. Controllable Image Synthesis

We first assess the SSC-GAN's capability of disentangling the three variables (z, b, y_z) and synthesizing images of fine-grained categories. The synthesized images shown in Figure 3 demonstrate how well the proposed approach control image semantics. The results suggest that the three variables are able to associate with the semantics of our interest, and we find that the associated semantics are human-interpretable. On CUB and Stanford-Cars, the variable $z/b/y_z$ controls object shape and pose/background/object appearance. On FS-100, the variable $z/b/y_z$ encodes facial expression and pose/image style/person's identity.

4.3. Model Analysis

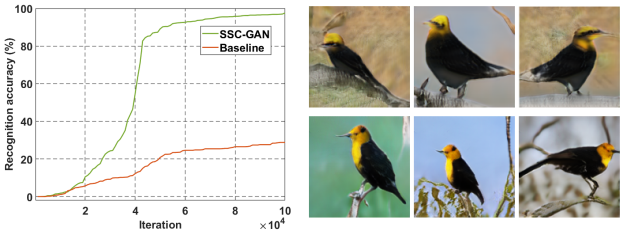
Effectiveness of model components. We quantitatively investigate what contributes to the performance margin between the baseline model and SSC-GAN. In this experiment, the baseline model is progressively enhanced in the following order: the class-semantic regularization $\mathcal{L}_y^{cntr}$ and $\mathcal{L}_y^{sem}$, the discriminator D_{x_b} and variable b-based regularization, and the encoder M and sample $\tilde{x}$-based regularization. The performance of the resulting models are evaluated in terms of four metrics in Table 1. The class-semantic regularization leads to a RA increase from 11.04% to 90.43% on CUB. In Figure 4, we also plot the RA scores of the synthesized images on a representative class in the training process. One can find that SSC-GAN is able to efficiently converge to a much better solution. As a result, the images synthesized by SSC-GAN hold more precise class semantics than the ones synthesized by Baseline. In addition, inclusion of D_{x_b} and M leads to a FID/cFID decrease of about 30/56 points. The results demonstrate that the techniques are effective in improving the class semantics, realism, and diversity of synthesized data.

Encoding cross-class content. To obtain more insights on the encoder M and associated regularization, we visualize the synthesized images $\tilde{x}$, which is based on the latent code $M(x)$ of real reference images x and the randomly specified b and y_z. Figure 5 shows that the synthesized images can have different background and class semantics from the reference images, but the shapes and poses between them are similar. The results suggest that M encodes the content independent of background and class semantics, which is consistent with what the variable z controls.

Associated with class-independent content. To give meaning to the variable b, we incorporate a prediction head H_b on top of the discriminator D_{x_b} to predict the value of b, given a synthesized image. We adopt a class activa-

Table 1. The results of the baseline model and variants on CUB.

Method	FID↓	cFID↓	IS↑	RA↑
Baseline	82.87	196.99	4.42±0.07	11.04
+ Class-semantic Reg.	50.35	157.97	4.45±0.03	90.43
+ D_{x_b} & b-based Reg.	30.48	113.42	4.59±0.05	92.49
+ M & $\tilde{x}$-based Reg.	20.03	101.58	4.68±0.04	97.85
Improvement	***-62.84***	***-95.41***	***+0.26***	***+86.81***



(a) RA curves (b) Synthesized images

Figure 4. Comparison between Baseline and SSC-GAN in synthesizing bird images of a specified class. (a) Recognition accuracy of synthesized images. (b) Representative images generated by Baseline (*upper row*) and SSC-GAN (*bottom row*).

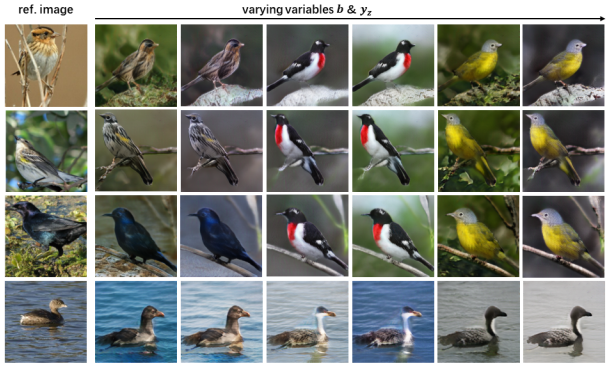


Figure 5. Synthesized images with the latent codes of real reference images (*first column*) and varying variables b and y_z.

tion mapping (CAM) method [53] to visualize the spatial regions where H_b focuses on. In Figure 6, we observe that H_b applies more attention on the background (hot areas).

Improving class-conditional distribution matching. Furthermore, we show the t-SNE [41] embedding of the features $f_{D_{x\_y}}$ associated with the hidden layer of the class-conditional discriminator D_{x_y}, given real labeled images and synthesized images. For simplicity, we randomly select 5 classes of CUB to visualize the data distribution in Figure 7. We find that SSC-GAN performs better than Baseline in aligning real and synthesized data on each class.

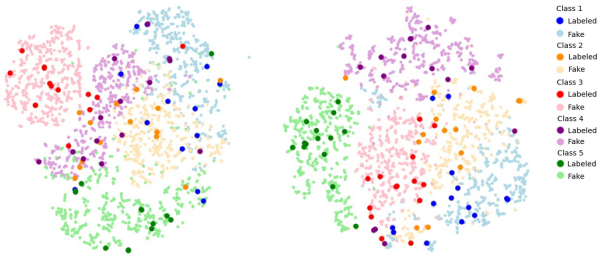
4.4. Comparison with State-of-the-arts

We perform a comparison between SSC-GAN and a number of competing GAN-based generative models without any advanced GAN's training strategies in Table 2.

Unsupervised models. The unsupervised competing methods include SN-GAN [34], FineGAN [37] and MixNMatch [32] as representative generic and fine-grained generative models. The unsupervised models are trained on the same data as SSC-GAN, without using the class labels of labeled data. The fine-grained generative models perform much better than SN-GAN, especially on CUB. Compared to FineGAN and MixNMatch, the superiority of SSC-GAN is still significant. On CUB/FS-100/Stanford-Cars, the FID score of SSC-GAN reaches 20.03/20.65/39.02, which is lower than that of MixNMatch by about 26/5/7 points. As shown in Figure 8, the images synthesized by our model have higher visual quality than FineGAN. This suggests that synthesis quality can greatly benefit from the limited supervision in object categories.

Semi-supervised models. We also conduct a comparison with a number of semi-supervised GANs: Triple-GAN [31], Triangle-GAN [15], EnhancedTGAN [46], and R^3-CGAN [33]. All the competing models are trained in the same semi-supervised setting and experiment configuration as SSC-GAN. Among the test datasets, synthesizing FS-100 images is a relatively easy task, and the FID score of

Figure 6. Examples to visualize where the prediction head H_b focuses on (*bottom row*), given the real images (*upper row*).



(a) Baseline (b) SSC-GAN

Figure 7. t-SNE visualization of real labeled instances and synthesized instances on 5 classes of CUB.

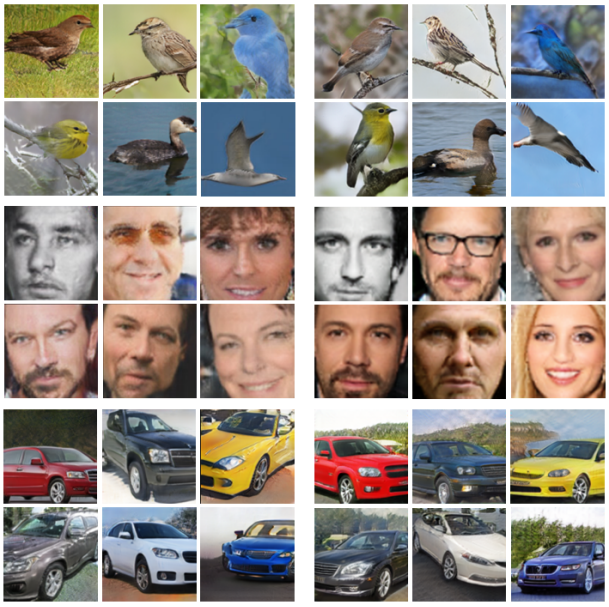
synthesized data drops from the previous best result 25.28 (achieved by R^3-CGAN) to 20.65. On all the three datasets, SSC-GAN is able to achieve lower FID scores and higher IS/RA scores than R^3-CGAN. On CUB, the performance of R^3-CGAN is less satisfactory. We consider that the disentanglement of class semantics and other variation factors is beneficial for capturing fine-grained categories and increasing the diversity of synthesized data.

4.5. Impact of Labeled Data

The existing semi-supervised generative models rarely take the levels of supervision into consideration. To verify that the proposed approach is robust to this factor, we define the proportion of labeled data as $\rho = |\mathbb{L}|/|\mathbb{L} \bigcup \mathbb{U}|$, and conduct experiments on CUB with ρ limited to $\{0.2, 0.3, 0.4, 0.5, 1\}$ ($\rho = 1$ means full supervision). Figure 9 demonstrate that SSC-GAN exceeds the baseline model by a large margin under all the supervision levels. In particular, the performance of the baseline model degrades drastically with $\rho < 0.4$, while the proposed approach performs steadily. This suggests that the adopted improvement strategies are effective in reducing the dependence on labeled data. Further, we train SSC-GAN in the setting of full supervision to provide an upper bound of semi-supervised generative learning. In Table 3, we find that the generation performance of SSC-GAN can be close to that of 'SSC-GAN *w/ Full Sup.*'.

Table 2. Comparison between SSC-GAN and state-of-the-art un(semi-)supervised GAN-based models in fine-grained image synthesis. $*$ indicates that an unsupervised model is trained on the same data as semi-supervised models, without using the class labels of labeled data.

	CUB			FS-100			Stanford-Cars		
Method	FID↓	IS↑	RA↑	FID↓	IS↑	RA↑	FID↓	IS↑	RA↑
SN-GAN* [34]	160.09	4.21±0.05	-	41.26	1.66±0.05	-	53.20	2.80±0.05	-
FineGAN* [37]	46.68	4.62±0.03	-	24.63	1.76±0.02	-	45.72	2.85±0.04	-
MixNMatch* [32]	45.59	**4.78±0.08**	-	25.63	1.71±0.05	-	45.94	2.60±0.05	-
Triple-GAN [31]	140.94	3.94±0.06	9.35	91.05	1.45±0.03	36.21	114.12	2.45±0.06	4.43
EnhancedTGAN [46]	133.57	4.17±0.03	9.16	57.58	1.57±0.02	62.69	105.20	2.43±0.05	3.48
Triangle-GAN [15]	96.42	4.36±0.05	9.01	35.49	1.71±0.04	94.99	61.44	2.77±0.10	4.74
R^3-CGAN [33]	88.62	4.43±0.06	8.60	25.28	1.73±0.02	74.30	44.57	3.05±0.04	5.48
SSC-GAN	**20.03**	4.68±0.04	**97.85**	**20.65**	**1.82±0.03**	**96.86**	**39.02**	**3.10±0.03**	**87.45**



(a) FineGAN (b) SSC-GAN

Figure 8. Visual comparison between FineGAN and SSC-GAN on CUB, FS-100 and Stanford-Cars.

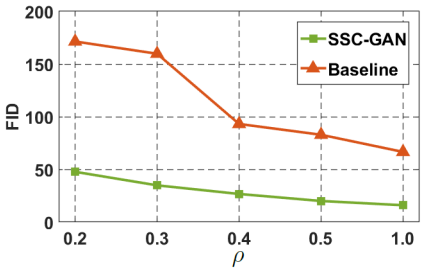


Figure 9. The impact of the levels of supervision on the final generation performance on CUB.

Table 3. The results of SSC-GAN in the semi-supervised and supervised settings.

	CUB		FS-100		Stanford-Cars	
Method	FID↓	RA↑	FID↓	RA↑	FID↓	RA↑
Baseline	82.87	11.04	30.63	88.30	50.78	5.32
SSC-GAN	20.03	97.85	20.65	96.86	39.02	87.45
w/ Full Sup.	*18.34*	*98.54*	*16.28*	*98.35*	*37.48*	*89.54*

5. Conclusion

We focus on class-conditional generative modeling for fine-grained object categories in semi-supervised scenarios, where only a small amount of labeled data can be accessed. Toward this end, we present a single-stage controllable GAN in this paper. Image generation is conditioned on class-independent variable, cross-class variable and class variable to model the factors of variation. We extend the structure of a generic semi-supervised GAN and apply effective regularizers to reduce the dependence of the model on labeled data, as well as enhance class separability. Benefiting from the regularization, the variables are disentangled and associated with corresponding image properties. Our design not only makes the image generation process controllable, but is also beneficial to the matching of the class-conditional distributions of real and synthesized data.

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] Rameen Abdal, Yipeng Qian, and Peter Wonka. Image2StyleGAN: how to embed images into the StyleGAN latent space? In *Proc. International Conference on Computer Vision*, 2019.

[2] Rameen Abdal, Yipeng Qian, and Peter Wonka. Image2StyleGAN++: how to edit the embedded images? In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

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

[5] Yaniv Benny and Lior Wolf. OneGAN: simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. In *Proc. European Conference on Computer Vision*, 2020.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natual image synthesis. In *Proc. International Conference on Learning Representation*, 2019.

[7] Lei Cai, Hongyang Gao, and Shuiwang Ji. Multi-stage variational auto-encoders for coarse-to-fine image generation. In *Proc. International Conference on Data Mining*, pages 630–638, 2019.

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

[9] Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. In *Proc. International Conference on Learning Representation*, 2017.

[10] Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. PuppeteerGAN: arbitrary portrait animation with semantic-aware appearance transformation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: diverse image synthesis for multiple domains. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

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

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

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

[16] Lianli Gao, Junchen Zhu, Jingkuan Song, Feng Zheng, and Heng Tao Shen. Lab2Pix: label-adaptive generative adversarial network for unsupervised image synthesis. *Proc. ACM International Conference on Multimedia*, 2020.

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

[18] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional GANs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

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

[20] Songyao Jiang, Zhiqiang Tao, and Yun Fu. Geometrically editable face image translation with adversarial networks. *IEEE Transactions on Image Processing*, 30:2771 – 2783, 2021.

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

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

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

[24] Diederik P. Kingma and Jimmy Lei Ba. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representation*, 2015.

[25] Diederik P. Kingma, Shakir Mohamed, Danilo J. Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Proc. Neural Information Processing Systems*, 2014.

[26] Diederik P. Kingma, Tim Salimans, Rafal Jzefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational autoencoders with inverse autoregressive flow. In *Proc. Neural Information Processing Systems*, pages 4736–4744, 2016.

[27] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *arXiv:1312.6114*, 2013.

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proc. IEEE Workshop on 3D Representation and Recognition*, 2013.

[29] Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans. In *International Conference on Machine Learning*, pages 3581–3590, 2019.

[30] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: towards diverse and interactive facial image manipulation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

[32] Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. MixNMatch: multifactor disentanglement and encoding for conditional image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

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

[35] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *Proc. IEEE International Conference on Image Processing*, 2014.

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

[37] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. FineGAN: unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[38] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output represenation using deep conditional generative models. In *Proc. Neural Information Processing Systems*, 2015.

[39] Casper Kaae Sonderby, Tapani Raiko, Lars Maaloe, Soren Kaae Sonderby, and Ole Winther. Ladder variational autoencoders. In *Proc. Neural Information Processing Systems*, 2016.

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

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. In *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011.

[43] Jiayu Wang, Wengang Zhou, Guo-Jun Qi, Zhongqian Fu, Qi Tian, and Houqiang Li. Transformation GAN for unsupervised image synthesis and representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[44] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

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

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

[47] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. LR-GAN: layered recursive generative adversarial networks for image generation. In *Proc. International Conference on Learning Representation*, 2017.

[48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. International Conference on Computer Vision*, pages 5907–5915, 2017.

[49] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN++: realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019.

[50] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained GANs for generation with limited data. In *Proc. International Conference on Machine Learning*, 2020.

[51] Shenyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *Proc. Neural Information Processing Systems*, 2020.

[52] Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: information maximizing variational autoencoders. *arXiv:1706.02262*, 2017.

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.