DivCo: Diverse Conditional Image Synthesis via Contrastive Generative Adversarial Network

Rui Liu[†]   Yixiao Ge[†]   Ching Lam Choi[†‡]   Xiaogang Wang[†]   Hongsheng Li[†§]

[†]CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
[‡]NVIDIA AI Technology Center, NVIDIA, Hong Kong   [§]School of CST, Xidian University
`{ruiliu@link, yxge@link, cchoi@link, hsli@ee}.cuhk.edu.hk`

Abstract

Conditional generative adversarial networks (cGANs) target at synthesizing diverse images given the input conditions and latent codes, but unfortunately, they usually suffer from the issue of mode collapse. To solve this issue, previous works [47, 22] mainly focused on encouraging the correlation between the latent codes and their generated images, while ignoring the relations between images generated from various latent codes. The recent MSGAN [27] tried to encourage the diversity of the generated image but only considers "negative" relations between the image pairs.

In this paper, we propose a novel DivCo framework to properly constrain both "positive" and "negative" relations between the generated images specified in the latent space. To the best of our knowledge, this is the first attempt to use contrastive learning for diverse conditional image synthesis. A novel latent-augmented contrastive loss is introduced, which encourages images generated from adjacent latent codes to be similar and those generated from distinct latent codes to be dissimilar. The proposed latent-augmented contrastive loss is well compatible with various cGAN architectures. Extensive experiments demonstrate that the proposed DivCo can produce more diverse images than state-of-the-art methods without sacrificing visual quality in multiple unpaired and paired image generation tasks. Training code and pretrained models are available at https://github.com/ruiliu-ai/DivCo.

1. Introduction

Generative adversarial network (GAN) [9] has shown great potential to capture complex distributions and generate high-dimensional samples since its first introduction in 2014. The follow-up years witnessed its great progresses and successes in synthesizing realistic high-resolution images [35, 28, 1, 18, 19, 3]. Based on GANs, conditional generative adversarial networks (cGANs) [29] were proposed, which focus not only on producing realistic images but more on preserving the input conditional information.
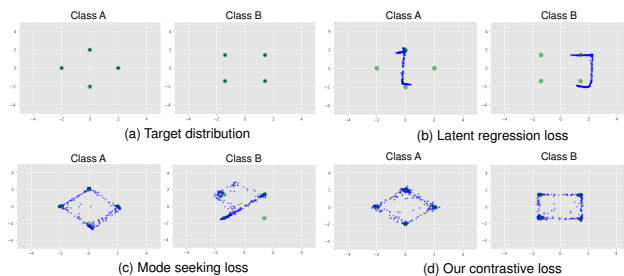

Figure 1. 2D toy experiment for demonstrating the generation distribution learned with different regularization terms. (a) The ground-truth distributions for the two different classes, each of which is a Gaussian mixture model. (b) The generated samples from the learned distribution by the latent regression loss [47] shows its weakness in covering all modes of the ground truths. (c) The generated samples from the learned distribution by the mode seeking loss [27] shows similar patterns regardless of different classes. (d) The generated samples from the learned distribution by our proposed latent-augmented contrastive loss demonstrates an unbiased distribution which is properly dependent on both conditional input and latent codes.

tion. For example, ACGANs [32] generated diverse images conditioned on class labels. Pix2pix [16] and Cycle-GAN [46] translated images across two characteristic domains to change their visual styles under paired and unpaired settings, respectively. There exist other cGANs that condition on input images, *e.g.* super-resolution [21], style transfer [23, 22], inpainting [41], denoising [4], *etc.*, as well as cGANs conditioned on text descriptions [44, 43, 40].

Real-world scenarios expect the synthetic samples to be diverse and able to manipulate flexibly. However, all the applications mentioned above suffer from the problem of mode collapse to some extent, even though randomly sampled latent codes are added as additional inputs. To deal with this shortcoming, many works attempted to enhance the correlation between input latent codes and output images to ensure that the latent codes have control over the generated images. BicycleGAN [47] and DRIT [22] adopted a latent regression loss term, which encourages the mod-

el to recover the input latent code from the generated images. However, the effect of this term on boosting the generation diversity is far from satisfactory (see Fig. 1(b)), which is due to the fact that this term only considers the relation between individual latent codes and their generated images. The relations between images generated from various latent codes are more valuable but were neglected. MSGAN [27] tried to tackle the problem and proposed a mode seeking loss, which aims to improve generation diversity by maximizing the dissimilarity of two arbitrary images. However, the distance between two sampled latent codes may be close to each other and their synthesized images should not be pushed away. By imposing such strong yet sub-optimal constraints between pairs of generated images, the learned distribution easily turns out to be biased, *i.e.* the generation results only depend on latent codes while ignoring the conditional input (see Fig. 1(c)). Such a phenomenon can also be observed in our later experiments (refer to Fig. 4(b) and Fig. 5).

We argue that the unsatisfaction of MSGAN [27] derives from its strategy on always treating any image pairs as "negative" pairs while ignoring "positive" pairs. However, they should be equally crucial for the generator to correctly capture the semantics of various latent codes. Towards this end, we attempt to learn unbiased distributions by considering the "positive" and "negative" relations simultaneously in the form of contrastive learning. Contrastive learning was widely applied in self-supervised representation learning tasks [33, 39, 5, 11] and recently showed its great potential in conditional image synthesis [34, 17] by maintaining the correspondences between each generated image and its conditional input. In this work, we further demonstrate how to adapt it to *diverse* conditional image synthesis by our newly proposed latent-augmented contrastive loss.

Specifically, conditioned on the same class code, the introduced latent-augmented contrastive loss forces the visual relations of the generated images to be correlated to the distances between their input latent codes, *i.e.* "positive" images with close latent codes should be similar while "negative" images with distinct latent codes should be far away from each other in the feature space. Note that the critical view ("positive" or "negative") selection in our proposed framework is achieved via a novel latent augmentation scheme, *i.e.* a positive code is sampled within a small hyper-sphere around the query code in the latent space while negative ones are sampled outside that hyper-sphere. As a result, the problem of mode collapse can be alleviated to large extent by regularizing the generator on better understanding the structure of latent space. Better performance on both learned distributions (Fig. 1(d)) and generation results (Section 4) indicate the effectiveness of our method.

Our contributions are summarized as three-fold:

- We for the first time adapt contrastive learning to encourage diverse image synthesis. The proposed Div-

Co learning scheme can be readily integrated into existing conditional generative adversarial networks with marginal modifications.

- A novel latent-augmented contrastive loss is proposed to discriminate the latent representations of generated samples in a contrastive manner. The issue of mode collapse in cGANs has been much alleviated.

- Extensive experiments in different conditional generation tasks demonstrate that our proposed method helps existing frameworks improve the performance of diverse image synthesis without sacrificing the visual quality of the generated images.

2. Related Work

Conditional generative adversarial network was first presented in [29]. By feeding class labels into generator and discriminator simultaneously, it successfully generate images with respect to the given class labels. To further enhance the relation between input class labels and generated results, ACGAN [32] introduces an auxiliary classifier with cross entropy loss for categorizing both real and generated samples correctly. Similar scheme is also used in semi-supervised GANs [31, 6, 25]. After that, self-attention GAN and its follow-up BigGAN verify the effectiveness of conditional inputs on stabilizing and improving the training of GANs [30, 42, 3]. Similarly, CocoGAN feeds into coordinate conditions into the generator [14]. By changing the modality of conditions into texts, text-to-image generation realizes to synthesize images with respect to the given sentences [44, 43, 40]. Nevertheless, image-to-image translation methods benefit from the development of cGANs most due to the large quantity of applications based on image processing. Pix2pix [16] and CycleGAN [46] translate different styles between two characteristic domains under paired and unpaired settings respectively. There are other applications of cGANs that condition on images such as super-resolution [21], inpainting [41], denoising [4], domain translation [13, 8, 26] and so on [24, 15]. In this work, we endorse these conditional image synthesis tasks with the capability of generating more diverse samples with marginal modifications to their original frameworks.

Contrastive learning was found effective in state-of-the-art self-supervised learning methods [33, 39, 5, 11, 38]. They generally cast the unsupervised visual representation learning as an instance discrimination task, where each unlabeled training sample is treated as a unique class. A contrastive loss [10, 33] is therefore introduced to make representations of different views from the same image closer and keep representations of different images away. How to properly create the positive pairs becomes crucial in these methods, which can be different randomly augmented versions of a single image [5], a same image from various modalities [38], and so on.
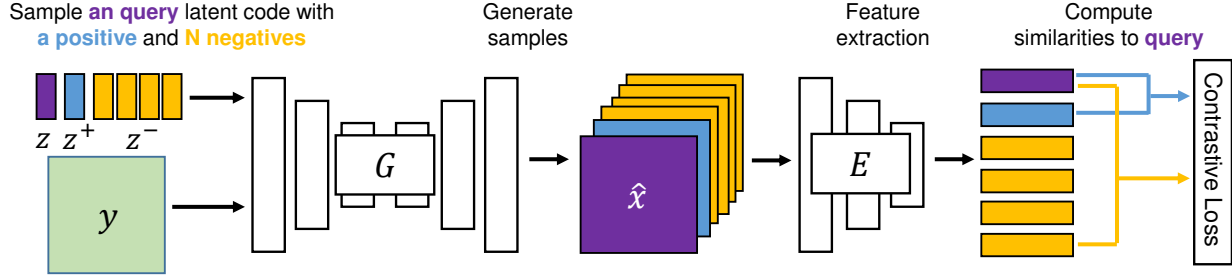
Figure 2. The proposed DivCo framework for image-to-image translation for achieving diverse conditional image synthesis. For a conditional image input y, we sample a query latent code z with a positive code z^+ and N negatives z^- from a prior Gaussian distribution, which are fed into the generator G to obtain the generated samples $\hat{x}$, $\hat{x}^+$, $\{\hat{x}_i^-\}_{i=1}^N$. The generated samples are passed into an auxiliary encoder E to obtain their corresponding feature representations f, f^+, $\{f_i^-\}_{i=1}^N$ and are supervised by the contrastive loss. Here the encoder E shares weights with the encoding layers of the generator G_{enc}. Note that we omit the original cGAN losses for better illustration.

In generative modeling, some recent works investigate the use of contrastive loss for different objectives [34, 2, 17]. CUT [34] aimed to replace cycle-consistency loss with a patch-wise contrastive loss for improving the performance of unpaired image-to-image translation. They considered positive pairs as images patches at the same location in two image domains and negative ones as patches from different locations within a single domain. TUNIT [2] and ContraGAN [17] adopted contrastive losses for preserving the learned style (domain class) codes and categorical labels on the generated images respectively. Note that all the above methods cannot be adopted for diverse image synthesis and cannot deal with mode collapse as well. Our proposed DivCo is a totally different and novel contrastive learning strategy, compared to the existing methods.

3. Method

Given an image dataset $\mathcal{X}$ and its associated condition set $\mathcal{Y}$, we would like to learn a generator G that synthesizes diverse images in the domain of $\mathcal{X}$ following the given conditions $\mathcal{Y}$. The generated images are expected to be realistic and diverse simultaneously. Our proposed DivCo learning scheme can operate regardless of the exact forms of the condition dataset $\mathcal{Y}$, *e.g.* prior images for paired or unpaired image-to-image translation tasks (Fig. 2), class labels for class label-conditional generation tasks (Fig. 3), *etc.* The key innovation of our DivCo lies in the latent-augmented contrastive regularization term, which effectively avoids mode collapse and encourages more diverse conditional image synthesis. Note that this term can be seamlessly integrated into the training of any existing conditional generative adversarial network (cGANs) of different generation tasks without modifying their architectures.

Formally, given an arbitrary condition $y \in \mathcal{Y}$ and a latent code $z \sim \mathcal{N}(0, 1)$ sampled from a prior Gaussian distribution, the generator G takes both y and z as inputs and yields an output image $\hat{x} = G(z, y)$. Ideally, $\hat{x}$ should be determined by the condition y and also be influenced by the
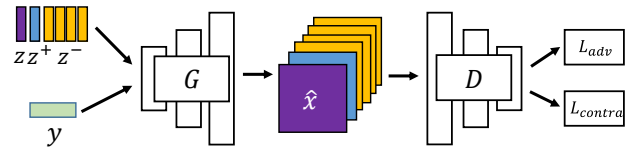

Figure 3. The proposed DivCo framework for class-conditioned image generation. Here y represents conditional class label. Here the encoder E shares weights with the encoding layers of discriminator D.

latent code z.

Adversarial loss. The loss is adopted to ensure that $\hat{x}$ is vivid enough to be indistinguishable from real images $\mathcal{X}$ by a discriminator D. G and D play a minimax game and finally converge to a Nash equilibrium where the synthetic images cannot be distinguished from real images by D any more. The adversarial loss is formulated as

$$\begin{aligned}
\mathcal{L}_{\text{adv}}(G, D) = {} & \mathbb{E}_{x \sim \mathcal{X}, y \sim \mathcal{Y}} \left[\log D(x, y) \right] \\
& + \mathbb{E}_{y \sim \mathcal{Y}, z \sim \mathcal{N}(0,1)} \left[\log \left(1 - D(G(z, y), y) \right) \right].
\end{aligned} \quad (1)$$

However, it can only encourage the realism rather than the diversity of the generated images.

Diverse image synthesis via contrastive learning. We aim to properly regularize cGANs on diverse conditional image synthesis with properly designed learning objectives instead of modifying the original architectures to adapt to different conditional generation tasks.

Intuitively, to avoid mode collapse, we should strengthen the influence of z to the generated image $\hat{x}$ while preserving the fitness of the image $\hat{x}$ to the conditional input y. In other words, the variations of $\hat{x}$ should be intently correlated to the manipulation of z. To achieve this goal, we propose to adopt contrastive learning, where the goal is to associate a query image and its "positive" sample, in contrast to other "negatives" samples at the same time for supervising the image generation process. The most critical design in contrastive learning is how to select the "positive" and "negative" pairs.

Specifically, given a certain condition y, we wish im-

ages generated from close ("positive") latent codes to be visually similar and images from far-away ("negative") latent codes to be visually dissimilar. For a query generated image $\hat{x} = G(z, y)$ generated from a latent code z, we denote its "positive" code as z^+ and the corresponding generated image can be represented as $\hat{x}^+ = G(z^+, y)$. Similarly, the generated images for a series of "negative" latent code $\{z_i^-\}_{i=1}^N$ can be represented as $\hat{x}_i^- = G(z_i^-, y)$ for $i = 1, \ldots, N$. Taking advantages of an auxiliary encoder E (will be discussed), we extract feature representations for generated images, i.e., $f = E(\hat{x})$, $f^+ = E(\hat{x}^+)$, $f_i^- = E(\hat{x}_i^-)$. These representations are expected to be discriminative and can be adopted in contrastive learning to supervise the image generator G.

A novel regularization term, namely *latent-augmented contrastive loss*, is proposed to regularize the generated images' feature representations f, f^+, f^- by bringing f, f^+ closer and spreading f, f^- away. The regularization term is formulated in the form of contrastive loss as

$$\mathcal{L}_{\text{contra}}(G) = \mathbb{E}_{y \sim \mathcal{Y}, z \sim \mathcal{N}(0,1)} \tag{2}$$
$$\left[-\log \frac{\exp\left(\langle f, f^+ \rangle / \tau\right)}{\exp\left(\langle f, f^+ \rangle / \tau\right) + \sum_{i=1}^N \exp\left(\langle f, f_i^- \rangle / \tau\right)} \right],$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two normalized feature vectors to measure their similarity, and τ is a temperature hyper-parameter for scaling the similarity. For each query generated image $\hat{x}$, we generate a positive sample $\hat{x}^+$ and multiple negative samples $\{\hat{x}_i^-\}_{i=1}^N$ in training.

We create positive and negative samples, $\hat{x}^+$ and $\{\hat{x}_i^-\}$, for the latent-augmented contrastive loss by manipulating the input latent codes, $\hat{z}^+$ and $\{\hat{z}_i^-\}$. We observe that the strategy for creating $\hat{z}^+$ and $\{\hat{z}_i^-\}$ has critical impact to the contrastive learning (similar observations can be found in the self-supervised representation learning tasks [5, 11] and image synthesis [34, 2]). We propose a *latent-augmentation* strategy to generate positive and negative samples for contrastive learning on-the-fly. Note that such a newly proposed augmentation strategy is conducted on the latent space rather than the image manifold as existing unsupervised learning methods [5, 11], to adapt to our generation tasks. Specifically, given a query latent code $z \sim \mathcal{N}(0,1)$, we define a small hyper-sphere with radius R centered at the query code z. Consequently, a positive latent code z^+ can be randomly sampled as a vector within the sphere $z^+ = z + \delta$, where δ is a randomly sampled vector from a uniform distribution $\delta \sim \mathcal{U}[-R, R]$, while the negative latent codes can be sampled as random vectors outside the sphere within the latent space, i.e., $z_i^- \sim \{z_i^- \mid z_i^- \sim \mathcal{N}(0,1) \cap |z_i^- - z| \succ R\}$ for $i = 1, \ldots, N$ where $\succ$ is element-wise greater-than operator.

The generator G itself in cGANs is able to encode discriminative feature representations as discussed in [34].

Therefore, we directly adopt the encoding layers of the generator as our auxiliary encoder E, as shown in Fig. 2. Note that since class-conditioned image generation tasks do not require encoding layers in the generator, we use the encoding layers of the conditional discriminator instead, which are also able to encode discriminative features, as shown in Fig. 3. By exploiting the off-the-shelf generators or discriminators for extracting feature representations in the contrastive learning, we could avoid modifying the original architectures of different conditional generation methods, showing the plug-and-play functionality of our method.

Overall training objective. Our proposed latent-augmented contrastive loss can be readily integrated into existing cGANs for different conditional generation tasks, *i.e.* simply adding our proposed latent contrastive loss with their original training objectives. The overall training objective can be formulated as

$$\max_D \min_G \mathcal{L}_{\text{adv}}(G, D) + \lambda_{\text{contra}} \mathcal{L}_{\text{contra}}(G) + \lambda_{\text{opt}} \mathcal{L}_{\text{opt}}(G), \tag{3}$$

where λ_{opt} and λ_{contra} are hyper-parameters balancing different losses in different generation tasks. $\mathcal{L}_{\text{opt}}$ represents an optional regularization term of different conditional generation tasks.

$\mathcal{L}_{\text{opt}}$ is not required for class label-conditional generation tasks. For paired image-to-image translation, the $\mathcal{L}_{\text{opt}}$ regularization term can be formed as a pix-to-pix reconstruction loss

$$\mathcal{L}_{\text{opt}}(G) = \mathbb{E}_{x \sim \mathcal{X}, y \sim \mathcal{Y}, z \sim \mathcal{N}(0,1)} \left[\|G(z, y) - x\|_1 \right]. \tag{4}$$

For unpaired image-to-image translation, $\mathcal{L}_{\text{opt}}$ can be formed as the cyclic reconstruction loss,

$$\mathcal{L}_{\text{opt}}(G) = \mathbb{E}_{x \sim \mathcal{X}, y \sim \mathcal{Y}, z \sim \mathcal{N}(0,1)} \Big[\|F(z, G(z, y)) - y\|_1$$
$$+ \|G(z, F(z, x)) - x\|_1 \Big], \tag{5}$$

where F is an auxiliary generator which learns the inverse mapping of G from $\mathcal{X}$ to $\mathcal{Y}$.

Discussion. In this work, we for the first time investigate adopting contrastive learning for diverse conditional image synthesis. We propose a novel latent augmentation strategy for creating "positive" and "negative" samples in the latent space, while all the previous contrastive losses [11, 5, 34, 2] focus on image-based augmentation strategies that cannot be used for the diverse image synthesis task.

In diverse image synthesis, previous regularization terms [47, 22, 27] on encouraging diversity synthesis failed on jointly modeling positive and negative relations between the generated images, which are sub-optimal solutions. In addition, unlike mode seeking loss that actually constrains the generator to an opposite direction to pixel-level reconstruction losses, which easily breaks down (see Fig. 4), our
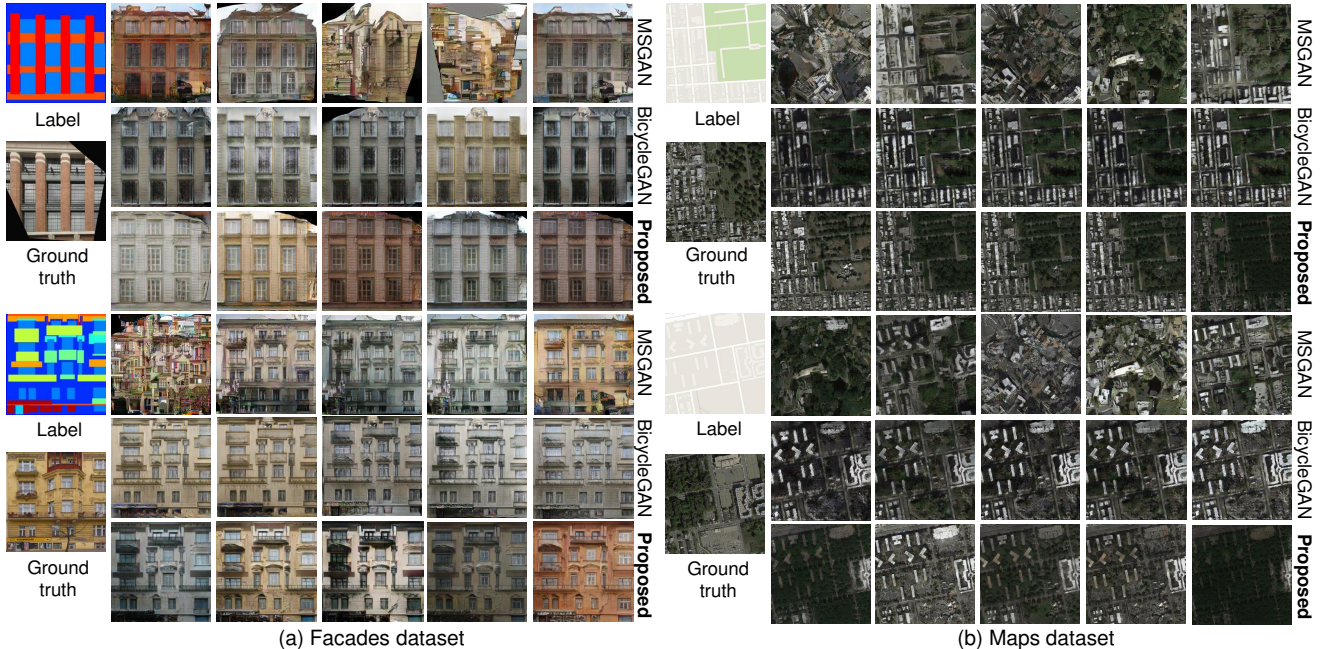
| (a) Facades dataset | (b) Maps dataset |

Figure 4. **Diversity comparison** in the paired image-to-image translation task on Facades and Maps datasets. MSGAN produces diverse images but sacrifices the visual quality and even sometimes cannot condition on the input labels. BicycleGAN produces more realistic images but its results lack diversity. Our proposed method synthesizes both realistic and diverse images.

Table 1. Quantitative results of paired image-to-image translation on Facades and Maps dataset.

Facades			
Metric	BicycleGAN [47]	MSGAN [27]	Proposed
NDB ↓	8.00 ± 1.00	7.60 ± 0.94	$\mathbf{7.30} \pm 0.66$
JSD ↓	0.0427 ± 0.0045	0.0392 ± 0.0088	$\mathbf{0.0374} \pm 0.0098$
LPIPS ↑	0.2355 ± 0.0067	$\mathbf{0.2677} \pm 0.0259$	0.2531 ± 0.0108
FID ↓	80.34 ± 1.46	78.32 ± 3.32	$\mathbf{75.96} \pm 1.25$
Maps			
Metric	BicycleGAN [47]	MSGAN [27]	Proposed
NDB ↓	8.90 ± 1.86	8.60 ± 1.98	$\mathbf{7.80} \pm 2.35$
JSD ↓	0.0956 ± 0.0030	0.1066 ± 0.0023	$\mathbf{0.0806} \pm 0.0021$
LPIPS ↑	0.2473 ± 0.0226	0.2306 ± 0.0287	$\mathbf{0.2538} \pm 0.0382$
FID ↓	130.74 ± 2.63	169.97 ± 8.62	$\mathbf{109.54} \pm 4.41$

latent-augmented contrastive loss can be combined with original cGAN losses in a harmonious manner, leading to better performance in diverse image synthesis.

4. Experiment

4.1. Implementation details

We evaluate our proposed DivCo framework with extensive qualitative and quantitative experiments. We integrate the proposed latent-augmented contrastive loss in three conditional generation tasks, including paired image-to-image translation, unpaired image-to-image translation and class label-conditioned generation. We would maintain their o-

riginal network architectures and supervision signals. We only add our latent-augmented contrastive loss to their optimization objectives and adjust the relative weights between the loss terms. Empirically, the number of negative samples N is set to 10, the balancing weight for latent-augmented contrastive loss λ_{contra} is set to 1 and the temperature τ is set to 1 throughout our experiments if not particularly indicated. For the radius R which depends on the dimension of latent code to some extent, is set to 0.01 in image-to-image translation tasks and set to 0.001 in class-conditioned image generation task.

Baselines. For paired image-to-image translation tasks, we choose BicycleGAN [47] as our baseline. For unpaired image-to-image translation, DRIT [22] is chosen as our baseline. Both of them utilize a latent regression loss for building correlation between latent codes and generated images. We remove their original latent regression loss and then add mode seeking loss [27] and our proposed latent-augmented contrastive loss separately for fair comparison. As for class-conditioned generation, we take DCGAN [35] as our baseline following [27]. Since its generator do not require encoding layers, we take all but the topmost layers of the discriminator as our auxiliary encoder for calculating latent-augmented contrastive loss.

Datasets. We use Facades [7] and Maps [16] datasets for evaluating paired image-to-image translation performance. For unpaired image-to-image translation, we take Yosemite [46] and Dog2Cat [22] datasets for evaluation. CIFAR10 [20] is utilized for evaluating class-conditioned
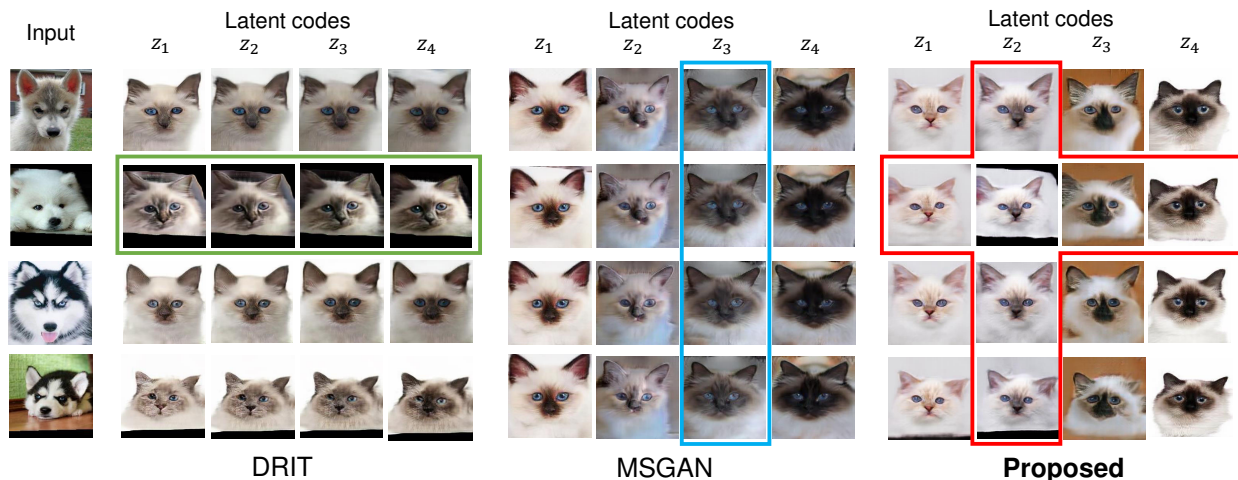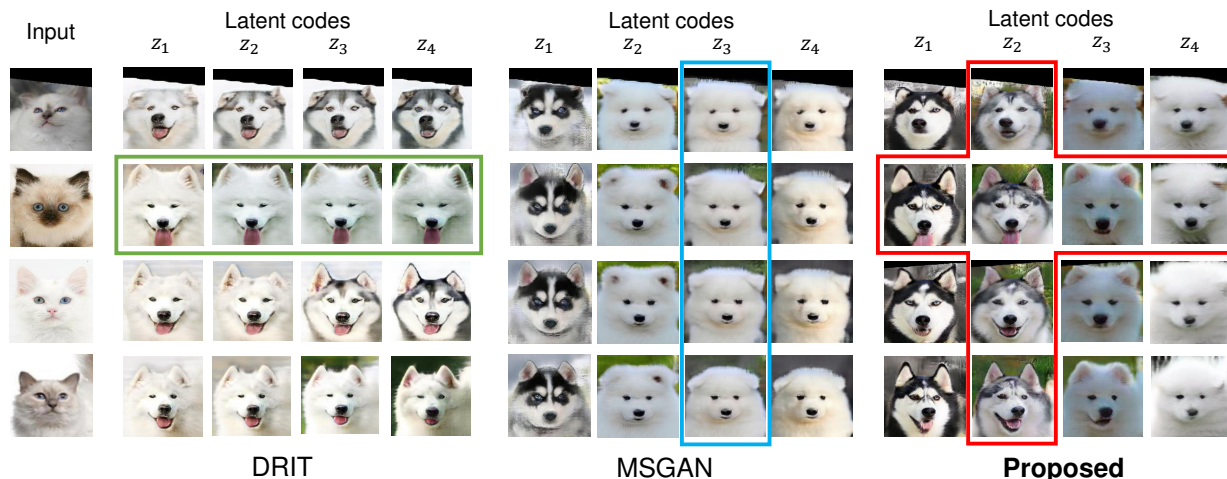
Figure 5. **Diversity Comparison** in the unpaired image-to-image translation task on shape-variant dataset Dog⇔Cat. We manipulate the latent codes from z_1 to z_4 where z_1 is an all-negative-one vector, z_4 is an all-one vector, and z_2 and z_3 are the interpolation of z_1 and z_4 with equal intervals. DRIT produces images without shape variation (mode collapse). MSGAN produces images that are badly conditioned on the input image. Our proposed method synthesizes diverse images by either changing the input condition or manipulating the input latent code.

generation results.

Evaluation metrics. We evaluate the performance of the compared methods via calculating following evaluation metrics. *NDB* and *JSD* are two bin-based evaluation metrics for measuring the similarities between generated distributions and real distributions [36]. Lower value means higher similarity. *Fréchet Inception Distance (FID)*, proposed in [12], is used to measure the quality and diversity of a set of generated images compared to real images by feeding them into an Inception network [37]. Lower value means higher visual quality and higher diversity. *LPIPS* is used to measure the diversity of an image set by computing the average feature distance of all pairs of images [45]. Higher value means higher inter-sample diversity.

4.2. Paired image-to-image translation

We take BicycleGAN [47] as our baseline and separately replace the latent regression loss with mode seeking loss [27] and our proposed latent-augmented contrastive loss for fair comparison. We show our generation results on Facades dataset in Fig. 4(a). The generation results of MSGAN are diverse enough for each input condition, but the visual quality drops drastically. On the contrary, the generation results of BicycleGAN and our method looks more realistic. In addition, our method's results show better generation diversity. Similar comparisons could be found on the Maps dataset, as shown in Fig. 4(b). There is another interesting finding that MSGAN produces similar patterns on

Table 2. Quantitative results of unpaired image-to-image translation on Cat⇔Dog and Yosemite (Winter⇔Summer) dataset.

Dataset	Dog2Cat			Cat2Dog		
	DRIT [22]	MSGAN [27]	Proposed	DRIT [22]	MSGAN [27]	Proposed
NDB ↓	5.20 ± 1.72	4.20 ± 1.32	$\mathbf{3.60} \pm 0.80$	7.7 ± 1.00	6.30 ± 1.13	$\mathbf{5.90} \pm 1.04$
JSD ↓	0.0362 ± 0.0035	0.0285 ± 0.0042	$\mathbf{0.0238} \pm 0.0039$	0.0440 ± 0.0036	0.0381 ± 0.0043	$\mathbf{0.0345} \pm 0.0063$
LPIPS ↑	0.1932 ± 0.0059	0.2057 ± 0.0052	$\mathbf{0.2188} \pm 0.0045$	0.2544 ± 0.0145	0.2615 ± 0.0156	$\mathbf{0.2686} \pm 0.0214$
FID ↓	30.40 ± 1.82	27.46 ± 1.89	$\mathbf{24.83} \pm 2.40$	21.54 ± 2.34	20.78 ± 1.09	$\mathbf{19.25} \pm 2.14$
Dataset	Winter2Summer			Summer2Winter		
	DRIT [22]	MSGAN [27]	Proposed	DRIT [22]	MSGAN [27]	Proposed
NDB ↓	5.80 ± 2.34	5.00 ± 1.56	$\mathbf{4.90} \pm 1.30$	6.30 ± 1.96	5.70 ± 1.01	$\mathbf{5.40} \pm 0.91$
JSD ↓	0.0584 ± 0.0079	0.0401 ± 0.0097	$\mathbf{0.0338} \pm 0.0059$	0.0625 ± 0.0052	0.0595 ± 0.0080	$\mathbf{0.0573} \pm 0.0057$
LPIPS ↑	0.1801 ± 0.0021	0.1816 ± 0.0043	$\mathbf{0.1869} \pm 0.0053$	0.1721 ± 0.0072	0.1795 ± 0.0080	$\mathbf{0.1926} \pm 0.0093$
FID ↓	47.58 ± 2.12	46.37 ± 1.90	$\mathbf{45.79} \pm 2.45$	56.02 ± 2.10	53.32 ± 1.67	$\mathbf{51.57} \pm 1.92$

Figure 6. **Diversity comparison** in the unpaired image-to-image translation task on shape-invariant dataset Winter2Summer. Only the generation results from our proposed method shows the variation from sunshine appearing to totally disappearing, compared to two other methods.

the 3-th and 4-th columns regardless of the input conditional label maps, which is consistent to the toy demonstration in Fig. 1.

The quantitative results on Facades and Maps datasets are recorded in Table 1. As one can see, our method performs best in terms of all the evaluation metrics except for LPIPS on Facades. The reason is that LPIPS only measure the diversity score of a set of images by their average feature distance while the realism is not taken into consideration.

4.3. Unpaired image-to-image translation

In unpaired image-to-image translation, we choose DRIT [22] as our baseline and replace its original latent regression loss with mode seeking loss and our latent con-

trastive loss for encouraging the diversity of the generation results. We choose a shape-variant dataset Cat↔Dog and a shape-invariant dataset Winter↔Summer for comparison.

The synthesized images on Cat↔Dog dataset are shown in Fig. 5. We highlight the advantages of our approach's generation results over other methods by the rectangles in different colors. For the generation results from DRIT, the green rectangle shows its incapability of synthesizing diverse images according to various latent codes. For the results from MSGAN, the blue rectangle highlights its incapability of synthesizing diverse images with different conditional inputs, which could also be observed in Fig. 1 and Fig. 4(b). The synthesized images from our proposed method show much higher diversity than those of previous works, as highlighted in the red cross. Our generated results also fit the different input conditions much better than MSGAN. More diverse generation results can also be observed in Winter2Summer dataset in Fig. 6.

The quantitative results of different approaches on the two datasets are summarized in Table 2. As we see, the proposed method performs best in all evaluation metrics under different conditions compared to two other state-of-the-art diverse image synthesis methods, which demonstrate the effectiveness of our method on this task.

4.4. Class-conditioned image generation

We choose DCGAN [35] as our baseline for comparing the performance on class-conditioned image generation. Since there is no encoder-like architecture in the generator of DCGAN and the dimension of intermediate discriminator feature does not match that of latent code, the latent regression loss is not applicable. Table 3 shows the quantitative results of NDB and JSD for each class. For calculating NDB and JSD, all training samples are first clustered into 50 groups, and each generated sample is then assigned to a nearest group. The NDB score and JSD are then calculated according to the group proportions. We also calculate FID for all generated images, as shown in Table 4. Our method performs best in terms of all these evaluation metrics.

Table 3. NDB and JSD of the generated images by different compared methods on CIFAR10.

Metric	Method	airplane	automobile	bird	cat	deer
NDB↓	DCGAN [35]	17.90 ± 1.56	18.40 ± 1.82	13.80 ± 2.36	14.50 ± 1.90	18.90 ± 2.18
	MSGAN [27]	17.40 ± 2.80	17.10 ± 2.07	14.10 ± 2.62	16.50 ± 2.85	20.70 ± 2.45
	Proposed	$\mathbf{16.30} \pm 2.75$	$\mathbf{16.90} \pm 1.92$	$\mathbf{12.80} \pm 2.14$	$\mathbf{10.70} \pm 2.49$	$\mathbf{14.60} \pm 1.96$
JSD↓	DCGAN [35]	0.0098 ± 0.0001	0.0125 ± 0.0001	0.0082 ± 0.0001	0.0086 ± 0.0001	0.0069 ± 0.0001
	MSGAN [27]	0.0095 ± 0.0001	0.0121 ± 0.0001	0.0083 ± 0.0002	0.0154 ± 0.0001	0.0072 ± 0.0001
	Proposed	$\mathbf{0.0091} \pm 0.0001$	$\mathbf{0.0119} \pm 0.0001$	$\mathbf{0.0076} \pm 0.0001$	$\mathbf{0.0059} \pm 0.0002$	$\mathbf{0.0058} \pm 0.0001$
		dog	frog	horse	ship	truck
NDB↓	DCGAN [35]	14.50 ± 1.83	24.00 ± 1.55	24.60 ± 2.36	18.50 ± 2.51	15.90 ± 1.33
	MSGAN [27]	11.20 ± 1.16	22.30 ± 2.90	22.90 ± 1.83	16.40 ± 3.36	15.00 ± 2.69
	Proposed	$\mathbf{11.00} \pm 2.24$	$\mathbf{20.20} \pm 2.22$	$\mathbf{20.90} \pm 2.76$	$\mathbf{16.30} \pm 2.29$	$\mathbf{14.60} \pm 2.28$
JSD↓	DCGAN [35]	0.0088 ± 0.0001	0.0150 ± 0.0001	0.0171 ± 0.0001	0.0112 ± 0.0001	0.0088 ± 0.0001
	MSGAN [27]	0.0072 ± 0.0003	0.0134 ± 0.0001	0.0154 ± 0.0001	0.0096 ± 0.0001	0.0084 ± 0.0001
	Proposed	$\mathbf{0.0058} \pm 0.0002$	$\mathbf{0.0124} \pm 0.0001$	$\mathbf{0.0126} \pm 0.0001$	$\mathbf{0.0085} \pm 0.0001$	$\mathbf{0.0078} \pm 0.0002$

Table 4. FID of the generated images by different compared methods on CIFAR10 dataset.

Method	DCGAN [35]	MSGAN [27]	Proposed
FID ↓	28.34 ± 0.12	27.09 ± 0.14	$\mathbf{25.21} \pm 0.20$

4.5. Sensitivity Analysis

We perform sensitivity analysis for 3 hyper-parameters in our framework: weight of latent-augmented loss λ_{contra}, temperature for scaling similarity τ, and the radius of hyper-sphere R in the class-conditioned image generation task. We empirically set $\lambda_{\text{contra}} = 1$, $\tau = 1$, and $R = 0.001$. We take FID as our metric for balancing both realism and diversity. The FID results are plotted in Fig. 7.

As we can see, when the weight λ_{contra} decreases, the performance drops slightly and gradually approach the FID value of original DCGAN (see Table 4). When λ_{contra} is greater than 3, the performance drops drastically as λ_{contra} increases, as $\mathcal{L}_{\text{contra}}$ overwhelms the original cGAN loss, leading to images of lower realism. The larger the radius of hyper-sphere R is, the larger the distance between query and positive latent codes tends to be, leading to a more biased definition of "positive" and "negative". However, when R becomes smaller and gradually approaches 0, the effect of modeling positive relations between images tends to be weaker and make the proposed contrastive loss performs similarly to the mode seeking loss in MSGAN. As for the temperature τ, we find that only a proper value leads to the best performance. Either higher or lower values would make the latent-augmented contrastive loss generate inferior results.
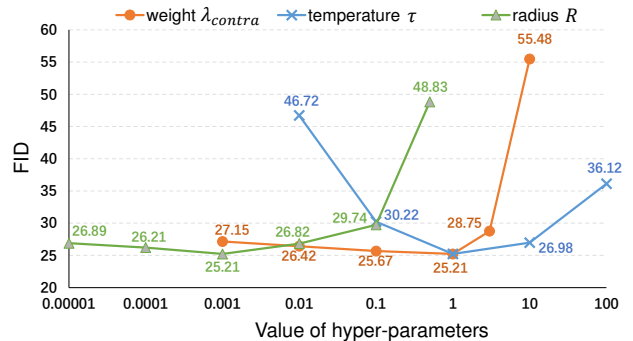
Figure 7. Sensitivity analysis for the relative weight of latent-augmented loss λ_{contra}, temperature for scaling similarity τ, and the radius of hyper-sphere R under class-conditioned image generation task.

5. Conclusion

In this paper we propose a novel framework DivCo for dealing with mode collapse issue in conditional generative adversarial networks. We adapt contrastive learning to diverse image synthesis and propose a novel latent-augmented contrastive loss. The proposed method could be readily integrated into existing GAN-based conditional generation approaches by integrating the latent-augmented contrastive regularization term to their original optimization objective. Experiments on multiple settings with various datasets verify the superiority of our method to other state-of-the-arts, both qualitatively and quantitatively.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 1

[2] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. *arXiv: 2006.06500*, 2020. 3, 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 2

[4] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *CVPR*, 2018. 1, 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 4

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016. 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 2

[9] 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

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 4

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[14] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *ICCV*, 2019. 2

[15] A. Ignatov, J. Patel, R. Timofte, B. Zheng, X. Ye, L. Huang, X. Tian, S. Dutta, K. Purohit, P. Kandula, M. Suin, A. N. Rajagopalan, Z. Xiong, J. Huang, G. Dong, M. Yao, D. Liu, M. Hong, W. Lin, Y. Qu, J. Choi, W. Park, M. Kim, R. Liu, X. Mao, C. Yang, Q. Yan, W. Sun, J. Fang, M. Shang, F. Gao, S. Ghosh, P. K. Sharma, A. Sur, and W. Yang. Aim 2019 challenge on bokeh effect synthesis: Methods and results. In *ICCVW*, 2019. 2

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 5

[17] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In *NeurIPS*, 2020. 2, 3

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 5

[21] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2

[22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 1, 4, 5, 7

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*. 2017. 1

[24] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *arXiv: 2008.02793*, 2020. 2

[25] Rui Liu, Yu Liu, Xinyu Gong, Xiaogang Wang, and Hongsheng Li. Conditional adversarial generative flow for controllable image synthesis. In *CVPR*, 2019. 2

[26] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *CVPR*, 2020. 2

[27] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019. 1, 2, 4, 5, 6, 7, 8

[28] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *arXiv preprint arXiv:1611.04076*, 2016. 1

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1, 2

[30] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 2

[31] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv: 1606.01583*, 2016. 2

[32] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 1, 2

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[34] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for conditional image synthesis. In *ECCV*, 2020. 2, 3, 4

[35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 1, 5, 7, 8

[36] Eitan Richardson and Yair Weiss. On gans and gmms. In *NeurIPS*, 2018. 6

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2

[39] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. In *CVPR*, 2018. 2

[40] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. 2018. 1, 2

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 1, 2

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Spectral normalization for generative adversarial networks. In *ICML*, 2019. 2

[43] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv: 1710.10916*, 2017. 1, 2

[44] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 1, 2

[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2, 5

[47] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 1, 4, 5, 6