Variational Conditional GAN for Fine-grained Controllable Image Generation

Mingqi Hu MINGQ@SEU.EDU.CN
Southeast University, Nanjing 211189, China

Deyu Zhou D.ZHOU@SEU.EDU.CN
Southeast University, Nanjing 211189, China

Yulan He YULAN.HE@WARWICK.AC.UK
University of Warwick, Coventry CV4 7AL, UK

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

In this paper, we propose a novel variational generator framework for conditional GANs to catch semantic details for improving the generation quality and diversity. Traditional generators in conditional GANs simply concatenate the conditional vector with the noise as the input representation, which is directly employed for upsampling operations. However, the hidden condition information is not fully exploited, especially when the input is a class label. Therefore, we introduce a variational inference into the generator to infer the posterior of latent variable only from the conditional input, which helps achieve a variable augmented representation for image generation. Qualitative and quantitative experimental results show that the proposed method outperforms the state-of-the-art approaches and achieves the realistic controllable images.

Keywords: Controllable Image Generation, Variational Inference, Generative Adversarial Networks (GANs)

1. Introduction

Generating controllable images from natural language has many applications, including text illustration, computer-aided design and second-language learning. In recent years, it has gained increasing interests in the research community (Mansimov et al., 2016). However, due to the challenges faced in language understanding and cross-modal transformation, it is far from being solved (Reed et al., 2016b).

Since directly generating images from text descriptions is difficult, a relatively easier problem is to generate an image from a class label which indicates an image category such as '*Plane*', '*Cat*' and '*Coat*', which is called *class-conditional image generation*. Recently, with the emergence of Conditional Generative Adversarial Network (CGAN) (Mirza and Osindero, 2014), there has been remarkable progress in this field (Denton et al., 2015; Odena et al., 2017; Miyato and Koyama, 2018). To the best of our knowledge, most generators in CGAN feed the class condition information (i.e., an class label vector), c, by simply concatenating c with noise φ to form an input which is then directly fed into the upsampling operations to generate an image. Another generative model, Conditional Variational Auto-Encoder (CVAE) (Sohn et al., 2015), also shows some promise for conditional image

generation (Mansimov et al., 2016; Yan et al., 2016). However, in such approaches, the generated images are usually blurry because of direct sampling from prior and the element-wise measures used.

In this paper, inspired by the variation inference employed in VAE, we propose a novel variational generator framework for CGAN. More concretely, in our proposed framework, variational inference only depending on the conditional vector c is introduced into the generator to infer the distribution of latent variable z, which represents the shared semantics across both text and image modalities. As the images are drawn from the inferred posterior in the generation phase, it overcomes the problem of mismatching from prior sampling in CVAE. Also, the reconstruction loss is augmented by minimizing the distance between the fake and true distributions through adversarial training. From another perspective, instead of upsampling from the concatenated representation in CGAN, the condition information is relatively fully exploited by posterior inference and a variable augmented representation (drawn from latent distribution) is supplied for image generation. By introducing variational inference, the fine-grained images with more visual details and richer diversity under the conditional constraint can be generated. The mode collapse problem that all outputs moving toward one or some fixed point can be partially addressed.

The main contributions of this paper are summarized as follows:

- We propose a novel variational generator framework for CGAN to improve the generation quality and diversity. The framework is flexible and can be applied in various tasks such as text-to-image generation. To the best of our knowledge, it is the first attempt to incorporate the variational inference only from the conditional input (without images) into the generator of CGAN, which guarantees images can be generated from the posterior after training.

- We propose a novel auxiliary classifier to better satisfy the class-conditional constraint. Experimental results show it accelerates adversarial training and avoid mode collapse problem than original version. We also incorporate a truncation technique as post-processing for latent space to further boost the generation performance.

- Qualitative and quantitative experiments are conducted on the class-conditional task, and results show that the proposed method outperforms the state-of-the-art approaches for class-conditional image generation. We also applied our method to the text-to-image generation task, the fine-grained images that match the sentence descriptions can be achieved.

2. Related work

Research on image generation based on natural language can be classified into two categories: *sentence-level image generation* and *class-conditional image generation*. *Sentence-level image generation* learns to generate related image from one sentence, which is also called text-to-image generation (Reed et al., 2016b). A number of end-to-end methods have been exploited to solve the problem. Mansimov et al. (2016) built an AlignDRAW model based on the recurrent variational auto-encoder to learn the alignment between text embeddings and the generating canvas. With CGAN, Reed et al. (2016b) generated plausible images for birds and flowers based on text descriptions. Following this way, Hong et al.

(2018) generated more fine-grained images by decomposing the generation process into multiple steps. However, the text descriptions that are used for generation usually have simple grammatical structures only with single entity (e.g. "*This bird is red and brown in color, with a stubby beak.*"). Moreover, there is no effective quantitative metric to measure the consistency of the generated image and a given text description. The end-to-end methods also lack the interpretability.

Class-conditional image generation (van den Oord et al., 2016; Odena et al., 2017) is dedicated to generate images from an image class label (i.e., an entity). It is relatively easier compared to sentence-level image generation since it does not require the understanding of semantic meanings encoded in sentences. Some quantitative metrics such as Inception score (IS) and Frechet Inception distance (FID) have been exploited to evaluate the quality and variety of the generated images. Based on CVAE, Yan et al. (2016) generated face images from the visual attributes extracted from a text description, through disentangling foreground and background of image. However a disadvantage of CVAE is that, the generated samples are often blurry because of the injected prior noise in the test phase and imperfect element-wise measures such as the squared error. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have shown promising performance for generating sharper images. Its variant, conditional GAN, has become a general framework for cross-modality transformation for various tasks (e.g., image-to-image translation and image captioning (Sun et al., 2019)) by using conditional information for the discriminator and generator. Odena et al. (2017) proposed a novel approach to incorporate the class-conditional information into the discriminator by adding an auxiliary classifier. Similarly, Miyato and Koyama (2018) proposed a projection-based discriminator to further improve the generation quality. Instead of improving the discriminator. Zhang et al. (2018) used the self-attention mechanism in the generator to make the generated image look more globally coherent.

However, to the best of our knowledge, most of methods based on CGAN generate images through a series of upsampling operations from the condition and noise. The concatenation representation of input is relatively simple and the hidden condition information is not fully exploited. It would be interesting to take into account the hidden semantics behind the conditional input to generate images. Recently, Bao et al. (2017) proposed a framework called CVAE-GAN, whcih uses the generator in CGAN as the decoder in CVAE to combine them and a feature matching loss to reconstruct images in feature space. However, it is still a CVAE-based framework, which can not overcome the blurry side effect by CVAE/VAE. Different from CVAE-GAN, we introduce a variational inference from control condition into the generator and the encoder can be reused in test phase for posterior inference, which we believe is a natural way to incorporate both advantages of VAE and GAN.

3. Methodology

We cast class-conditional image generation as a conditional likelihood maximization problem and define the problem setting in Section 3.1, followed by the proposed framework in Section 3.2. After that, we describe the training procedure in Section 3.3 and conditional data generation in Section 3.4.

3.1. Problem Setting

Given the condition variable $c \in \mathbb{R}^{N_c}$ (i.e., entities or image classes) and latent variable $z \in \mathbb{R}^{N_z}$, we aim to build a conditional generative model $p_\theta(x|z)$ to generate a realistic image $x \in \mathbb{R}^{N_x}$ conditioned on z, which is the hidden semantics behind c.

A traditional way of model learning is to maximize the variational lower bound of the conditional log-likelihood $\log p_\theta(x|c)$. Specifically, an auxiliary distribution $q_\phi(z|x,c)$ is introduced to approximate the true posterior $p_\theta(z|x,c)$. The conditional log-likelihood can be formulated below (Yan et al., 2016):

$$\begin{aligned} \log p_\theta(x|c) &= KL(q_\phi(z|x,c)||p_\theta(z|x,c)) + \mathcal{L}(x,c;\theta,\phi), \\ \mathcal{L}(x,c;\theta,\phi) &= -KL(q_\phi(z|x,c)||p_\theta(z)) - \mathbb{E}_{q_\phi(z|x,c)} L(g_\theta(c,z),x) \end{aligned} \tag{1}$$

where the Kullback-Leibler divergence $KL(q_\phi(z|x,c)||p_\theta(z))$ as a regularization loss reduces the gap between the prior $p_\theta(z)$ and the auxiliary posterior $q_\phi(z|x,c)$, and $L(g_\theta(c,z),x)$ is the reconstruction loss (e.g., ℓ_2 loss) between the generated image $g_\theta(c,z)$ and real image x. An encoder network $f(x,c)$ and a decoder network $g(c,z)$ are built for $q_\phi(z|x,c)$ and $p_\theta(x|c,z)$, respectively. During training, image is generated based on $g(c,z)$ where z is drawn from the inferred posterior $q_\phi(z|x,c)$. However, in the test phase, latent samples z are all drawn from the prior distribution $p_\theta(z)$ (usually, z is a Gaussian noise) instead of the posterior distribution as the real image x is not given. Obviously, there is a mismatch between the latent sample $z \sim p_\theta(z)$ and the condition c, which might generate unclear images. This mismatch problem is inherent as the latent distribution is not modeled explicitly.

We can however assume that the latent space has the local aggregated property, depending on the different conditions (manifold hypothesis). Based on the assumption, we directly infer the latent variable posterior $p_\theta(z|c)$ only depending on condition c (without image x) so that the images can be drawn from the inferred posterior $q_\phi(z|c)$ instead of the prior. Such modification brings another problem: there is no ground-truth x_{c_i} to calculate reconstruction loss $L(g_\theta(z),x_{c_i})$ directly as a specific condition c_i usually corresponds to a variety of real images x_{c_i}. Thus an adversarial loss $-\log D(g_\theta(z)|x,c)$ given by a discriminator D is introduced as the reconstruction loss to help reduce the distance of the generated images and the ground-truth. The total objective function of the encoder-decoder network is modified as below:

$$\mathcal{L}(x,c,\varphi;\ \theta,\phi) = -KL(q_\phi(z|c,\varphi)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|c,\varphi)} \log D(g_\theta(z)|x,c) \tag{2}$$

where noise φ is introduced as a part of the input of the encoder to compensate the disturbance caused by x.

3.2. The Proposed Framework

To solve the objective function defined in Eq. 2, we design a novel model architecture named Variational Conditional GAN (VCGAN). The VCGAN architecture as shown in Figure 1 contains the following three modules: (1) *Encoder Network* $F_\phi(c,\varphi)$, takes noise φ and condition c as input and encodes them as latent variable z; (2) *Decoder Network* $G_\theta(z)$, is designed to learn the distribution of real image x given the latent variable z; (3) *Discriminator Network* $D(x)$, as a supervisor to judge the image x and supply the reconstruction loss for the *encoder-decoder network* $G(c,\varphi)$ (the proposed variational generator framework). We explain the structure and function of each module of VCGAN in more details below:

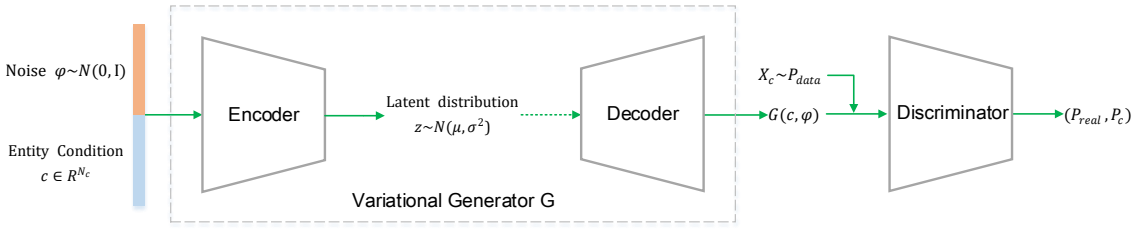


Figure 1: The Architecture of the Variational Conditonal GAN (VCGAN). The dashed line represents the sampling operation.

Encoder $F_\phi(c, \varphi)$: The encoder aims to conduct posterior inference of latent variable z given the condition variable $c \in \mathbb{R}^{N_c}$ with noise variable $\varphi \sim \mathcal{N}(0, I)$. The posterior $q_\phi(z|c, \varphi)$ is assumed as a diagonal Gaussian where the mean and covariance are parametrized by encoder $F_\phi(c, \varphi)$.

$$\begin{aligned} &\varphi \sim \mathcal{N}(0, I) \\ &(\mu, diag(\sigma^2)) = F_\phi(c, \varphi) \\ &z \sim \mathcal{N}(\mu, diag(\sigma^2)) \end{aligned} \tag{3}$$

where noise φ is drawn from standard multivariate Gaussian, the mean μ and covariance $diag(\sigma^2)$ of the latent variable z are estimated by the encoder.

The structure detail of the encoder is shown in Figure 2, we employed three linear layers in our experiments for simplicity. More complex neural network architecture could be used in the encoder for inference.

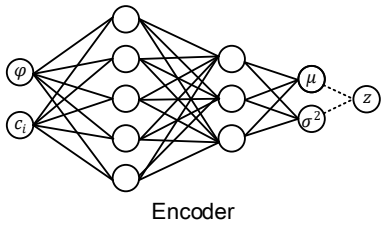


Figure 2: The architecture of the Encoder in VCGAN. Each sample is drawn from the latent distribution by the *reparameterization trick*.

Decoder $G_\theta(z)$: The decoder learns a function $f : \mathbb{Z} \to \mathbb{X}$, where $\mathbb{Z}$ represents the low dimensional latent space and $\mathbb{X}$ represents the high dimensional pixel space. The decoder takes the latent variable z drawn from the posterior $q_\phi(z|c, \varphi)$ as input and decodes it as the image x. The decoding process can be formulated below:

$$z \sim q_\phi(z|c, \varphi) \ ; \ x = G_\theta(z) \tag{4}$$

Discriminator $D(x)$: To calculate the reconstruction loss, we introduce the discriminator as a supervisor to reduce the distance of the generated images and the ideal ground-truth. Specifically it judges the perceptual fidelity of the generated image x and the consistency with corresponding condition c. In order to satisfy the above two requirements, we design a novel objective based on a auxiliary classifier (Odena et al., 2017). The discriminator outputs both a probability distribution over sources s, $p(s|x)$, and a probability distribution over the conditions c, $p(c|x)$. The objective function of $D(x)$ includes two parts: the log-likelihood of the correct source, L_s, and the log-likelihood of the correct condition, L_c.

$$L_s = \mathbb{E}[\log p(s = real|x_{real})] + \mathbb{E}[\log p(s = fake|x_{fake})] \tag{5}$$

$$L_c = \mathbb{E}[\log p(c = c_x|x_{real})] + \mathbb{E}[\log p(c = Others|x_{fake})] \tag{6}$$

where Eq. 5 is the standard GAN objective (Goodfellow et al., 2014), it minimizes the negative log-likelihood for the binary classification task (*is the sample true or fake?*) and is equivalent to minimize the Jensen-Shannon divergence between the true data distribution P and the model distribution Q.

It should be pointed out that we introduce $\mathbb{E}[\log p(c = Others|x_{fake})]$ in Eq. 6 instead of the term $\mathbb{E}[\log p(c = c_x|x_{fake})]$ as in original version. This is to introduce an additional class "*Others*" to represent that the image x does not belong to any of the known conditions. Experimental results show that labeling the x_{fake} with "*Others*" makes the training converge faster.

We inline the modified classifier to the discriminator as a one-ve-many multi-task learning by sharing the hidden layers to improve each other. Suppose there are a total of K classes or conditions, then we have $K + 2$ output units in the discriminator network. We use a Sigmoid activation function in the first output unit and the Softmax function over the remaining K+1 output units (include an extra class "*Others*"). The probability of the image x belonging to sources s and conditions c can be written below:

$$\begin{aligned} P(s = real|x) &= sigmoid\left(h(x) \cdot \boldsymbol{W}_0 + b_0\right) \\ P(c = c_i|x) &= \frac{\exp\left(h(x) \cdot \boldsymbol{W}_i + b_i\right)}{\sum_{i=1}^{K+1} \exp\left(h(x) \cdot \boldsymbol{W}_i + b_i\right)} \end{aligned} \tag{7}$$

where $h(x)$ is the hidden layer representation of image x, $\boldsymbol{W} \in \mathbb{R}^{H\times(K+2)}$ denotes the weight matrix of the output layer, $\boldsymbol{W}_0$ is the weights of the first output unit.

3.3. Training

The final loss functions of the proposed framework are given as follows:

$$\begin{aligned} \mathcal{L}_D = &- \left(\mathbb{E}_{x\sim P}[\log D(x)_0] + \mathbb{E}_{x\sim Q}[\log(1 - D(x)_0)]\right) \\ &- \left(\mathbb{E}_{x\sim P}[\log D(x)_1] + \mathbb{E}_{x\sim Q}[\log D(x)_1]\right) \end{aligned} \tag{8}$$

$$\mathcal{L}_G = - \mathbb{E}_{x\sim Q}[\log D(x)_0] - \mathbb{E}_{x\sim Q}[\log D(x)_1] + KL(q(z|c,\varphi)||p(z)) \tag{9}$$

where P is true data distribution and Q is the generated data distribution. The $D(x)_0$ denotes the probability of image x being real, the $D(x)_1$ denotes the the probability over the label condition (including the "*Others*" class). The KL divergence of the latent prior $p(z)$ from $q(z|c,\varphi)$ is added to $\mathcal{L}_G$ in Eq. 9, as the regularization loss for constraining the latent z.

Based on the assumptions that the latent posterior is a diagonal Gaussian with mean μ and standard deviation σ and the prior is a standard Gaussian with mean 0 and standard deviation 1, in which case the KL term in Eq. 9 becomes

$$KL(q(z|c,\varphi)||p(z)) = -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2\right) \tag{10}$$

In the training stage, we first optimize the discriminator $D(x)$ under the loss $\mathcal{L}_D$ with the fixed generator, and then optimize the generator $G(c,\varphi)$ under the loss $\mathcal{L}_G$ (note the generated image is labeled with the fed condition c not "Others" when training G) with

the fixed discriminator. The above two steps alternates (adversarial training) by minibatch stochastic gradient descent.

As GAN training is relatively unstable and hard to optimize, we investigate some techniques to stabilize the adversarial training. We use batch normalizationin both the generator and the discriminator, and only use spectral normalization (Miyato et al., 2018) in the discriminator to make the training more stable.

3.4. Conditional Image Generation

An image x can be generated controllably from a learned VCGAN model by picking latent sample z from a inferred posterior $q(z|c, \varphi)$, then running the decoder to generate the image x fulfilling the condition c. The generation process in test phase can be described in the following:

$$\begin{aligned} &\varphi \sim \mathcal{N}(0, I) \\ &q(z|\varphi, c) = F_\phi(c, \varphi) \\ &z \sim q(z|c, \varphi) \\ &x = G_\theta(z) \end{aligned} \tag{11}$$

Note that we reuse the encoder to obtain inferred posterior $q(z|\varphi, c)$ not the fixed prior $p(z)$ as in CVAE.

We also incorporate a *Truncation Technique* as a post-processing step for further enhancing the quality of the generated images. Specifically, we draw the latent sample z from a truncated posterior (e.g., a truncated Gaussian distribution) then feed it into the decoder. The latent variable z will be resampled if it exceeds the truncation range. Using this technique, we can achieve a higher generation quality than sampling from the whole latent space.

4. Experiments

We first conduct extensive class-conditional experiments on two popular datasets, CIFAR10 (Krizhevsky and Hinton, 2009) and FASHION-MNIST (Xiao et al., 2017), which are widely employed by many state-of-the-art conditional image generation approaches. The two datasets both include ten different categories of objects. Also, we conduct sentence-level experiments on another two datasets, CUB (Wah et al., 2011) and Oxford Flower (Nilsback and Zisserman, 2008), which include 10 captions for each image provided by Reed et al. (2016a). They include many subcategories of birds and flowers. The statistics of the these datasets are presented in Table 1.

Datasets	CIFAR10	FASHION-MNIST	CUB	Oxford Flower
Training images	50,000	60,000	8,855	7,034
Test images	10,000	10,000	2,933	1,155
Resolution	32x32	28x28	128x128	128x128
Class labels	"Plane","Car","Bird", "Cat","Deer","Dog","Frog", "Horse","Ship","Truck"	"Tshirt","Trouser","Pullover", "Dress","Coat","Sandal","Shirt", "Sneaker","Bag","Ankle boot"	200	102
Captions	–	–	10	10

Table 1: Statistics of the datasets.

Two quantitative metrics, Inception score (IS) (Salimans et al., 2016) and Frechet Inception distance (FID) (Heusel et al., 2017) are employed. IS uses an pre-trained Inception net on ImageNet to calculate the statistic of the generated images, which is defined as:

$$\text{IS} = \exp\left(\mathbb{E}_{x\sim Q}[KL(p(y|x)||p(y))]\right) \tag{12}$$

where $p(y|x)$ is the conditional class distribution given by Inception net, and $p(y) = \int_x p(y|x)p(x)$ is the marginal class distribution. The higher IS indicates the generated images contain clear recognizable objects. However, as pointed out in (Zhang et al., 2018), IS can not assess the perceptual fidelity of details and intra-class diversity. FID is a more principled metric, which can detect intra-class mode collapse (Kurach et al., 2018). By assuming that the sample embeddings follow a multivariate Gaussian distribution, FID measures the Wasserstein-2 distance between two Gaussian distributions, which is defined as:

$$\text{FID} = \|\mu_{x_1} - \mu_{x_2}\|_2^2 + \text{Tr}\left(\Sigma_{x_1} + \Sigma_{x_2} - 2\left(\Sigma_{x_1}\Sigma_{x_2}\right)^{1/2}\right) \tag{13}$$

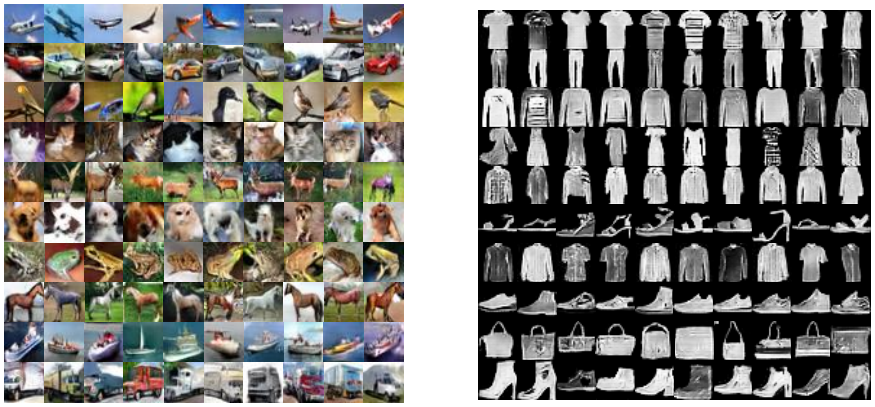
where x_1 and x_2 are samples from P and Q. The lower FID value, the closer distance between the synthetic and the real data distributions. In all our experiments, 50k samples divided into 10 groups are randomly generated to compute the Inception scores and 10k samples are generated to compute FID.

Seven state-of-the-art approaches chosen as the baselines:

- DCGAN (Radford et al., 2015), a deep convolutional architecture of GAN.
- LSGAN (Mao et al., 2017), which uses a least-squares loss instead of the standard GAN loss which minimizes the Pearson χ^2 divergence between P and Q.
- AC-GAN (Odena et al., 2017), an auxiliary classifier is introduced into the discriminator for class-constraint.
- WGAN (Arjovsky et al., 2017), which minimizes the Wasserstein distance between P and Q.
- WGAN-GP (Gulrajani et al., 2017), which uses the Gradient Penalty as a soft penalty for the violation of 1-Lipschitzness in WGAN.
- CVAE-GAN (Bao et al., 2017), a CAVE-based framework with a discriminator for feature matching.
- SNHGAN-Proj (Miyato and Koyama, 2018), a spectrally normalized GAN was combined with the projection-based discriminator with hinge loss.

The proposed model was implemented based on a recent robust architecture called SNDCGAN (Miyato et al., 2018). Another common architecture is "ResNet" (Gulrajani et al., 2017). It is deeper than SNDCGAN, which we use in the sentence-level experiments. The proposed encoder is implemented by three-layer FCs with 512, 256, and 128 units. The CVAE-GAN is also implemented by SNDCGAN with a convolutional image encoder for better comparsion with ours. For all experiments, we fix the size of the latent variable and noise variable to 128, encode the class conditions as one-hot representations. We choose

Adam solver with hyper-parameters set to $\beta_1 = 0.5, \beta_2 = 0.999$ and the learning rate $\alpha = 0.0002$ by default. The batch size is set to 100 and the balanced update frequencies (1:1) of discriminator and generator are employed. For sentence-level experiments, a smaller batch size (64) are employed. We use the char-CNN-RNN text encoder provided by Reed et al. (2016a) to encode each sentence into a 1024-d text embedding as the conditional vector.



(a) CIFAR10 samples. (b) FASHION-MNIST samples.

Figure 3: The generated samples controlled by the class labels with the proposed model.

4.1. Performance Comparison

To evaluate the effectiveness of the proposed model, the class-conditional experiments are conducted on the CIFAR10 and FASHION-MNIST datasets. The results are presented in the Table 2 and Table 3. Sample images generated by VCGAN are illustrated in Figure 3a and 3b.

Method	Inception Score ↑	FID ↓
DCGAN	6.58	–
AC-GAN	$8.25 \pm .07$	–
WGAN-GP (ResNet)	$8.42 \pm .10$	19.5
SNHGAN-Proj (ResNet)	8.62	17.5
CVAE-GAN	$7.92 \pm .10$	24.1
VCGAN	$\mathbf{8.90 \pm .16}$	**16.9**

Table 2: Comparison with the baselines on CIFAR10 (with truncation post-processing). Some results are collected from Gulrajani et al. (2017) and Miyato and Koyama (2018).

It can be observed from Table 2 and 3, the proposed approach (VCGAN) achieves the best IS and FID on both datasets. On FASHION-MNIST, FID is significantly improved from 15.9 to 13.8 by VCGAN. Furthermore, some baselines relied on a deep ResNet network, which is more complex than SNDCGAN we used and needs more training resources. It should be pointed out that instead of carefully selecting some modified versions of GAN loss function (e.g., Wasserstein loss or Hinge loss) in the baselines, our VCGAN simply used the original standard GAN loss. And also, VCGAN significantly outperforms the similar

Method	Inception Score ↑	FID ↓
DCGAN	$4.15 \pm .04$	–
WGAN	$3.00 \pm .03$	21.5
LSGAN	$4.45 \pm .03$	30.7
CVAE-GAN	$4.28 \pm .04$	15.9
VCGAN	$\mathbf{4.75 \pm .06}$	**13.8**

Table 3: Comparison with the baselines on FASHION-MNIST. Some results are collected from Nandy et al. (2018) and Lucic et al. (2018).

model, CVAE-GAN, on two datasets with similar settings. Figure 3a and Figure 3b shows the clear and realistc images with hige variety can be achieved by VCGAN.

Method	Inception Score ↑					FID ↓				
Truncation	normal	2σ	1.5σ	σ	0.5σ	normal	2σ	1.5σ	σ	0.5σ
CVAE	$4.00 \pm .03$	$4.12 \pm .05$	$4.23 \pm .05$	$4.52 \pm .05$	$5.06 \pm .02$	105.9	103.5	100.7	95.5	91.9
Concat-CGAN	$6.51 \pm .08$	$6.79 \pm .07$	$6.98 \pm .08$	$7.37 \pm .10$	$7.08 \pm .05$	34.7	32.3	30.6	30.4	45.9
CBN-CGAN	$7.33 \pm .10$	$7.35 \pm .10$	$7.33 \pm .08$	$7.46 \pm .07$	$7.29 \pm .09$	29.4	29.5	28.7	28.4	**30.0**
CVAE-GAN	$7.72 \pm .08$	$7.86 \pm .07$	$7.92 \pm .10$	$7.91 \pm .09$	$7.37 \pm .08$	25.5	24.3	24.1	28.0	45.6
VCGAN	$\mathbf{8.43 \pm .13}$	$\mathbf{8.62 \pm .10}$	$\mathbf{8.90 \pm .16}$	$\mathbf{8.80 \pm .13}$	$\mathbf{7.66 \pm 0.05}$	**17.6**	**16.9**	**17.3**	**20.5**	37.2

Table 4: Ablation on CIFAR10 with different truncated ranges.

4.2. Ablation Comparison on CIFAR10

To further evaluate the effectiveness of the proposed variational generator framework, we conduct an ablation comparison with CVAE and CGAN models on CIFAR10. We also make a comparison with CVAE-GAN. For CVAE, we follow the general setting by concatenating image and class condition as input and feeding it into the same architecture (without discriminator). For CGAN, there are two ways of feeding the condition c into the generator (without encoder). There is usually to simply concatenate c with the noise φ as input as stated before, which we called Concat-CGAN. Recently, Miyato and Koyama (2018) used conditional batch normalization (Dumoulin et al., 2017) instead of batch normalization to feed condition information into the each layer of the generator (except the output layer), which we called CBN-CGAN. The same loss functions in section 3.3 except KL term are employed for Concat-CGAN and CBN-CGAN. The truncation technique mentioned is applied on all models and the truncated ranges are set by three-sigma rule, e.g., the range 2σ means samples are sampled (or re-sampled) from $\mu \pm 2\sigma$, *normal* means no truncation for the original latent space (noise space).

The results of various approaches on CIFAR10 are reported in Table 4. It can be observed that our VCGAN significantly outperforms all opponents on different metrics and ranges. It shows that VCGAN has a more powerful generation capability and can perform latent posterior inference only from the class condition to achieve richer semantic details. Compared with CGAN, VCGAN achieves much lower FID values. It shows that by incorporating the variational inference into the generator, VCGAN can better approximate the real image distribution and owns richer diversity of images. Compared with CVAE-GAN, VCGAN is simple in framework and losses but more effective in results, which we attribute to the CGAN-based framework and the reusable encoder for posterior inference.

Figure 4 shows the samples generated by the five different approaches from various z (or noise). The samples generated by CVAE seem very blurry even not related with the class. As stated before, CVAE suffers from the pixel-wise measure and the prior sampling. The Concat-CGAN and CBN-CGAN are more shaper, but there are much noise in the samples. Relatively, the CBN-CGAN performs better than Concat-CGAN. CVAE-GAN is much better compared with the former, however it is blurry and distorted in the detail. Our proposed VCGAN model achieves clearer and more natural images with more visual details and also does well in the diversity.

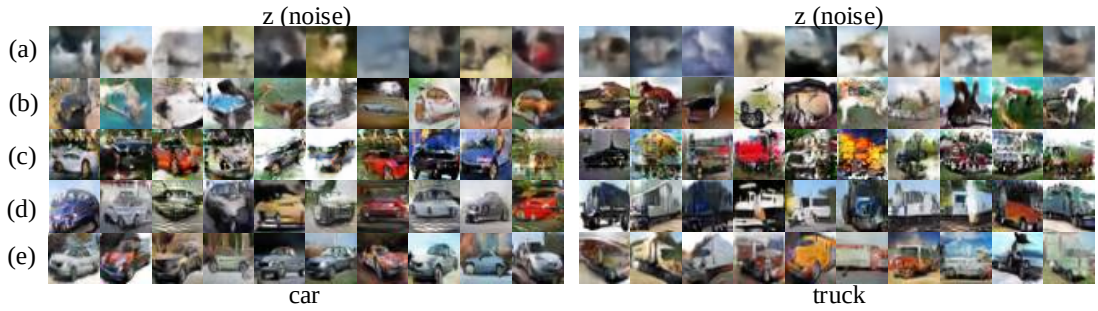
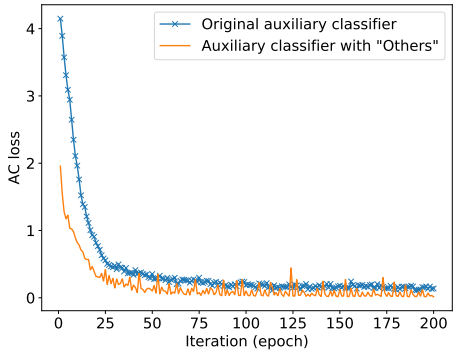


Figure 4: Comparison of the images generated by (a) CVAE; (b) Concat-CGAN; (c) CBN-CGAN; (d) CVAE-GAN; (e) VCGAN. Each column corresponds to a fixed latent sample z or noise.

4.3. Effect of Modified Auxiliary Classifier

We further investigate the effect of incorporating the modified auxiliary classifier (AC) loss on CIFAR10. The training curves are shown as Figure 5a. It can be observed that the modified version coverages faster and has lower classification error, which helps the discriminator accelerate training and get closer to the optimal state. It can also be found that the images generated by the original AC loss might lead to the mode collapse (Miyato and Koyama, 2018), which is hardly discovered in the experiments using the modified AC loss. Some collapsed images by original auxiliary classifier are shown in Figure 5b.



(a) Training curves of the modified auxiliary classifier and the original version under the same setting.

(b) The collapsed images generated by original auxiliary classifier under random class-conditional input.

Figure 5: The analysis for the modified auxiliary classifier on CIFAR10.

4.4. Sentence-level Image Generation

Although we believe it is relatively hard to generate an image from a sentence without the understanding for entity concepts, we also directly apply the proposed generator framework to the text-to-image generation task to evaluate the generalizability of the framework. Unlike class-conditional task, a text encoder is needed to embed the sentence semantics into the conditional vector. We use a hybrid character-level convolutional-recurrent network (Reed et al., 2016a) to calculate the semantic embeddings from text descriptions. The dimension of the sentence embedding is 1024. The dimension of noise and latent variable is 128. The output dimension of generator is set to 128 x 128. We conduct a pre-processing on CUB and Oxford Flower datasets, and all images are randomly cropped to 128 x 128 and flipped horizontally for data augmentation. Following the setup in (Reed et al., 2016b), we split CUB and Oxford Flower into class-disjoint training and test sets. we randomly pick an view (e.g., crop, flip) of the image and one of the captions as a pair for mini-batch training. The minibatch size is set to 64 and the model is trained for 300 epochs.

The Inception Scores on CUB is $5.00 \pm .11$ and on Oxford Flower is $2.95 \pm .03$. The generated samples controlled by text descriptions are presented in Figure 6. We also conduct class-conditional experiments on CUB and Oxford Flower. The generated samples of birds controlled by class labels are illustrated in Figure 7. The Inception Scores on CUB and Oxford Flower datasets are $6.43 \pm .09$ and $2.62 \pm .03$.

Figure 6: The generated 128 x 128 images controlled by text descriptions on CUB testset.

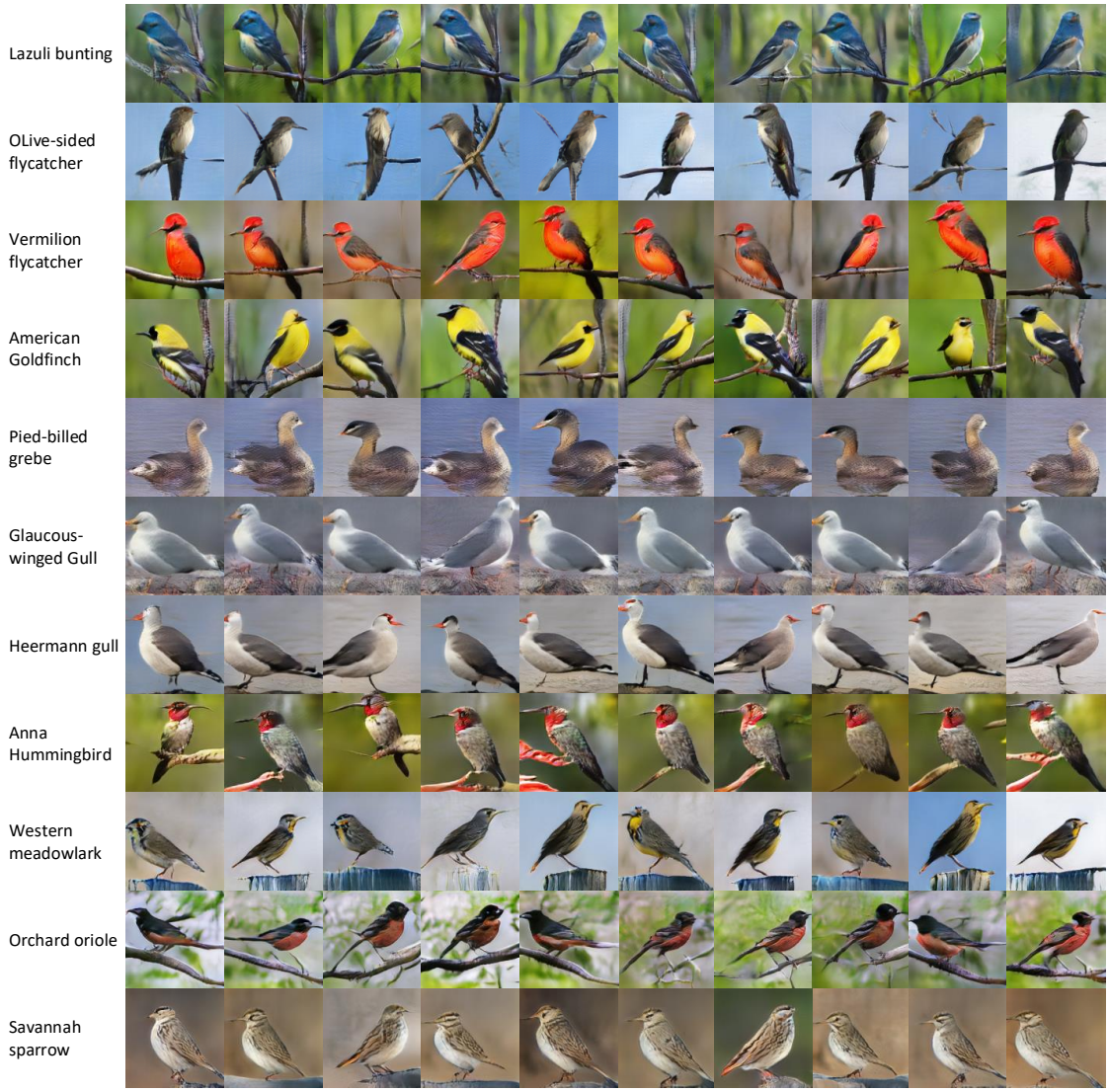


Figure 7: The generated 128 x 128 images controlled by class labels on CUB.

As shown in Figure 6, our model can achieve high resolution realistic images from text descriptions. The generated images have a fine-grained match with the descriptions such as the beak shape and the vivid parts. The Figure 7 also presents the high resolution class-conditional results, which show the realistic details (e.g., feather texture) and rich variety under the class labels. And the generated background is also clear and suitable.

To further explore the smooth property of latent data manifold learned by our model, the linear interpolation results between two different text descriptions are shown in Figure 8. The noise is fixed so that the images variety is only influenced by the sentence semantics. We can see the interpolated images have a gradual change in the colors of different parts. It is also noted that the interpolated images have a tiny variety in the pose and background

though the noise is fixed. We speculate that the learned latent distribution rather than a fixed representation introduce some disturbance to yield the richer variety.

Figure 8: Images generated by interpolating between two sentence embeddings (Left to right). The noise is fixed for each row.

5. Conclusion

In this paper, we have proposed a variational generator framework for conditional GANs to catch semantic details behind the conditional input and the fine-grained images with rich diversity can be achieved. The variational inference depending on the conditional input is introduced into the generator to generate images from the inferred latent posterior. Experimental results show that the proposed model outperforms the state-of-the-art approaches on the class-conditional task and we also demonstrate the generalizability of the proposed framework in the more complex sentence-level task. The truncation technique is incorporated to further boost the generation performance. Furthermore, the modified auxiliary classifier can make training converge faster and weaken mode collapse. In the future work, we hope to generate images with multiple entities from sentence descriptions by introducing the knowledge of entity concepts.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Jianmin Bao, Chen Dong, Wen Fang, Houqiang Li, and Hua Gang. Cvae-gan: Fine-grained image generation through asymmetric training. In *IEEE International Conference on Computer Vision*, 2017.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations*, 2017.

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*, pages 2672–2680, 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*, pages 5767–5777, 2017.

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*, pages 6626–6637, 2017.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018.

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

Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. The GAN landscape: Losses, architectures, regularization, and normalization. *CoRR*, abs/1807.04720, 2018.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 697–706, 2018.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *International Conference on Learning Representations*, 2016.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2813–2821. IEEE Computer Society, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

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

Jay Nandy, Wynne Hsu, and Mong Li Lee. Normal similarity network for generative modelling. *arXiv preprint arXiv:1805.05269*, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2642–2651. PMLR, 2017.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition*, 2016a.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1060–1069. PMLR, 2016b.

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*, pages 2234–2242, 2016.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

Zhaoyue Sun, Jiaze Chen, Hao Zhou, Deyu Zhou, Lei Li, and Mingmin Jiang. Graspsnooker: Automatic chinese commentary generation for snooker videos. In *IJCAI-19*, 2019.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.