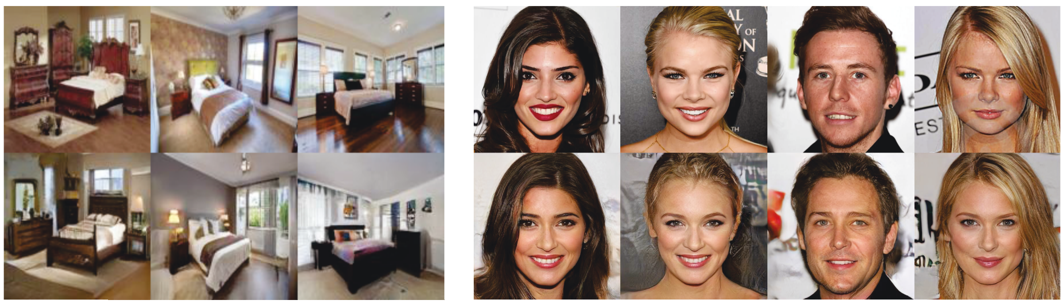
Dual Contradistinctive Generative Autoencoder

Gaurav Parmar*[1] Dacheng Li*[1] Kwonjoon Lee*[2] Zhuowen Tu[2]
[1]Carnegie Mellon University [2]UC San Diego
{gparmar,dacheng2}@andrew.cmu.edu {kwl042,ztu}@ucsd.edu



(a) DC-VAE (ours) Reconstruction results. Left: 128 × 128. Right: 512 × 512.

(b) DC-VAE (ours) Sampling results. Left: 128 × 128. Right: 512 × 512.

Figure 1: DC-VAE Reconstruction (top) and Sampling (bottom) on LSUN Bedroom [68] at resolution 128×128 (left) and CelebA-HQ [35] at resolution 512×512 (right).

Abstract

We present a new generative autoencoder model with dual contradistinctive losses to improve generative autoencoder that performs simultaneous inference (reconstruction) and synthesis (sampling). Our model, named dual contradistinctive generative autoencoder (DC-VAE), integrates an instance-level discriminative loss (maintaining the instance-level fidelity for the reconstruction/synthesis) with a set-level adversarial loss (encouraging the set-level fidelity for the reconstruction/synthesis), both being contradistinctive. Extensive experimental results by DC-VAE across different resolutions including 32×32, 64×64, 128×128, *and* 512×512 *are reported. The two contradistinctive losses in VAE work harmoniously in DC-VAE leading to a significant qualitative and quantitative performance enhancement over the baseline VAEs without architectural changes. State-of-the-art or competitive results among generative autoencoders for image reconstruction, image synthesis, image interpolation, and representation learning are observed. DC-VAE is a general-purpose VAE model, applicable to a wide variety of downstream tasks in computer vision and machine learning.*

1. Introduction

Tremendous progress has been made in deep learning for the development of various learning frameworks [40, 24, 17, 64]. Autoencoder (AE) [44, 27] aims to compactly represent and faithfully reproduce the original input signal by concatenating an encoder and a decoder in an end-to-end learning framework. The goal of AE is to make the encoded representation semantically *efficient* and *sufficient* to reproduce the input signal by its decoder. Autoencoder's generative companion, variational autoencoder (VAE) [38], additionally learns a variational model for the latent variables to capture the underlying sample distribution.

The key objective for a generative autoencoder is to maintain two types of fidelities: (1) an *instance-level fidelity* to make the reconstruction/synthesis faithful to the individual input data sample, and (2) a *set-level fidelity* to make the reconstruction/synthesis of the decoder faithful to the entire input data set. The VAE/GAN algorithm [42] combines a

* indicates equal contribution.
Code: https://github.com/mlpc-ucsd/DC-VAE.

reconstruction loss (for instance-level fidelity) with an adversarial loss (for set-level fidelity). However, the result of VAE/GAN is sub-optimal, as shown in Table 1.

The pixel-wise reconstruction loss in the standard VAE [38] typically results in blurry images with degenerated semantics. A possible solution to resolving the above conflict lies in two aspects: (1) turning the measure in the pixel space into induced feature space that is more semantically meaningful; (2) changing the L2 distance (per-pixel) into a *learned instance-level* distance function for the entire image (akin to generative adversarial networks which learn *set-level* distance functions). Taking these two steps allows us to design an instance-level classification loss that is aligned with the adversarial loss in the GAN model enforcing set-level fidelity. Motivated by the above observations, we develop a new generative autoencoder model with dual contradistinctive losses by adopting a discriminative loss performing instance-level classification (enforcing the instance-level fidelity), which is rooted in metric learning [41] and contrastive learning [21, 66, 63]. Combined with the adversarial losses for the set-level fidelity, both terms are formulated in the induced feature space performing contradistinction: (1) the instance-level contrastive loss considers each input instance (image) itself as a class, and (2) the set-level adversarial loss treats the entire input set as a positive class. We name our method dual contradistinctive generative autoencoder (DC-VAE) and make the following contributions:

- We develop a new algorithm, dual contradistinctive generative autoencoder (DC-VAE), by combining instance-level and set-level classification losses in the VAE framework, and systematically show the significance of these two loss terms in DC-VAE;
- The effectiveness of DC-VAE is illustrated in a number of tasks, including image reconstruction, image synthesis, image interpolation, and representation learning by reconstructing and sampling images across different resolutions including 32×32, 64×64, 128×128, and 512×512;
- Under the new loss term, DC-VAE attains a significant performance boost over the competing methods without architectural change, thus potentially providing a handy solution for many AE/VAE based modeling/representation applications. DC-VAE helps greatly reducing the performance gap for image synthesis between the baseline VAE to the competitive GAN models.

2. Related Work

Related work can be roughly divided into three categories: (1) generative autoencoder, (2) deep generative model, and (3) contrastive learning.

Generative autoencoder. Variational autoencoder (VAE) [38] points to an exciting direction of generative models by developing an Evidence Lower BOund (ELBO) objective. However, the VAE reconstruction/synthesis is known to be blurry. To improve the image quality, a sequence of VAE based models have been developed [42, 15, 30, 3, 69]. VAE/GAN [42] adopts an adversarial loss to improve the quality of the image, but its output for both reconstruction and synthesis (new samples) is still unsatisfactory. IntroVAE [30] adds a loop from the output back to the input and is able to attain image quality that is on par with some modern GANs in some aspects. However, its full illustration for both reconstruction and synthesis is unclear. PGA [69] adds a constraint to the latent variables.

Deep generative model. Pioneering works of [61, 20] alleviate the difficulty of learning densities by approximating likelihoods via classification (real (positive) samples vs. fake (pseudo-negative or adversarial) samples). Generative adversarial network (GAN) [17] builds on neural networks and amortized sampling (a decoder network that maps a noise into an image). The subsequent development after GAN [56, 2, 18, 35, 16] has led to a great leap forward in building decoder-based generative models. It has been widely observed that the adversarial loss in GANs contributes significantly to the improved quality of image synthesis. InfoGAN [9] aims to maximize mutual information between latent variables and the output of the generator. In contrast, DC-VAE can be understood as increasing mutual information between the input and output of the autoencoder. Furthermore, InfoGAN does not learn encoder. ALI/BiGAN [15, 14] additionally learn inference network (encoder), however they are shown to have poor reconstruction. Energy-based generative models [57, 67, 43, 32, 47] — which aim to directly model data density — are making a steady improvement for a simultaneously generative and discriminative single model.

Contrastive learning. From another angle, contrastive learning [21, 66, 23, 6] has lately shown its particular advantage in unsupervised training of CNN features. It overcomes the limitation in unsupervised learning where class label is missing by turning each image instance into one class. Thus, the softmax function in the standard discriminative classification training can be applied. Contrastive learning can be connected to metric learning [4, 10, 5].

In this paper, we aim to improve VAE [38] by introducing a contrastive loss [63] to address instance-level fidelity between the input and the reconstruction in the induced feature space. Unlike in self-supervised representation learning methods [63, 23, 6], where self-supervision requires generating a transformed input (via data augmentation operations), the reconstruction naturally fits into the contrastive term that encourages the matching between the reconstruction and the input image instance, while pushing the reconstruction away from the rest of the images in the entire training set. Thus, the instance-level and set-level contradistinctive terms collaborate with each to encourage the high fidelity of the reconstruction and synthesis. In Figure 3 and Table 1, we systematically show the significance of with and without the instance-level and the set-level contradistinctive terms. In addition, we explore multi-scale contrastive learning via two schemes in Section 4.2: 1) deep supervision for contrastive learning in different convolution layers, and 2) patch-based contrastive learning for fine-grained data fidelity. In the experiments, we show competitive results for the proposed DC-VAE in a number of benchmarks for three tasks, including image synthesis, image reconstruction, and representation learning.

3. Preliminaries: VAE and VAE/GAN

Variational autoencoder (VAE) Assume a given training set $S = \{\mathbf{x}_i\}_{i=1}^n$ where each $\mathbf{x}_i \in \mathbb{R}^m$. We suppose that each $\mathbf{x}_i$ is sampled from a generative process $p(\mathbf{x}|\mathbf{z})$. In the literature, vector $\mathbf{z}$ refers to latent variables. In practice, latent variables $\mathbf{z}$ and the generative process $p(\mathbf{x}|\mathbf{z})$ are unknown. The objectives of a variational autoencoder (VAE) [38] is to simultaneously train an inference network $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and a generator network $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. In VAE [38], the inference network is a neural network that outputs parameters for Gaussian distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_{\boldsymbol{\phi}}(\mathbf{x}), \Sigma_{\boldsymbol{\phi}}(\mathbf{x}))$. The generator is a deterministic neural network $f_{\boldsymbol{\theta}}(\mathbf{z})$ parameterized by $\boldsymbol{\theta}$. Generative density $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is assumed to be subject to a Gaussian distribution: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(f_{\boldsymbol{\theta}}(\mathbf{z}), \sigma^2 I)$. These models can be trained by minimizing the **negative** of evidence lower bound (ELBO) in Eq. (1) below.

$$\begin{aligned}&\mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \\ &-\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log(p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}))] + KL[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]\end{aligned} \tag{1}$$

where $p(\mathbf{z})$ is the prior, which is assumed to be $\mathcal{N}(0, I)$. The first term $-\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log(p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}))]$ reduces to standard pixel-wise reconstruction loss $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[||\mathbf{x} - f_{\boldsymbol{\theta}}(\mathbf{z})||_2^2]$ (up to a constant) due to the Gaussian assumption. The second term is the regularization term, which prevents the conditional $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ from deviating from the Gaussian prior $\mathcal{N}(0, I)$. The inference network and generator network are jointly optimized over training samples by:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})} \mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}). \tag{2}$$

where p_{data} is the distribution induced by the training set S.

VAE has an elegant formulation. However, it relies on a pixel-wise reconstruction loss, which is known not ideal to be reflective of perceptual realism [33, 31], often resulting in blurry images. From another viewpoint, it can be thought of as using a kernel density estimator (with an isotropic Gaussian kernel) in the pixel space. Although allowing efficient training and inference, such a non-parametric approach is overly simplistic for modeling the semantics and perception of natural images.

VAE/GAN Generative adversarial networks (GANs) [17] and its variants [56], on the other hand, are shown to be producing highly realistic images. The success was largely attributed to learning a fidelity function (often referred to as a discriminator) that measures how realistic the generated images are. This can be achieved by learning to contrast (classify) the set of training images with the set of generated images [61, 20, 17].

VAE/GAN [42] augments the ELBO objective (Eq. (2)) with the GAN objective. Specifically, the objective of VAE/GAN consists of two terms, namely the modified ELBO (Eq. (3)) and the GAN objective. To make the notations later consistent, we now define the set of given training images as $S = \{\mathbf{x}_i\}_{i=1}^n$ in which a total number of n unlabeled training images are present. For each input image $\mathbf{x}_i$, the modified ELBO computes the reconstruction loss in the *feature space* of the discriminator instead of the pixel space:

$$\begin{aligned}&\mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i) = \\ &\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}[||F_D(\mathbf{x}_i) - F_D(f_{\boldsymbol{\theta}}(\mathbf{z}))||_2^2] + KL[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})]\end{aligned} \tag{3}$$

where $F_D(\cdot)$ denotes the feature embedding from the discriminator D. Feature reconstruction loss (also referred to as perceptual loss), similar to that in style transfer [33]. The modified GAN objective considers both reconstructed images (latent code from $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$) and sampled images (latent code from the prior $p(\mathbf{z})$) as its fake samples:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i) = \log D(\mathbf{x}_i) + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\log(1 - D(f_{\boldsymbol{\theta}}(\mathbf{z}))\\ + \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}\log(1 - D(f_{\boldsymbol{\theta}}(\mathbf{z})).\end{aligned} \tag{4}$$

The VAE/GAN objective becomes:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \max_{D} \sum_{i=1}^{n} \left[\mathcal{L}_{\text{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i) + \mathcal{L}_{\text{GAN}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i)\right]. \tag{5}$$

4. Dual contradistinctive generative autoencoder (DC-VAE)

Here we want to address a question: *Is the degeneration of the synthesized images by VAE always the case once the decoder is joined with an encoder?* Can the problem be remedied by using a more informative loss?

Although improving the image qualities of VAE by integrating a set-level contrastive loss (GAN objective of Eq. (4)), VAE/GAN still does not accurately model instance-level fidelity. Inspired by the literature on instance-level classification [51], approximating likelihood by classification [61], and contrastive learning [21, 66, 23], we propose to model instance-level fidelity by contrastive loss (commonly referred to as InfoNCE loss) [63]. In DC-VAE, we perform

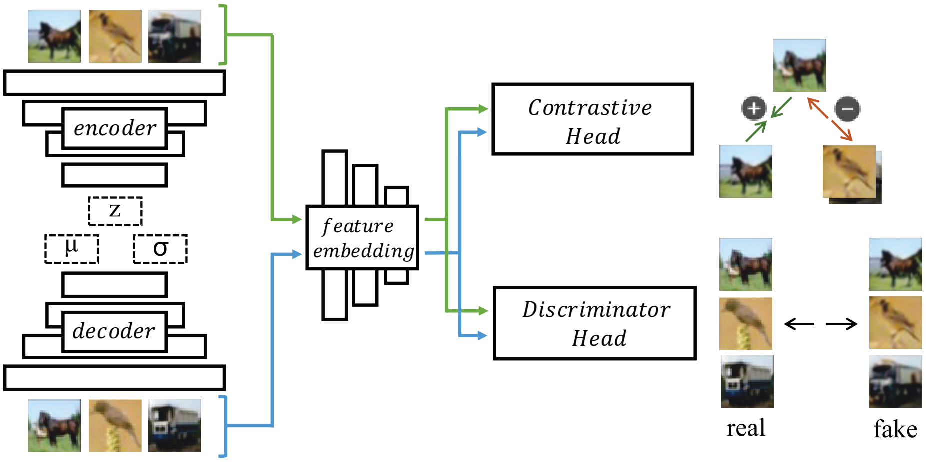


Figure 2: **Model architecture** for the proposed DC-VAE algorithm.

the following minimization and loosely call each term a loss.

$$\mathcal{L}_{\text{instance}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; i, \{\mathbf{x}_j\}_{j=1}^n) \triangleq -\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}\left[\log \frac{e^{h(\mathbf{x}_i, f_{\boldsymbol{\theta}}(\mathbf{z}))}}{\sum_{j=1}^n e^{h(\mathbf{x}_j, f_{\boldsymbol{\theta}}(\mathbf{z}))}}\right], \quad (6)$$

where i is an index for a training sample (instance), $\{\mathbf{x}_j\}_{j=1}^n$ is the union of positive samples and negative samples, $h(\mathbf{x}, \mathbf{y})$ is the critic function that measures compatibility between $\mathbf{x}$ and $\mathbf{y}$. Following the popular choice from [23], $h(\mathbf{x}, \mathbf{y})$ is the cosine similarity between the embeddings of $\mathbf{x}$ and $\mathbf{y}$: $h(\mathbf{x}, \mathbf{y}) = \frac{F_D(\mathbf{x})^\top F_D(\mathbf{y})}{||F_D(\mathbf{x})||_2||F_D(\mathbf{y})||_2}$. Note that unlike in contrastive self-supervised learning methods [63, 23, 6] where two views (independent augmentations) of an instance constitutes a positive pair, an input instance $\mathbf{x}_i$ and its reconstruction $f_{\boldsymbol{\theta}}(\mathbf{z})$ comprises a positive pair in DC-VAE. Likewise, the reconstruction $f_{\boldsymbol{\theta}}(\mathbf{z})$ and any instance that is not $\mathbf{x}_i$ can be a negative pair.

To bridge the gap between the instance-level contrastive loss (Eq. (6)) and log-likelihood in ELBO term (Eq. (1)), we observe the following connection.

Remark 1 *(From [50, 55]) The following objective is minimized, i.e., the optimal critic h is achieved, when $h(f_{\boldsymbol{\theta}}(\mathbf{z}), \mathbf{x}) = \log p(\mathbf{x}|\mathbf{z}) + c(\mathbf{x})$ where $c(\mathbf{x})$ is any function that does not depend on $\mathbf{z}$.*

$$I_{NCE} \triangleq \mathbb{E}_{\mathbf{x}_1,\cdots\mathbf{x}_K}\mathbb{E}_i[\mathcal{L}_{instance}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; i, \{\mathbf{x}_j\}_{j=1}^n)]. \quad (7)$$

It can be seen from [50, 55] that the contrastive loss of Eq. (6) *implicitly* estimates the log-likelihood $\log p_\theta(\mathbf{x}|\mathbf{z})$ required for the evidence lower bound (ELBO). Hence, we modify the ELBO objective of Eq. (1) as follows and name it as *implicit* ELBO (IELBO):

$$\mathcal{L}_{\text{IELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i) = \mathcal{L}_{\text{instance}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; i, \{\mathbf{x}_j\}_{j=1}^n) + KL[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})]. \quad (8)$$

Finally, the combined objective for the proposed DC-VAE algorithm becomes:

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \max_{D} \sum_{i=1}^n [\mathcal{L}_{\text{IELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i) + \mathcal{L}_{\text{GAN}}(\boldsymbol{\theta}, \boldsymbol{\phi}, D; \mathbf{x}_i)] \,. \quad (9)$$

The definition of $\mathcal{L}_{\text{GAN}}$ follows Eq. (4). Note here we also consider the term in Eq. (4) as contrasdistinctive since it tries to minimize the difference/discriminative classification between the input ("real") image set and the reconstructed/generated ("fake") image set. Below we highlight the significance of the two contradistinctive terms. Figure 2 shows the model architecture.

4.1. Understanding the loss terms

Instance-level fidelity. The first item in Eq. (8) is an instance-level fidelity term encouraging the reconstruction to be as close as possible to the input image while being different from all the rest of the images. A key advantage of the contrastive loss in Eq. (8) over the standard reconstruction loss in Eq. (3) is its relaxed and background instances aware formulation. In general, the reconstruction in Eq. (3) wants a perfect match between the reconstruction and the input, whereas the contrastive loss in Eq. (8) requests for being the most similar one among the training samples. This way, the contrastive loss becomes more cooperative with less conflict to the GAN loss, compared with the reconstruction loss. The introduction of the contrastive loss results in a significant improvement over VAE and VAE/GAN.

We further explain the difference between reconstruction and contrastive loss based on the input $\mathbf{x}$ and it's reconstruction $f_\theta(\mathbf{z})$. To simplify the notation, we use $\mathbf{x}$ instead of the output layer feature $F_D(\mathbf{x})$ (shown in Eq. 4)) for the illustration purpose. The reconstruction loss enforces the similarity between the reconstructed image and the input image $\min ||\mathbf{x} - f_\theta(\mathbf{z})||$ while the GAN loss computes an adversarial loss $\min \max_{\mathbf{w}} \log(\frac{1}{1+\exp\{-\mathbf{w}\cdot\mathbf{x}\}}) + \log(1 - \frac{1}{1+\exp\{-\mathbf{w}\cdot f_\theta(\mathbf{z})\}})$. $\mathbf{w}$ refers to the classifier parameter. The reconstruction loss term enforces pixel-wise/feature matching between input and the reconstruction, while the GAN loss encourages the reconstruction and input discriminatively non-separable; the two are measured in different ways resulting in a conflict. Our contrastive loss on the other hand, is also a discriminative term, it can be viewed as $\min -\log \frac{\exp\{(\mathbf{x}\cdot f_\theta(\mathbf{z}))\}}{\sum_{j=1}^{n}\exp\{(\mathbf{x}_j\cdot f_\theta(\mathbf{z}))\}}$. To compare the reconstruction loss with the contrastive loss: the former wants to have an exact match between the reconstruction with the input, whereas the later is more relaxed to be ok if no exact match but as the closest one amongst all the training samples.

In other words, the reconstruction wants a perfect match for the instance-level fidelity whereas the contrastive loss is asking for being the most similar one among the given training samples. Using the contrastive loss gives more room and creates less conflict with the GAN loss.

Set-level fidelity. The second item in Eq. (9) is a set-level fidelity term encouraging the entire set of synthesized images to be non distinguishable from the input image set. Having this term (Eq. (4)) is still important since the instance contrastive loss alone (Eq. (9)) will also lead to a degenerated situation: the input image and its reconstruction can be projected to the same point in the new feature space, but without a guarantee that the reconstruction itself lies on the valid "real" image manifold.

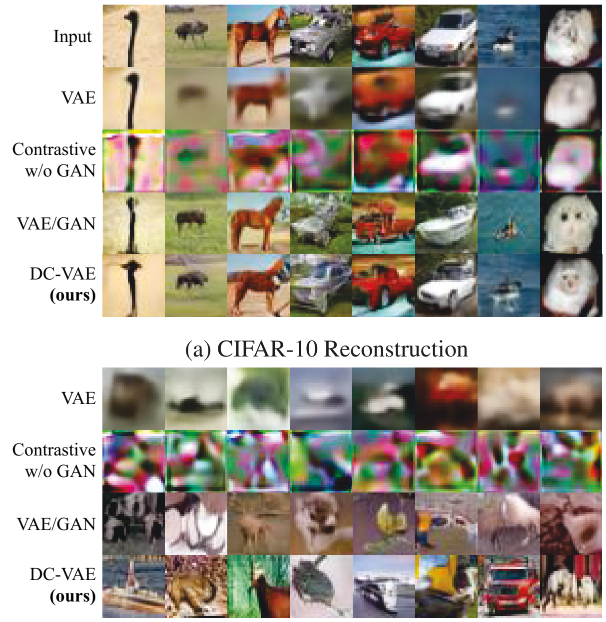
As shown in Figure 3 and Table 1 for the comparison with and without the individual terms in Eq. (9). We observe evident effectiveness of the proposed DC-VAE combining both the instance-level fidelity term (Eq. (6)) and the set-level fidelity term (Eq. (4)), compared with VAE (using pixel-wise reconstruction loss without the GAN objective), VAE-GAN (using feature reconstruction loss and the GAN objective), and VAE contrastive (using contrastive loss but without the GAN objective).

In the experiments, we show that both terms required to achieve faithful reconstruction (captured by InfoNCE loss) with perceptual realism (captured by the GAN loss).

4.2. Multi scale contrastive learning

Inspired by [46], we utilize information from feature maps at different scales. In addition to contrasting on the last layer of D in Eq. 9, we add contrastive objective on $f_l(\mathbf{z})$ where f_l is some function on top of an intermediate layer l of D. We do it in two different ways.

1. Deep supervision: We use 1×1 convolution to reduce the dimension channel-wise, and use a linear layer to obtain f_l.
2. Local patch: We use a random location across channel at layer l (size: $1\times1\times$d, where d is the channel depth).



(a) CIFAR-10 Reconstruction

(b) CIFAR-10 Samples

Figure 3: Qualitative results of CIFAR-10 [39] images (resolution 32×32) for experiments in Table 1 [39].

The intuition for the second is that in a convolutional neural network, one location at a feature map corresponds to a receptive area (patch) in the original image. Thus, by contrasting locations across channels in the same feature maps, we are encouraging the original image and the reconstruction to image have locally similar content, while encouraging them to have locally dissimilar content in other images. We use deep supervision for initial training, and add local patch after certain iterations. We observe that by using multi-scale contrastive learning, the perceptual distance for CIFAR-10 reconstruction improves from 72.2 to 52.9, while the FID score increases slightly from 15.7 to 17.9. While the decrease in the perceptual distance can be understood, we hypothesize that the increase for FID score is because multi-scale learning focuses on details, and thus creates a trade-off for the overall FID score.

5. Experiments

5.1. Implementation

Datasets To validate our method, we train our method on several different datasets — CIFAR-10 [39], STL-10 [11], CelebA [49], CelebA-HQ [35], and LSUN bedroom [68]. See the appendix for more detailed descriptions.

Network architecture For 32×32 resolution, we design the encoder and decoder subnetworks of our model in a similar way to the discriminator and generator found through neural architecture search in AutoGAN [16]. For the higher resolution experiments (128×128 and 512×512 resolution), we use Progressive GAN [35] as the backbone. Network architecture diagram is available in the appendix.

Training details The number of negative samples for contrastive learning is 8096 for all datasets. The latent dimension for the VAE decoder is 128 for CIFAR-10, STL-10, and 512 for CelebA, CelebA-HQ and LSUN Bedroom. Learning rate is 0.0002 with Adam parameters of $(\beta_1, \beta_2) = (0.0, 0.9)$ and a batch size of 128 for CIFAR-10 and STL-10. For CelebA, CelebA-HQ, LSUN Bedroom datasets, we use the optimizer parameters given in [35]. The contrastive embedding dimension used is 16 for each of the experiments.

5.2. Ablation Study

Table 1: **Ablation studies on CIFAR-10** for the proposed DC-VAE algorithm. We follow [33] and measure perceptual distance in an relu4_3 layer of a pretrained VGG network. ↓ means lower is better. ↑ means higher is better.

Method	FID↓/IS↑ Sampling	FID↓/IS↑ Reconstruction	Pixel↓ Distance	Perceptual↓ Distance
VAE	115.8 / 3.8	108.4 / 4.3	**21.8**	65.8
VAE/GAN	39.8 / 7.4	29.0 / 7.6	62.7	57.2
VAE-Contrastive	240.4 / 1.8	242 / 1.9	53.6	104.2
DC-VAE	**17.9 / 8.2**	**21.4 / 7.9**	45.9	**52.9**

To demonstrate the necessity of the GAN loss (Eq. 4) and contrastive loss (Eq. 8), we conduct four experiments with the same backbone. These experiments are: VAE (No GAN, no Contrastive), VAE/GAN (with GAN, no Contrastive), VAE-Contrastive (No GAN, with Contrastive, and ours (With GAN, with Contrastive). Here, GAN denotes Eq. 4, and Contrastive denotes Eq. 8.

Qualitative analysis From Figure 3, we see that without GAN and contrastive, images are blurry; Without GAN, the contrastive head can classify images, but not on the image manifold; Without Contrastive, reconstruction images are on the image manifold because of the discriminator, but they are different from input images. These experiments show that it is necessary to combine both instance-level and set-level fidelity, and in a contradistinctive manner.

Quantitative analysis In Table 1 we observe the same trend. VAE generates blurry images; thus the FID/IS (Inception Score) is not ideal. VAE-Contrastive does not generate images on the natural manifold; thus FID/IS is poor. VAE/GAN combines set-level and instance-level information. However the L2 objective is not ideal; thus the FID/IS is sub-optimal. For both reconstruction and sampling tasks, DC-VAE generates high fidelity images and has a favorable FID and Inception score. This illustrates the advantange of having a contradistinctive objective on both set level and instance level. To measure the faithfulness of the reconstructed image we compute the pixelwise L2 distance and the perceptual distance ([33]). For the pixel distance, VAE has the lowest reconstruction value because it directly optimizes this distance during training; our pixel-wise distance is better than VAE/GAN and VAE-Contrastive. For perceptual distance, our method outperforms the other three methods, which confirms that using contrastive learning helps the model retain more of the semantic features from the input in the reconstructed image.

Table 2: **Comparison on CIFAR-10 and STL-10.** Average Inception scores (IS) [58] and FID scores [25]. Results derived from [16]. †Result from [1]. *Result from [12].

	CIFAR-10		STL-10	
Method	**IS↑**	**FID↓**	**IS↑**	**FID↓**
Methods based on GAN:				
DCGAN [56]	6.6	-	-	-
ProbGAN [22]	7.8	24.6	8.9	46.7
WGAN-GP ResNet [18]	7.9	-	-	-
RaGAN [34]	-	23.5	-	-
SN-GAN [53]	8.2	21.7	9.1	40.1
MGAN [28]	8.3	26.7	-	-
Progressive GAN [35]	**8.8**	-	-	-
Improving MMD GAN [65]	8.3	16.2	**9.3**	37.6
PULSGAN [19]	-	22.3	-	-
AutoGAN [16]	8.6	**12.4**	9.2	**31.0**
Methods based on VAE:				
VAE	3.8	115.8	-	-
VAE/GAN	7.4	39.8	-	-
VEEGAN* [59]	-	95.2	-	-
WAE-GAN [60]	-	93.1	-	-
NCP-VAE [1]	-	24.08	-	-
NVAE† [62] Sampling	-	50.8	-	-
NVAE† [62] Reconstruction	-	**2.67**	-	-
DC-VAE Sampling (ours)	**8.2**	**17.9**	8.1	**41.9**
DC-VAE Recon. (ours)	7.9	21.4	**8.4**	43.6

5.3. Comparison to existing generative models

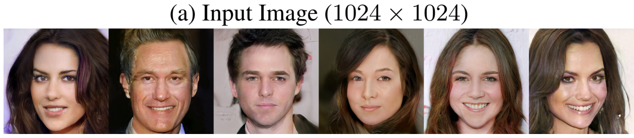
Table 2 gives a comparison of quantitative measurement for CIFAR-10 and STL-10 dataset. In general, there is a large difference in terms of FID and IS between GAN family and VAE family of models. Our model has state-of-the-art results in VAE family, and is comparable to state-of-the-art GAN models on CIFAR-10. Similarly Tables 3, 5, and 4 show that DC-VAE is able to generate images that are comparable to GAN based methods even on higher resolution datasets such as LSUN Bedrooms, CelebA, CelebA-HQ. Our method achieves state-of-the-art results on these datasets among VAE-based methods which focus on building better architectures. The qualitative comparison in Figure 5 and the reconstruction comparison in Table 8 show that our model yields more faithful reconstructions compared to existing state-of-the-art generative auto-encoder methods.

5.4. Latent Space Representation: Image and style interpolation

We further validate the effectiveness of DC-VAE for representation learning. One benefit of having an AE/VAE framework compared with just a decoder as in GAN [17] is to be able to directly obtain the latent representation from the input images. The encoder and decoder modules in VAE allows us to readily perform image/style interpolation by

Figure 4: DC-VAE Samples on LSUN Bedroom [68] (left) at resolution 128 × 128 and CelebA-HQ [35] at resolution 512 × 512 (right)

(a) Input Image (1024 × 1024)



(b) IntroVAE Reconstruction (1024 × 1024)

(c) DC-VAE Reconstruction (ours, 512 × 512)

Figure 5: Comparison of DC-VAE (resolution 512 × 512) with IntroVAE [30] (resolution 1024 × 1024). Zoom in for a better visualization.

Table 3: **Quality of Image generation (FID) comparison on LSUN Bedrooms.** †128×128 resolution. ‡256×256 resolution. ↓ means lower is better.

Method	FID↓ (Sampling)	FID↓ (Reconstruction)
Progressive GAN‡ [35]	**8.3**	-
SNGAN† [53] (from [7])	16.0	-
SSGAN†[7]	13.3	-
StyleALAE‡ [54]	17.13	15.92
DC-VAE † (ours)	14.3	**10.57**

mixing the latent variables of different images and reconstruct/synthesize new ones. We demonstrate qualitative results on image interpolation (Fig. 6), style interpolation and image editing (Fig. 7). We directly use the trained DC-VAE model without disentanglement learning [36]. We also quantitatively compare the latent space disentanglement through the perceptual path length (PPL) [36] (Table 7). We observe that DC-VAE learns a more disentangled latent space representation than the backbone Progressive GAN [35] and StyleALAE [54] that use a much more capable StyleGAN [36] backbone.

Table 4: **FID comparison on CelebA-HQ for 256x256 resolution.** ↓ means lower is better.

Method	FID↓
StyleALAE [54]	19.21
NVAE [62] (from [1])	40.26
NCP-VAE [1]	24.79
DC-VAE (ours)	**15.81**

Table 5: **FID comparison on CelebA.** *64×64 resolution. †128×128 resolution. ↓ means lower is better.

Method	FID↓
Methods based on GAN:	
PresGAN* [12]	29.1
LSGAN [52] (from [29])	53.9
COCO-GAN † [48]	**5.7**
ProGAN† [35] (from [48])	7.30
Methods based on VAE:	
VEE-GAN† [59] (from [12])	46.2
WAE-GAN* [60]	42
NCP-VAE* [1]	**5.3**
DC-VAE† (ours) Reconstruction	**14.3**
DC-VAE† (ours) Sampling	**19.9**

5.5. Latent Space Representation: Classification

To show that our model learns a good representation, we measure the performance on the downstream MNIST classification task [13]. The VAE models were trained on MNIST dataset [45]. We feed input images into our VAE encoder and get the latent representation. Then we train a linear classifier on the latent representation to classify the classes of the input images. Results in Table 6 show that our model gives the lowest classification error in most cases. This experiment demonstrates that our model not only gains

Figure 6: Interpolation results generated by DC-VAE (ours) on CelebA-HQ [35] images (512 × 512, left) and LSUN Bedroom [68] images (128 × 128, right). (Zoom in for a better visualization.)

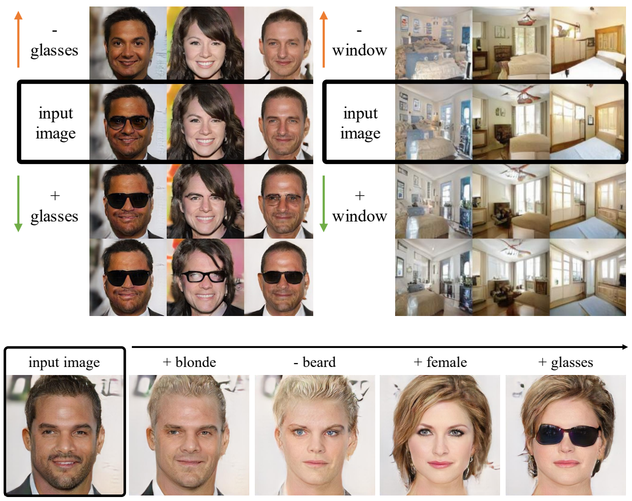


Figure 7: Latent traversal on CelebA-HQ [35] (resolution 512 × 512) and LSUN Bedroom [68] (resolution 128 × 128) and example image editing on CelebA-HQ [35] images.

Table 6: **Comparison to prior VAE-based representation learning methods.** Classification error on MNIST dataset. ↓: lower is better. 95 % confidence intervals are from 5 trials. Results derived from [13].

Method	$d_{\mathbf{z}} = 16$ ↓	$d_{\mathbf{z}} = 32$ ↓	$d_{\mathbf{z}} = 64$ ↓
VAE [38]	2.92%±0.12	3.05%±0.42	2.98%±0.14
β-VAE(β=2) [26]	4.69%±0.18	5.26%±0.22	5.40%±0.33
FactorVAE(γ=5) [37]	6.07%±0.05	6.18%±0.20	6.35%±0.48
β-TCVAE (α=1,β=5,γ=1) [8]	1.62%±0.07	1.24%±0.05	1.32%±0.09
Guided-VAE [13]	1.85%±0.08	1.60%±0.08	1.49%±0.06
Guided-β-TCVAE [13]	1.47%±0.12	**1.10%±0.03**	1.31%±0.06
DC-VAE (Ours)	**1.30%±0.035**	1.27%±0.037	**1.29%±0.034**

the ability to do faithful synthesis and reconstruction, but also gains better representation ability on the VAE side.

Table 7: PPL Comparison of on CelebA-HQ [35].

Method	Backbone	PPL Full↓
StyleALAE [54]	StyleGAN [36]	33.29
ProGAN [35]	ProGAN [35]	40.71
DC-VAE (ours)	ProGAN [35]	**24.66**

Table 8: **Reconstruction Comparison of on CelebA-HQ** [35] validation set. We follow [33] and measure perceptual distance in an relu4_3 layer of a pretrained VGG network. ↓ means lower is better.

Method	Backbone	Pixel↓ Distance	Perceptual↓ Distance
StyleALAE [54]	StyleGAN [36]	0.117	40.40
DC-VAE (ours)	ProGAN [35]	**0.072**	**38.63**

6. Conclusion

In this paper, we have developed dual contradistinctive generative autoencoder (DC-VAE), a new framework that integrates an instance-level discriminative loss (InfoNCE) with a set-level adversarial loss (GAN) into a single variational autoencoder framework. Our experiments show state-of-the-art or competitive results in several tasks, including image synthesis, image reconstruction, representation learning for image interpolation, and representation learning for classification. DC-VAE is a general-purpose VAE model and it points to a encouraging direction that attains high-quality synthesis (decoding) and inference (encoding).

Acknowledgment. This work is funded by NSF IIS-1717431 and NSF IIS-1618477. We thank Qualcomm Inc. for an award support. The work was performed when G. Parmar and D. Li were with UC San Diego.

References

[1] Jyoti Aneja, Alexander Schwing, Jan Kautz, and Arash Vahdat. Ncp-vae: Variational autoencoders with noise contrastive priors, 2020.

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

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Neural Information Processing Systems*, 1993.

[5] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, 11:1109–1135, Mar. 2010.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020.

[7] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *CVPR*, pages 12154–12163, 2019.

[8] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

[9] 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.

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, page 539–546, 2005.

[11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011.

[12] Adji B Dieng, Francisco JR Ruiz, David M Blei, and Michalis K Titsias. Prescribed generative adversarial networks. *arXiv:1910.04302*, 2019.

[13] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *CVPR*, 2020.

[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[15] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017.

[16] Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *ICCV*, 2019.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[18] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 2017.

[19] T. Guo, C. Xu, J. Huang, Y. Wang, B. Shi, C. Xu, and D. Tao. On positive-unlabeled classification in gan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8382–8390, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[20] Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(null):307–361, Feb. 2012.

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

[22] Hao He, Hao Wang, Guang-He Lee, and Yonglong Tian. Probgan: Towards probabilistic gan with theoretical guarantees. In *ICLR*, 2019.

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

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[25] 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 *Advances in Neural Information Processing Systems*, 2017.

[26] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[27] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, 1994.

[28] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Mgan: training generative adversarial nets with multiple generators. In *ICLR*, 2018.

[29] Yedid Hoshen, Ke Li, and Jitendra Malik. Non-adversarial image synthesis with generative latent nearest neighbors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5811–5819, 2019.

[30] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in Neural Information Processing Systems*, pages 52–63, 2018.

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

[32] Long Jin, Justin Lazarow, and Zhuowen Tu. Introspective classification with convolutional nets. In *Advances in Neural Information Processing Systems*, 2017.

[33] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[34] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

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

[36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[37] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018.

[38] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[41] Brian Kulis et al. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364, 2012.

[42] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

[43] Justin Lazarow, Long Jin, and Zhuowen Tu. Introspective neural networks for generative modeling. In *ICCV*, 2017.

[44] Yann LeCun. *Modeles connexionnistes de lapprentissage*. PhD thesis, PhD thesis, These de Doctorat, Universite Paris 6, 1987.

[45] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[46] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.

[47] Kwonjoon Lee, Weijian Xu, Fan Fan, and Zhuowen Tu. Wasserstein introspective neural networks. In *CVPR*, 2018.

[48] 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*, pages 4512–4521, 2019.

[49] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[50] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *EMNLP*, 2018.

[51] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[52] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

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

[54] Stanislav Pidhorskyi, Donald Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders, 2020.

[55] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

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

[57] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In David van Dyk and Max Welling, editors, *AISTATS*, volume 5, pages 448–455, 16–18 Apr 2009.

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

[59] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017.

[60] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[61] Zhuowen Tu. Learning generative models via discriminative approaches. In *CVPR*, 2007.

[62] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020.

[63] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[65] Wei Wang, Yuan Sun, and Saman Halgamuge. Improving mmd-gan training with repulsive loss function. In *ICLR*, 2019.

[66] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[67] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *ICML*, 2016.

[68] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[69] Zijun Zhang, Ruixiang Zhang, Zongpeng Li, Yoshua Bengio, and Liam Paull. Perceptual generative autoencoders. In *ICLR*, 2019.