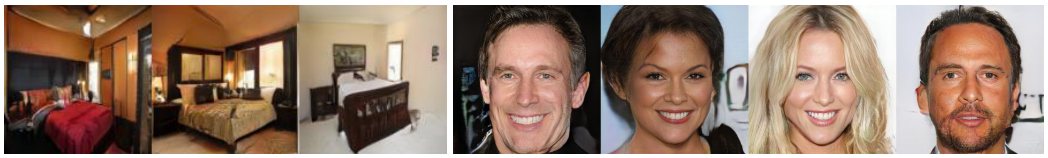
DUAL CONTRADISTINCTIVE GENERATIVE AUTOENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a new generative autoencoder model with dual contradistinctive losses to improve generative autoencoder that performs simultaneous inference (reconstruction) and synthesis (generation). We name our model dual contradistinctive generative autoencoder (DC-VAE) that 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. There also exists a mathematical connection between the instance-based classification and instance-level conditional distribution. DC-VAE achieves competitive results in three tasks, including image synthesis, image reconstruction, and representation learning. DC-VAE is applicable to various tasks in computer vision and machine learning.

(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 Yu et al. (2015) at resolution 128×128 (left) and CelebA-HQ (Karras et al., 2018) at resolution 512×512 (right).

1 INTRODUCTION

Tremendous progress has been made in deep learning for the development of various learning frameworks (Krizhevsky et al., 2012; He et al., 2016; Goodfellow et al., 2014; Vaswani et al., 2017). Autoencoder (AE) (LeCun, 1987; Hinton & Zemel, 1994) 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) (Kingma & Welling, 2014), additionally learns a variational model for the latent variables to capture the underlying sample distribution. For the encoder and decoder models separately, tremendous progress has been made in image classification with deep convolutional neural network (CNN) (Krizhevsky et al., 2012; He et al., 2016) (an encoder) and in image generation with generative adversarial network (GAN) (Goodfellow et al., 2014) (a decoder).

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 (Larsen et al., 2016) combines a reconstruction loss with an adversarial loss. However, the result of VAE/GAN is sub-optimal, as shown in Table 1.

The pixel-wise reconstruction loss in the standard VAE (Kingma & Welling, 2014) 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 (Kulis et al., 2012) and contrastive learning (Hadsell et al., 2006; Wu et al., 2018; van den Oord et al., 2018). 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 three tasks altogether, including image synthesis, image reconstruction, and representation learning.

2 Related work

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

Variational autoencoder (VAE) (Kingma & Welling, 2014) points to an exciting direction of generative models by developing an Evidence Lower BOund (ELBO) objective (Higgins et al., 2017; Ding et al., 2020). However, the VAE reconstruction/synthesis is known to be blurry. To improve the image quality, a sequence of VAE based models have been developed (Larsen et al., 2016; Dumoulin et al., 2017; Huang et al., 2018; Brock et al., 2018; Zhang et al., 2019). VAE/GAN (Larsen et al., 2016) 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 Huang et al. (2018) 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 (Zhang et al., 2019) adds a constraint to the latent variables.

Pioneering works of (Tu, 2007; Gutmann & Hyvärinen, 2012) 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) (Goodfellow et al., 2014) builds on neural networks and amortized sampling (a decoder network that maps a noise into an image). The subsequent development after GAN (Radford et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Karras et al., 2018; Gong et al., 2019; Dumoulin et al., 2017; Donahue et al., 2017) 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. Energy-based generative models (Salakhutdinov & Hinton, 2009; Xie et al., 2016; Jin et al., 2017; Lee et al., 2018) — which aim to directly model data density — are making a steady improvement for a simultaneously generative and discriminative single model.

From another angle, contrastive learning (Hadsell et al., 2006; Wu et al., 2018; He et al., 2020; Chen et al., 2020) 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 (Bromley et al., 1993; Chopra et al., 2005; Chechik et al., 2010).

In this paper, we aim to improve VAE (Kingma & Welling, 2014) by introducing a contrastive loss (van den Oord et al., 2018) to address instance-level fidelity between the input and the reconstruction in the induced feature space. Unlike in self-supervised representation learning methods (van den Oord et al., 2018; He et al., 2020; Chen et al., 2020), 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, 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 4.1: 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 dual contradistinctive generative autoencoder (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) (Kingma & Welling, 2014) 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 (Kingma & Welling, 2014), 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.

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

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 (Johnson et al., 2016; Isola et al., 2017), 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) (Goodfellow et al., 2014) and its variants (Radford et al., 2016), 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 (Tu, 2007; Gutmann & Hyvärinen, 2012; Goodfellow et al., 2014).

VAE/GAN (Larsen et al., 2016) 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:

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

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 (Johnson et al., 2016). 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:

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

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 (Malisiewicz et al., 2011), approximating likelihood by classification (Tu, 2007), and contrastive learning (Hadsell et al., 2006; Wu et al., 2018; He et al., 2020), we propose to model instance-level fidelity by contrastive loss (commonly referred to as InfoNCE loss) (van den Oord et al., 2018). In DC-VAE, we perform the following minimization and loosely call each term a loss.

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

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 (He et al., 2020), $h(\mathbf{x}, \mathbf{y})$ 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 (van den Oord et al., 2018; He et al., 2020; Chen et al., 2020) 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 show the following observation.

Theorem 1 *(From (Ma & Collins, 2018; Poole et al., 2019)) 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 $\mathbf{z}$.*

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

From Theorem 1, we see 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):

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

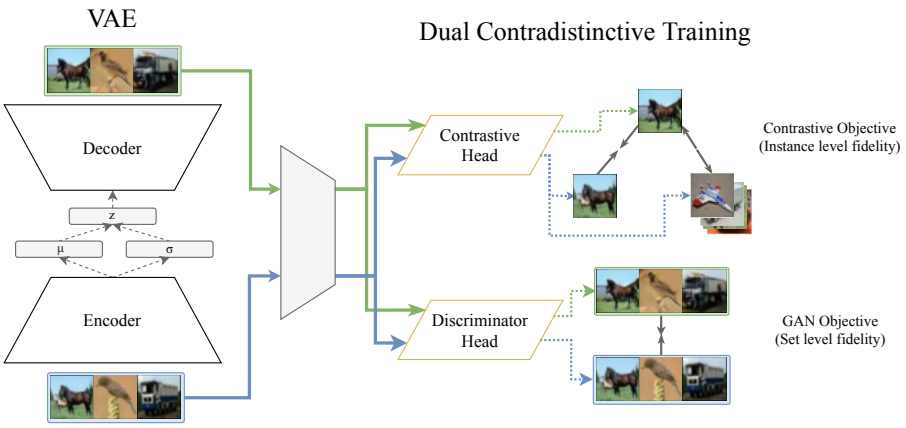


Figure 2: **Model architecture** for the proposed DC-VAE algorithm.

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

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

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.

- **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 in which only matching the reconstruction, and the input instance is enforced.
- **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).

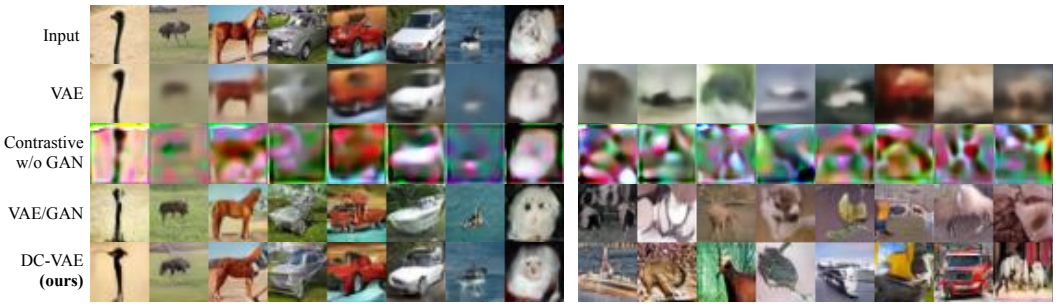
4.1 MULTI SCALE CONTRASTIVE LEARNING

Inspired by (Lee et al., 2015), we utilize information from feature maps at different scales. In addition to contrasting on the last layer of D in Equation 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).

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.

5 EXPERIMENTS



(a) CIFAR-10 Reconstruction (b) CIFAR-10 Samples

Figure 3: Qualitative results of CIFAR-10 (Krizhevsky et al., 2009) images (resolution 32×32) for experiments in Table 1 (Krizhevsky et al., 2009).

5.1 IMPLEMENTATION

Datasets To validate our method, we train our method on several different datasets — CIFAR-10 (Krizhevsky et al., 2009), STL-10 (Coates et al., 2011), CelebA (Liu et al., 2015), CelebA-HQ (Karras et al., 2018), and LSUN bedroom (Yu et al., 2015). 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 (Gong et al., 2019). For the higher resolution experiments (128×128 and 512×512 resolution), we use Progressive GAN (Karras et al., 2018) 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 (Karras et al., 2018). 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 (Johnson et al., 2016) 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 (Johnson et al. (2016)). For the pixel distance, VAE has the lowest 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 other three, which confirms that using contrastive learning helps reconstruct images semantically.

Table 2: **Comparison on CIFAR-10 and STL-10.** Average Inception scores (Salimans et al., 2016) and FID scores (Heusel et al., 2017). Results derived from (Gong et al., 2019). Table style derived from (Lee et al., 2019). †Result from Aneja et al. (2020). *Result from (Dieng et al., 2019).

	CIFAR-10		STL-10	
Method	**Inception Score↑**	**FID↓**	**Inception Score↑**	**FID↓**
Methods based on GAN:				
DCGAN (Radford et al., 2016)	6.6	-	-	-
ProbGAN (He et al., 2019)	7.8	24.6	8.9	46.7
WGAN-GP ResNet (Gulrajani et al., 2017)	7.9	-	-	-
RaGAN (Jolicoeur-Martineau, 2018)	-	23.5	-	-
SN-GAN (Miyato et al., 2018)	8.2	21.7	9.1	40.1
MGAN (Hoang et al., 2018)	8.3	26.7	-	-
Progressive GAN (Karras et al., 2018)	**8.8**	-	-	-
Improving MMD GAN (Wang et al., 2019)	8.3	16.2	**9.3**	37.6
PULSGAN (Guo et al., 2020)	-	22.3	-	-
AutoGAN (Gong et al., 2019)	8.6	**12.4**	9.2	**31.0**
Methods based on VAE:				
VAE	3.8	115.8	-	-
VAE/GAN	7.4	39.8	-	-
VEEGAN* (Srivastava et al., 2017)	-	95.2	-	-
WAE-GAN (Tolstikhin et al., 2017)	-	93.1	-	-
NVAE† (Vahdat & Kautz, 2020) Sampling	-	50.8	-	-
NVAE† (Vahdat & Kautz, 2020) Reconstruction	-	**2.67**	-	-
DC-VAE Sampling (ours)	**8.2**	**17.9**	8.1	**41.9**
DC-VAE Reconstruction (ours)	7.9	21.4	**8.4**	43.6

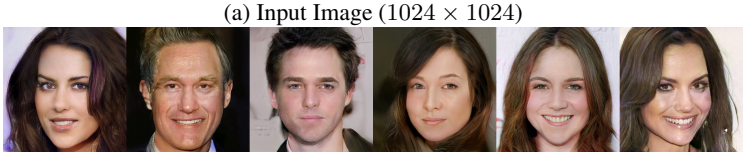
5.3 Comparison to existing generative models

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 Table 4 shows that DC-VAE is able to generate images that are comparable to GAN based methods even on higher resolution datasets.

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 Goodfellow et al. (2014) 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 mixing the latent variables of different images and reconstruct/synthesize new ones. We demonstrate qualitative results on image interpolation (Fig. 5, Appendix Fig. 9), style interpolation (Appendix Fig. 10) and image editing (Fig. 6). We directly use the trained DC-VAE model without disentanglement learning Karras et al. (2019). Additional latent space analysis and the method used for interpolation and editing is

(a) Input Image (1024×1024)



(b) IntroVAE Reconstruction (1024×1024)

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

Figure 4: Comparison of DC-VAE (resolution 512×512) with IntroVAE (Huang et al., 2018) (resolution 1024×1024).

Table 3: Quality of image generation (FID) comparison on LSUN Bedrooms. $^{\dagger}128\times128$ resolution. $^{\P}256\times256$ resolution. $\downarrow$ means lower is better.

Method	FID↓
Progressive GAN‡ (Karras et al., 2018)	**8.3**
SNGAN† (Miyato et al., 2018) (from (Chen et al., 2019))	16.0
SSGAN†(Chen et al., 2019)	13.3
StyleALAE¶ (Pidhorskyi et al., 2020) Reconstruction	15.92
StyleALAE¶ (Pidhorskyi et al., 2020) Sampling	17.13
DC-VAE† (ours) Reconstruction	**10.57**
DC-VAE† (ours) Sampling	14.3

Table 4: FID on CelebA. $^{*}64\times64$ resolution.

Method	FID↓
Methods based on GAN:	
PresGAN* (Dieng et al., 2019)	29.1
LSGAN (Mao et al., 2017) (from (Hoshen et al., 2019))	53.9
COCO-GAN † (Lin et al., 2019)	**5.7**
PGAN† (Karras et al., 2018) (from (Lin et al., 2019))	7.30
Methods based on VAE:	
VEE-GAN† (Srivastava et al., 2017) (from (Dieng et al., 2019))	46.2
WAE-GAN* (Tolstikhin et al., 2017)	42
DC-VAE† (ours) Reconstruction	**14.3**
DC-VAE† (ours) Sampling	**19.9**

provided in the Appendix. We also quantitatively compare the latent space disentanglement through the perceptual path length (PPL) (Karras et al., 2019) (Table 6). We observe that DC-VAE learns a more disentangled latent space representation than the backbone Progressive GAN (Karras et al., 2018) and StyleALAE (Pidhorskyi et al., 2020) that uses a much more capable StyleGAN (Karras et al., 2019) backbone.

Figure 5: Interpolation results generated by DC-VAE (ours) on CelebA-HQ (Karras et al., 2018) images (512×512, left) and LSUN Bedroom Yu et al. (2015) images (128×128, right). More images can be seen in Appendix Figure 9. (Zoom in for a better visualization.)

5.5 Latent Space Representation: Clssification

To show that our model learns a good representation, we measure the performance on the downstream MNIST classification task (Ding et al., 2020). The VAE models were trained on MNIST dataset (LeCun, 1998). 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 5 show that our model gives the lowest classification error in most cases. This experiment demonstrates that our model not only gains the ability to do faithful synthesis and reconstruction, but also gains better representation ability on the VAE side.

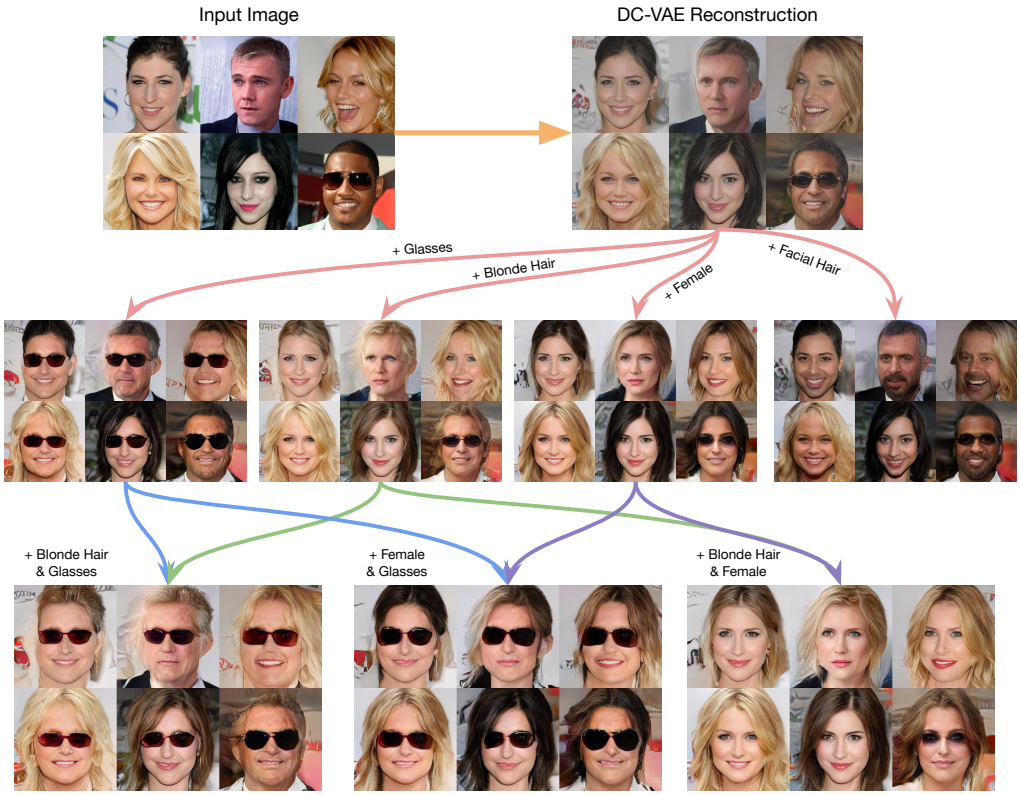


Figure 6: Image editing on CelebA-HQ (Karras et al., 2018) validation set images (resolution 512×512). The method used to generate these is outlined in Appendix A.2.

Table 5: **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 (Ding et al., 2020).

Method	$d_{\mathbf{z}} = 16 \downarrow$	$d_{\mathbf{z}} = 32 \downarrow$	$d_{\mathbf{z}} = 64 \downarrow$
VAE (Kingma & Welling, 2014)	2.92%±0.12	3.05%±0.42	2.98%±0.14
β-VAE(β=2) (Higgins et al., 2017)	4.69%±0.18	5.26%±0.22	5.40%±0.33
FactorVAE(γ=5) (Kim & Mnih, 2018)	6.07%±0.05	6.18%±0.20	6.35%±0.48
β-TCVAE (α=1,β=5,γ=1) (Chen et al., 2018)	1.62%±0.07	1.24%±0.05	1.32%±0.09
Guided-VAE (Ding et al., 2020)	1.85%±0.08	1.60%±0.08	1.49%±0.06
Guided-β-TCVAE (Ding et al., 2020)	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**

Method	**Backbone**	**PPL Full↓**
StyleALAE (Pidhorskyi et al., 2020)	StyleGAN	33.29
Progressive GAN (Karras et al., 2018)	Progressive GAN	40.71
DC-VAE (ours)	Progressive GAN	**24.66**

Table 6: PPL Comparison of on CelebA-HQ Karras et al. (2018)

6 Conclusion

In this paper, we have proposed 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 competitive results for a single model in several tasks, including image synthesis, image reconstruction, representation learning for image interpolation, and representation learning for classification. DC-VAE points to a encouraging direction that attains high-quality synthesis (decoding) and inference (encoding).

REFERENCES

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

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

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

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.

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, March 2010. ISSN 1532-4435.

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.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12154–12163, 2019.

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

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pp. 539–546, 2005. ISBN 0769523722.

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

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

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

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

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

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

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.

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.

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)*, pp. 8382–8390, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00841. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00841.

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, February 2012. ISSN 1532-4435.

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

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

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

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

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.

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.

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

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

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*, pp. 5811–5819, 2019.

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*, pp. 52–63, 2018.

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

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

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

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

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

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*, pp. 4401–4410, 2019.

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

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

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

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

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

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

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

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

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

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

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4512–4521, 2019.

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

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

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

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*, pp. 2794–2802, 2017.

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

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

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

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

Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In David van Dyk and Max Welling (eds.), *AISTATS*, volume 5, pp. 448–455, 16–18 Apr 2009.

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.

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*, pp. 3308–3318, 2017.

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

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

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL `http://arxiv.org/abs/1807.03748`.

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.

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

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

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

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.

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

A APPENDIX

A.1 Additional reconstruction results

Figure 7: Additional CelebA-HQ (Karras et al., 2018) reconstruction images (resolution 512×512) generated by DC-VAE (ours)

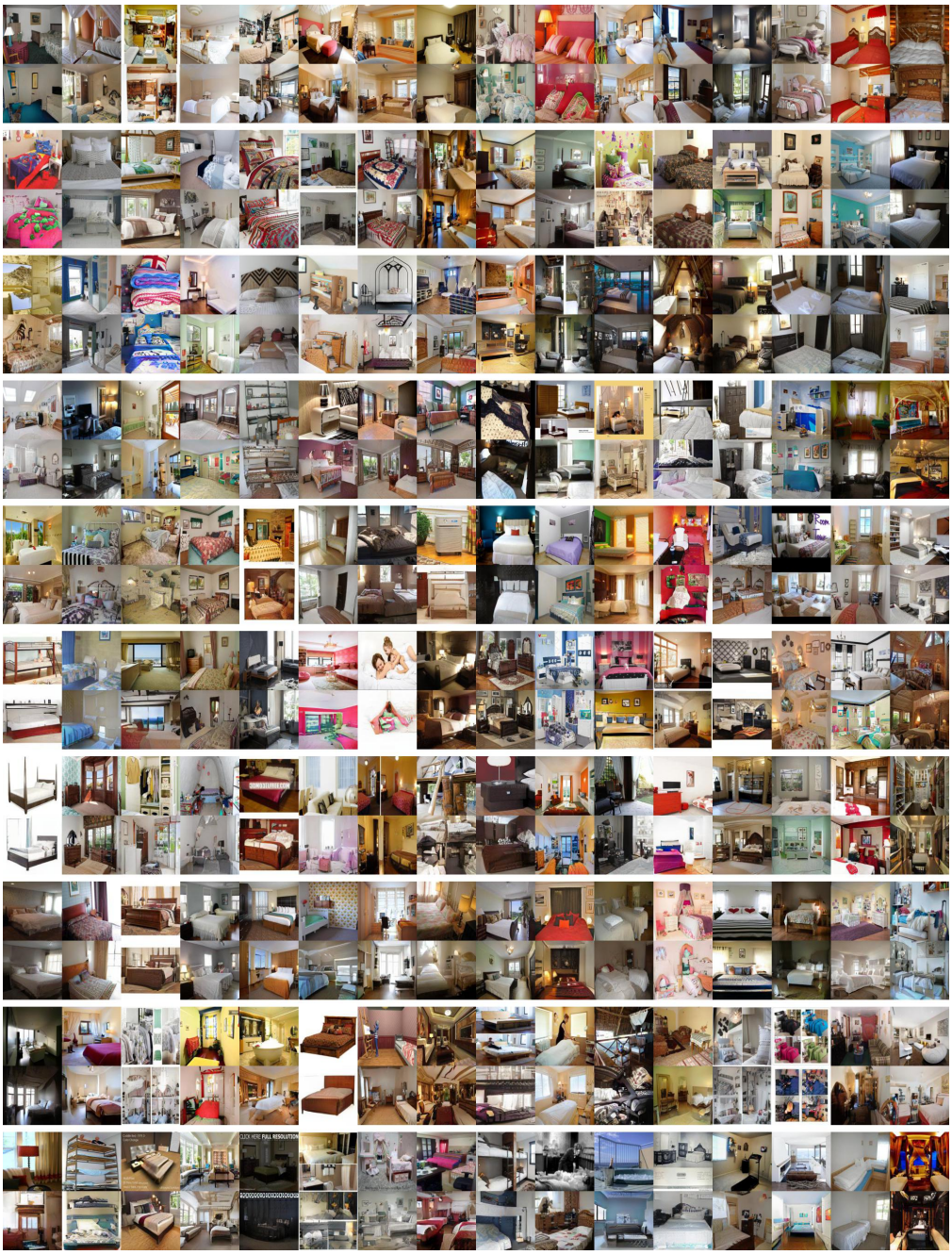


Figure 8: Additional LSUN Bedroom (Yu et al., 2015) reconstruction images (resolution 128×128)

A.2 **Analysing the latent space**

In this section we analyse the smoothness of the latent space learnt by DC-VAE. In Figure 9 we qualitatively show the high resolution (512×512) CelebA-HQ Karras et al. (2018) images generated by an evenly spaced linear blending between two latent vectors. In Fig. 6 we show that DC-VAE is able to perform meaningful attribute editing on images while retaining the original identity. To perform image editing, we first need to compute the direction vector in the latent space that correspond to a desired attribute (e.g. has glasses, has blonde hair, is a woman, has facial hair). We compute these attribute direction vectors by selecting 20 images that have the attribute and 20 images that do not have the attribute, obtaining the corresponding pairs of 20 latent vectors, and calculating the

difference of the mean. The results in Fig. 6 show that these direction vectors can be added to a latent vector to add a diverse combination of desired image attributes while retaining the original identity of the individual.

Additionally we corroborate the above qualitative results quantitatively by inspecting the Perceptual Path Length (PPL) Karras et al. (2019) of our learn DC-VAE Decoder (Tab. 6) to measures the disentanglement of the latent space. We note that although ProgressiveGAN (ours base model) has a better FID score, DC-VAE has a lower PPL score which indicated that the latent space learnt is more disentangled.

Figure 9: Additional latent space interpolations on CelebA-HQ (Karras et al., 2018) (resolution 512×512)

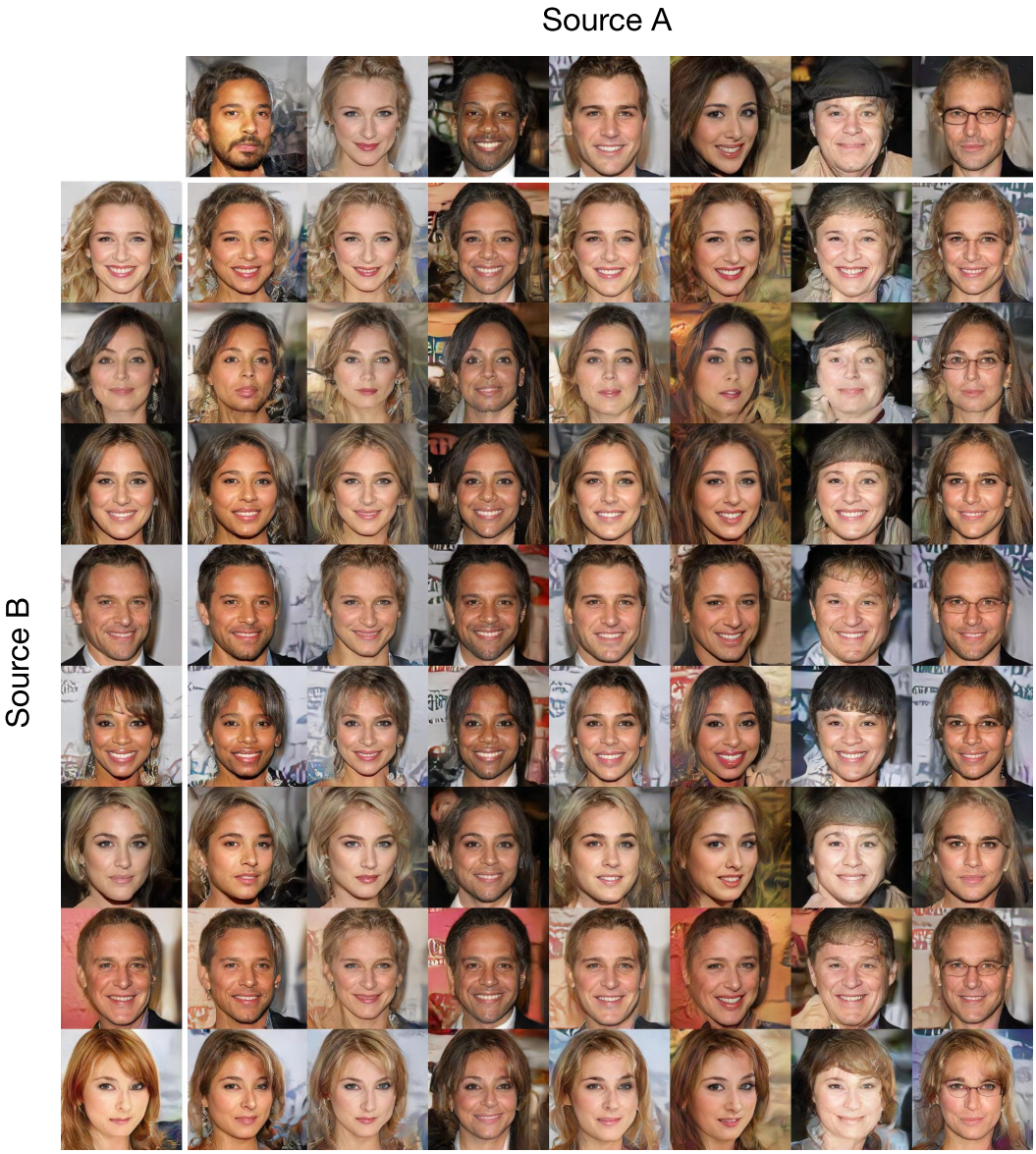


Figure 10: Latent Mixing results on CelebA-HQ Karras et al. (2018). Each combined image in the grid is generated by replacing an arbitrary subset of Source A latent with the corresponding Source B latent.

A.3 **Effect of negative samples**

In this section we analyse the effect of varying the number of negative samples used for contrastive learning. The figure 11 shows the reconstruction error on the CIFAR-10 Krizhevsky et al. (2009) test set as the negative samples is varied. We observe that a higher number of negative samples results in better reconstruction. We choose 8096 for all of our experiments because of memory constraints.

A.4 **Datasets used**

CIFAR-10 comprises 50,000 training images and 10,000 test images with a spatial resolution of 32×32. STL-10 is a similar dataset that contains 5,000 training images and 100,000 unlabeled

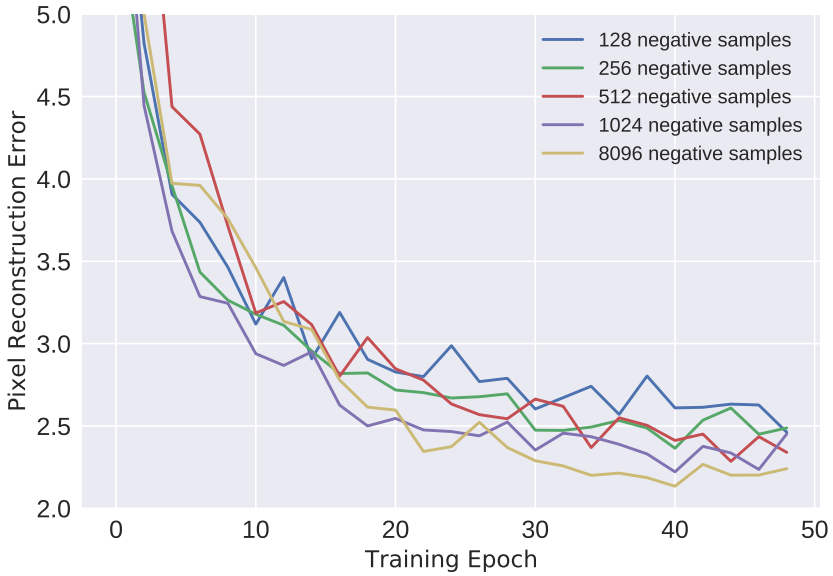


Figure 11: Pixel reconstruction error on CIFAR-10 Krizhevsky et al. (2009) test set for varying number of negative samples

images at 96×96 resolution. We follow the procedure in AutoGAN(Gong et al., 2019) and resize the STL-10 images to 32×32. The CelebA dataset has 162,770 training images and 19,962 testing images, CelebA-HQ contains 29,000 training images with 1,000 test images of size 1024×1024, and LSUN Bedroom has approximately 3M images. We resize all images progressively in these three datasets from (4×4) to (512×512) for the progressive training.

A.5 Network architecture diagrams

In Figures 15 we show the detailed network architecture of DC-VAE for input resolutions of 32×32. Note that the comparison results shown in Figure 3 and Table 1 in the main paper, for VAE, VAE/GAN, VAE w/o GAN, and our proposed DC-VAE are all based on the same network architecture (shown in Figure 15 here), for a fair comparison.

The network architectures shown in Figure 15 are adapted closely from the networks discovered by (Gong et al., 2019) through Neural Architecture Search. The DC-VAE developed in our paper is not tied to any particular CNN architecture. We choose the AutoGAN architecture (Gong et al., 2019) to start with a strong baseline. The decoder in Figure 15 matches the generator in (Gong et al., 2019). The encoder is built by modifying the output shape of the final linear layer in the discriminator of AutoGAN (Gong et al., 2019) to match the latent dimension and adding spectral normalization. The discriminator is used both for classifying real/fake images, and contrastive learning. For each layer we choose, we first apply 1x1 convolution and a linear layer, and then use this feature as an input to the contrastive module. For experiments at 32×32, we pick two different positions: the output of second residual conv block (lower level) and the output of the first linear layer (higher level). For experiments on higher resolution datasets we use a Progressive GAN (Karras et al., 2018) Generator and Discriminator as our backbone and apply similar modifications as described above.

A.6 Training infrastructure

A.7 Further details about the representation learning experiments

As seen in Table 4 in the main paper, we show the representation capability of DC-VAE following the procedure outlined in (Ding et al., 2020). We train our model on the MNIST dataset (LeCun, 1998) and measure the transferability though a classification task on the latent embedding vector. Specifically, we first pretrain the DC-VAE model on the training split of the MNIST dataset. Following

that we freeze the DC-VAE model and train a linear classifier that takes latent embedding vector as the input and predicts the class label of the original image.

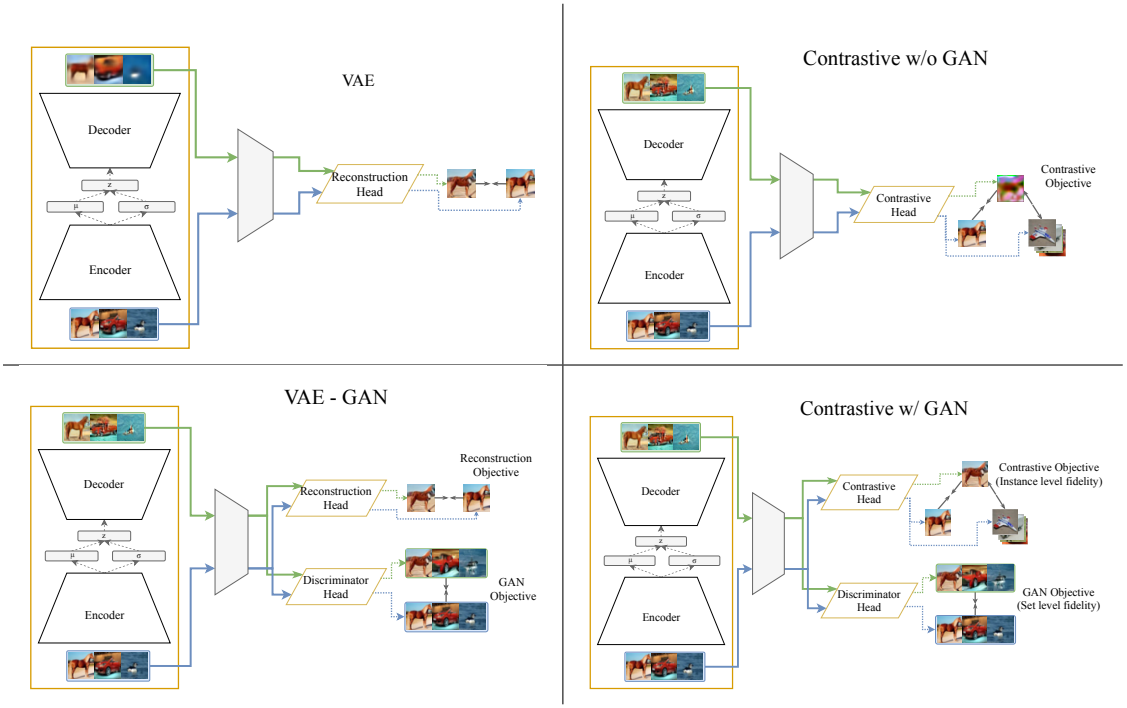
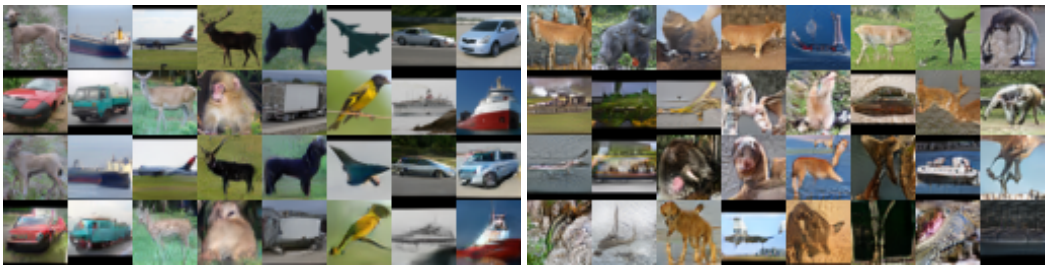


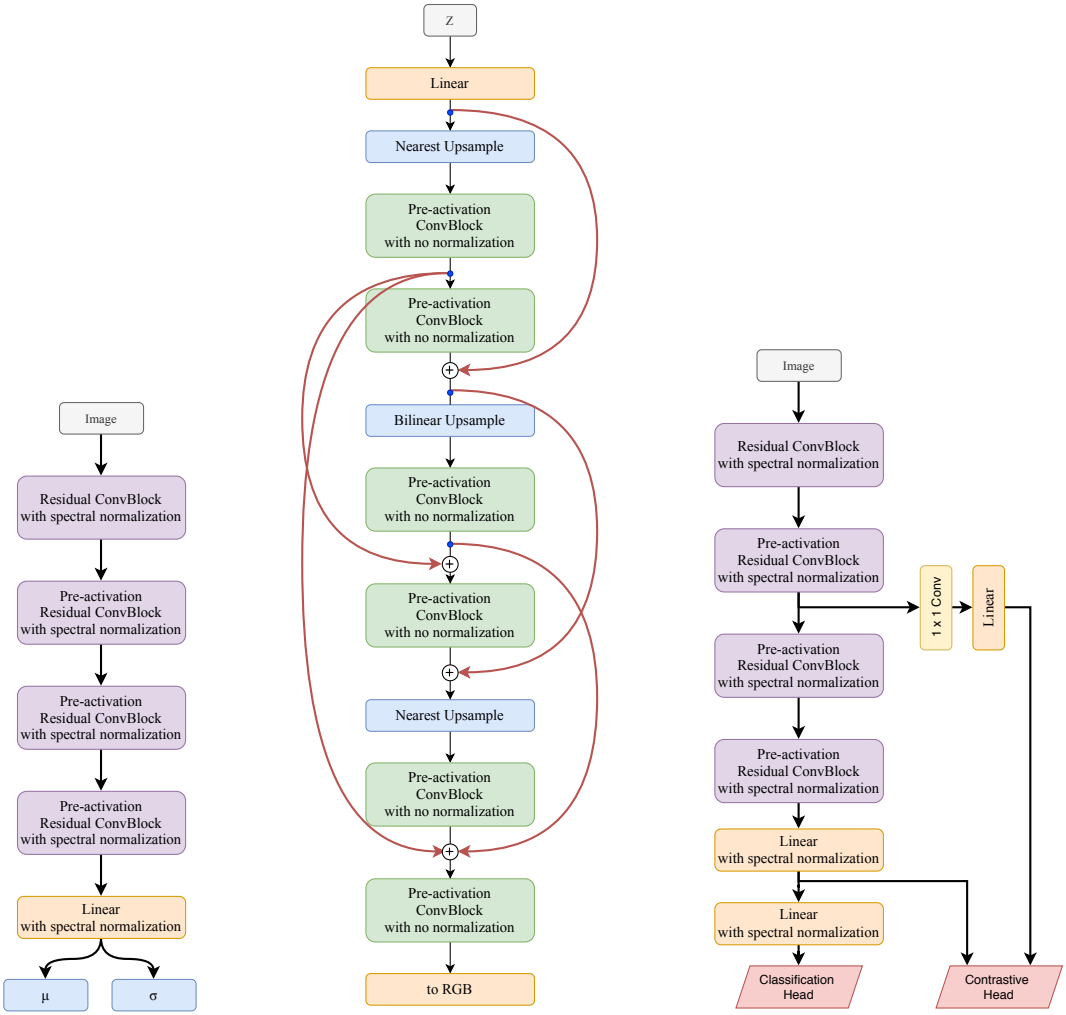
Figure 12: Visualization of the effect of adding each instance level and set level objectives. Table 1 and Figure 3 contain FID (Heusel et al., 2017) results and qualitative comparisons on the CIFAR-10 (Krizhevsky et al., 2009) that correspond to these settings.

Figure 13: DC-VAE synthesis images on LSUN images (Yu et al., 2015) (resolution 128×128)



(a) STL10 Reconstructions generated by DC-VAE (b) STL10 Samples generated by DC-VAE

Figure 14: DC-VAE reconstruction (a) and synthesis results (b) on STL10 (Coates et al., 2011) images (resolution 32×32). In (a) the top two rows are input images and the bottom two rows are the corresponding reconstruction images.



(a) Encoder (b) Decoder (c) Discriminator

Figure 15: Network architecture of DC-VAE for resolution 32×32 for CIFAR-10 (Krizhevsky et al., 2009) and STL-10 (Coates et al., 2011). (a) is the Encoder. (b) is the Decoder. (c) is the Discriminator.