Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis

Qi Mao[*1], Hsin-Ying Lee[*2], Hung-Yu Tseng[*2], Siwei Ma[1,3], Ming-Hsuan Yang[2,4]

[1]Institute of Digital Media, Peking University   [2]University of California, Merced
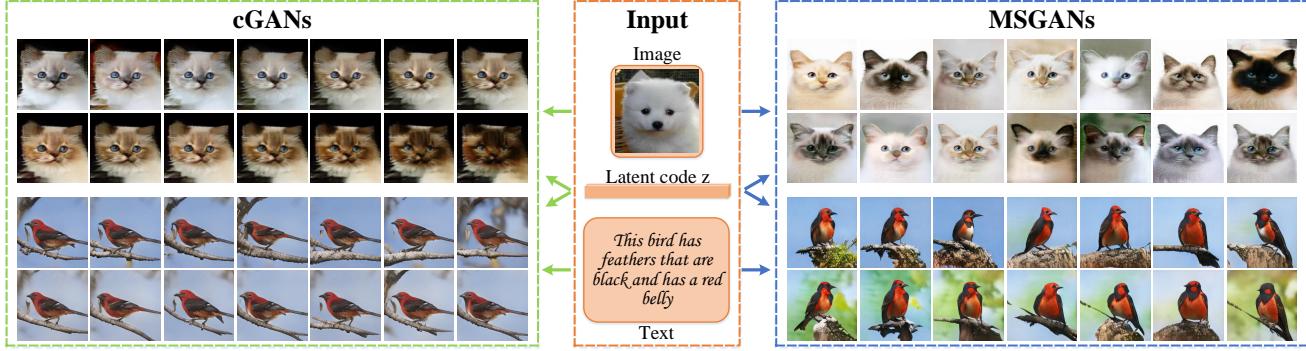[3]Peng Cheng Laboratory   [4]Google Cloud

Figure 1: **Mode seeking generative adversarial networks (MSGANs).** (*Left*) Existing conditional generative adversarial networks tend to ignore the input latent code **z** and generate images of similar modes. (*Right*) We propose a simple yet effective mode seeking regularization term that can be applied to arbitrary conditional generative adversarial networks in different tasks to alleviate the mode collapse issue and improve the diversity.

Abstract

Most conditional generation tasks expect diverse outputs given a single conditional context. However, conditional generative adversarial networks (cGANs) often focus on the prior conditional information and ignore the input noise vectors, which contribute to the output variations. Recent attempts to resolve the mode collapse issue for cGANs are usually task-specific and computationally expensive. In this work, we propose a simple yet effective regularization term to address the mode collapse issue for cGANs. The proposed method explicitly maximizes the ratio of the distance between generated images with respect to the corresponding latent codes, thus encouraging the generators to explore more minor modes during training. This mode seeking regularization term is readily applicable to various conditional generation tasks without imposing training overhead or modifying the original network structures. We validate the proposed algorithm on three conditional image synthesis tasks including categorical generation, image-to-image translation, and text-to-image synthesis with different baseline models. Both qualitative and quantitative results demonstrate the effectiveness of the proposed regularization method for improving diversity without loss of quality.

1. Introduction

Generative adversarial networks (GANs) [8] have been shown to capture complex and high-dimensional image data with numerous applications effectively. Built upon GANs, conditional GANs (cGANs) [21] take external information as additional inputs. For image synthesis, cGANs can be applied to various tasks with different conditional contexts. With class labels, cGANs can be applied to categorical image generation. With text sentences, cGANs can be applied to text-to-image synthesis [23, 32]. With images, cGANs have been used in tasks including image-to-image translation [10, 11, 15, 17, 34, 35], semantic manipulation [30] and style transfer [16].

For most conditional generation tasks, the mappings are in nature multimodal, i.e., a single input context corresponds to multiple plausible outputs. A straightforward approach to handle multimodality is to take random noise vectors along with the conditional contexts as inputs, where the contexts determine the main content and noise vectors are responsible for variations. For instance, in the dog-to-cat image-to-image translation task [15], the input dog images decide contents like orientations of heads and positions of facial landmarks, while the noise vectors help the generation of different species. However, cGANs usually suffer from the *mode collapse* [8, 26] problem, where generators

1

only produce samples from a single or few modes of the distribution and ignore other modes. The noise vectors are ignored or of minor impacts, since cGANs pay more attention to learn from the high-dimensional and structured conditional contexts.

There are two main approaches to address the mode collapse problem in GANs. A number of methods focus on discriminators by introducing different divergence metrics [1, 19] and optimization process [6, 20, 26]. The other methods use auxiliary networks such as multiple generators [7, 18] and additional encoders [2, 4, 5, 27]. However, mode collapse is relatively less studied in cGANs. Some recent efforts have been made in the image-to-image translation task to improve diversity [10, 15, 35]. Similar to the second category with the unconditional setting, these approaches introduce additional encoders and loss functions to encourage the one-to-one relationship between the output and the latent code. These methods either entail heavy computational overheads on training or require auxiliary networks that are often task-specific that cannot be easily extended to other frameworks.

In this work, we propose a mode seeking regularization method that can be applied to cGANs for various tasks to alleviate the mode collapse problem. Given two latent vectors and the corresponding output images, we propose to maximize the ratio of the distance between images with respect to the distance between latent vectors. In other words, this regularization term encourages generators to generate dissimilar images during training. As a result, generators can explore the target distribution, and enhance the chances of generating samples from different modes. On the other hand, we can train the discriminators with dissimilar generated samples to provide gradients from minor modes that are likely to be ignored otherwise. This mode seeking regularization method incurs marginal computational overheads and can be easily embedded in different cGAN frameworks to improve the diversity of synthesized images.

We validate the proposed regularization algorithm through an extensive evaluation of three conditional image synthesis tasks with different baseline models. First, for categorical image generation, we apply the proposed method on DCGAN [22] using the CIFAR-10 [13] dataset. Second, for image-to-image translation, we embed the proposed regularization scheme in Pix2Pix [11] and DRIT [15] using the facades [3], maps [11], Yosemite [34], and cat$\rightleftharpoons$dog [15] datasets. Third, for text-to-image synthesis, we incorporate StackGAN++ [32] with the proposed regularization term using the CUB-200-2011 [29] dataset. We evaluate the diversity of synthesized images using perceptual distance metrics [33].

However, the diversity metric alone cannot guarantee the similarity between the distribution of generated images and the distribution of real data. Therefore, we adopt two recently proposed bin-based metrics [24], the *Number of Statistically-Different Bins* (NDB) metric which determines the relative proportions of samples fallen into clusters predetermined by real data, and the *Jensen-Shannon Divergence* (JSD) distance which measures the similarity between bin distributions. Furthermore, to verify that we do not achieve diversity at the expense of realism, we evaluate our method with the Fréchet Inception Distance (FID) [9] as the metric for quality. Experimental results demonstrate that the proposed regularization method can facilitate existing models from various applications achieving better diversity without loss of image quality. Figure 1 shows the effectiveness of the proposed regularization method for existing models.

The main contributions of this work are:

- We propose a simple yet effective mode seeking regularization method to address the mode collapse problem in cGANs. This regularization scheme can be readily extended into existing frameworks with marginal training overheads and modifications.
- We demonstrate the generalizability of the proposed regularization method on three different conditional generation tasks: categorical generation, image-to-image translation, and text-to-image synthesis.
- Extensive experiments show that the proposed method can facilitate existing models from different tasks achieving better diversity without sacrificing visual quality of the generated images.

Our code and pre-trained models are available at https://github.com/HelenMao/MSGAN/.

2. Related Work

Conditional generative adversarial networks. Generative adversarial networks [1, 8, 19, 22] have been widely used for image synthesis. With adversarial training, generators are encouraged to capture the distribution of real images. On the basis of GANs, *conditional* GANs synthesize images based on various contexts. For instances, cGANs can generate high-resolution images conditioned on low-resolution images [14], translate images between different visual domains [10, 11, 15, 17, 34, 35], generate images with desired style [16], and synthesize images according to sentences [23, 32]. Although cGANs have achieved success in various applications, existing approaches suffer from the mode collapse problem. Since the conditional contexts provide strong structural prior information for the output images and have higher dimensions than the input noise vectors, generators tend to ignore the input noise vectors, which are responsible for the variation of generated images. As a result, the generators are prone to produce images with similar appearances. In this work, we aim to address the mode collapse problem for cGANs.
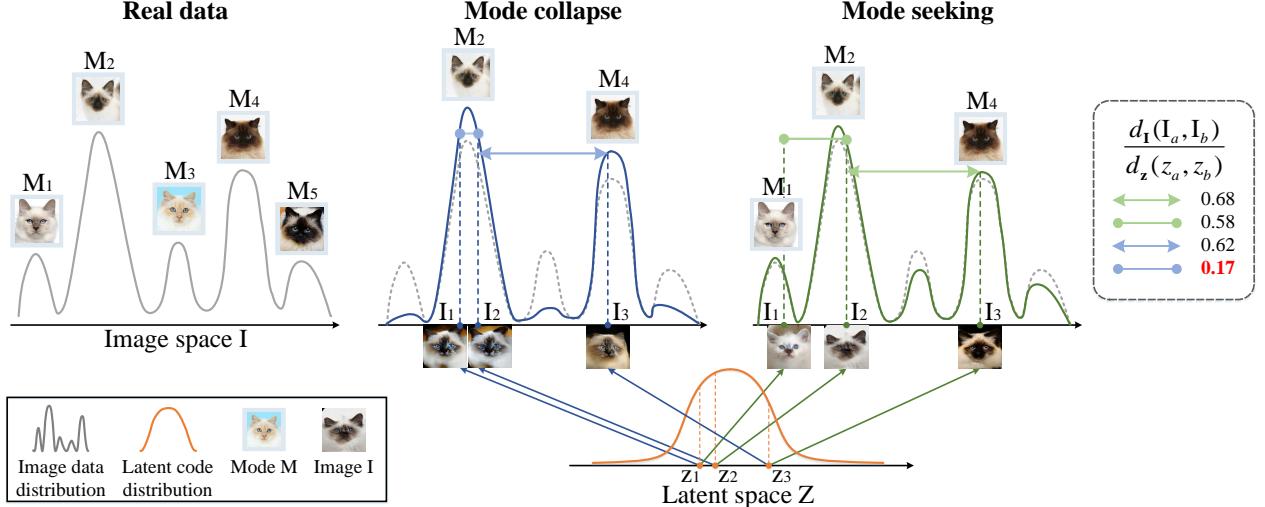
Figure 2: Illustration of motivation. Real data distribution contains numerous modes. However, when mode collapse occurs, generators only produce samples from a few modes. From the data distribution when mode collapse occurs, we observe that for latent vectors z_1 and z_2, the distance between their mapped images I_1 and I_2 will become shorter in a disproportionate rate when the distance between two latent vectors is decreasing. We present on the right the ratio of the distance between images with respect to the distance of the corresponding latent vectors, where we can spot an anomalous case (colored in red) where mode collapse occurs. The observation motivates us to leverage the ratio as the training objective explicitly.

Reducing mode collapse. Some methods focus on the discriminator with different optimization process [20] and divergence metrics [1, 19] to stabilize the training process. The minibatch discrimination scheme [26] allows the discriminator to discriminate between whole mini-batches of samples instead of between individual samples. In [6], Durugkar et al. use multiple discriminators to address this issue. The other methods use auxiliary networks to alleviate the mode collapse issue. ModeGAN [2] and VEE-GAN [27] enforce the bijection mapping between the input noise vectors and generated images with additional encoder networks. Multiple generators [7] and weight-sharing generators [18] are developed to capture more modes of the distribution. However, these approaches either entail heavy computational overheads or require modifications of the network structure, and may not be easily applicable to cGANs.

In the field of cGANs, some efforts [10, 15, 35] have been recently made to address the mode collapse issue on the image-to-image translation task. Similar to ModeGAN and VEEGAN, additional encoders are introduced to provide a bijection constraint between the generated images and input noise vectors. However, these approaches require other task-specific networks and objective functions. The additional components make the methods less generalizable and incur extra computational loads on training. In contrast, we propose a simple regularization term that imposes no training overheads and requires no modifications of the

network structure. Therefore, the proposed method can be readily applied to various conditional generation tasks. Recently, the concurrent work [31] also adopts a loss term similar to our work for reducing mode collapse for cGANs.

3. Diverse Conditional Image Synthesis

3.1. Preliminaries

The training process of GANs can be formulated as a mini-max problem: a discriminator D learns to be a classifier by assigning higher discriminative values to the real data samples and lower ones to the generated ones. Meanwhile, a generator G aims to fool D by synthesizing realistic examples. Through adversarial training, the gradients from D will guide G toward generating samples with the distribution similar to the real data one.

The mode collapse problem with GANs is well known in the literature. Several methods [2, 26, 27] attribute the missing mode to the lack of penalty when this issue occurs. Since all modes usually have similar discriminative values, larger modes are likely to be favored through the training process based on gradient descent. On the other hand, it is difficult to generate samples from minor modes.

The mode missing problem becomes worse in cGANs. Generally, conditional contexts are high-dimensional and structured (*e.g.*, images and sentences) as opposed to the noise vectors. As such, the generators are likely to focus on the contexts and ignore the noise vectors, which account for diversity.

3

3.2. Mode Seeking GANs

In this work, we propose to alleviate the missing mode problem from the generator perspective. Figure 2 illustrates the main ideas of our approach. Let a latent vector $\mathbf{z}$ from the latent code space Z be mapped to the image space I. When mode collapse occurs, the mapped images are collapsed into a few modes. Furthermore, when two latent codes $\mathbf{z}_1$ and $\mathbf{z}_2$ are closer, the mapped images $\mathbf{I}_1 = G(c, \mathbf{z}_1)$ and $\mathbf{I}_2 = G(c, \mathbf{z}_2)$ are more likely to be collapsed into the same mode. To address this issue, we propose a mode seeking regularization term to directly maximize the ratio of the distance between $G(c, \mathbf{z}_1)$ and $G(c, \mathbf{z}_2)$ with respect to the distance between $\mathbf{z}_1$ and $\mathbf{z}_2$,

$$\mathcal{L}_{\mathrm{ms}} = \max_G \left(\frac{d_{\mathbf{I}}(G(c, \mathbf{z}_1), G(c, \mathbf{z}_2))}{d_{\mathbf{z}}(\mathbf{z}_1, \mathbf{z}_2)} \right), \qquad (1)$$

where $d_*(\cdot)$ denotes the distance metric.

The regularization term offers a virtuous circle for training cGANs. It encourages the generator to explore the image space and enhances the chances for generating samples of minor modes. On the other hand, the discriminator is forced to pay attention to generated samples from minor modes. Figure 2 shows a mode collapse situation where two close samples, $\mathbf{z}_1$ and $\mathbf{z}_2$, are mapped onto the same mode M_2. However, with the proposed regularization term, $\mathbf{z}_1$ is mapped to $\mathbf{I}_1$, which belongs to an unexplored mode M_1. With the adversarial mechanism, the generator will thus have better chances to generate samples of M_1 in the following training steps.

As shown in Figure 3, the proposed regularization term can be easily integrated with existing cGANs by appending it to the original objective function.

$$\mathcal{L}_{\mathrm{new}} = \mathcal{L}_{\mathrm{ori}} + \lambda_{\mathrm{ms}} \mathcal{L}_{\mathrm{ms}}, \qquad (2)$$

where $\mathcal{L}_{\mathrm{ori}}$ denotes the original objective function and λ_{ms} the weights to control the importance of the regularization. Here, $\mathcal{L}_{\mathrm{ori}}$ can be as a simple loss function. For example, in categorical generation task,

$$\mathcal{L}_{\mathrm{ori}} = \mathbb{E}_{c,\mathbf{y}}[\log D(c, \mathbf{y})] + \mathbb{E}_{c,\mathbf{z}}[\log\left(1 - D(c, G(c, \mathbf{z}))\right)], \qquad (3)$$

where $c, \mathbf{y}, \mathbf{z}$ denote class labels, real images, and noise vectors, respectively. In image-to-image translation task [11],

$$\mathcal{L}_{\mathrm{ori}} = \mathcal{L}_{\mathrm{GAN}} + \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{z}}[\|\mathbf{y} - G(\mathbf{x}, \mathbf{z})\|_1], \qquad (4)$$

where $\mathbf{x}$ denotes input images and $\mathcal{L}_{\mathrm{GAN}}$ is the typical GAN loss. $\mathcal{L}_{\mathrm{ori}}$ can be arbitrary complex objective function from any task, as shown in Figure 3 (b). We name the proposed method as Mode Seeking GANs (MSGANs).

4. Experiments

We evaluate the proposed regularization method through extensive quantitative and qualitative evaluation. We ap-
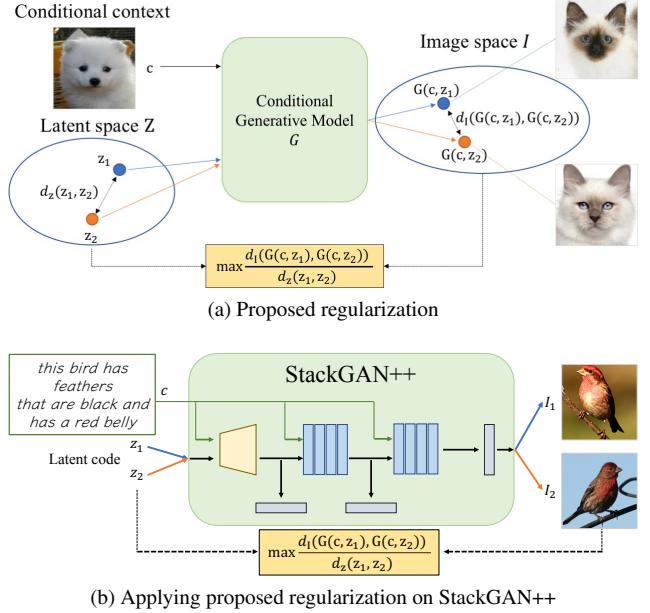

(a) Proposed regularization

(b) Applying proposed regularization on StackGAN++

Figure 3: **Proposed regularization.** (a) We propose a regularization term that maximizes the ratio of the distance between generated images with respect to the distance between their corresponding input latent codes. (b) The proposed regularization method can be applied to arbitrary cGANs. Take StackGAN++ [32], a model for text-to-image synthesis, as an example, we easily apply the regularization term regardless of the complex tree-like structure of the original model.

ply MSGANs to the baseline models from three representative conditional image synthesis tasks: categorical generation, image-to-image translation, and text-to-image synthesis. Note that we augment the original objective functions with the proposed regularization term while maintaining original network architectures and hyper-parameters. We employ L_1 norm distance as our distance metrics for both $d_{\mathbf{I}}$ and $d_{\mathbf{z}}$ and set the hyper-parameter $\lambda_{\mathrm{ms}} = 1$ in all experiments. More implementation and evaluation details, please refer to the appendixes.

4.1. Evaluation Metrics

We conduct evaluations using the following metrics. **FID.** To evaluate the quality of the generated images, we use FID [9] to measure the distance between the generated distribution and the real one through features extracted by Inception Network [28]. Lower FID values indicate better quality of the generated images.
LPIPS. To evaluate diversity, we employ LPIPS [33] following [10, 15, 35]. LIPIS measures the average feature distances between generated samples. Higher LPIPS score indicates better diversity among the generated images.

Table 1: **NDB and JSD results on the CIFAR-10 dataset.**

Metrics	Models	airplane	automobile	bird	cat	deer
NDB ↓	DCGAN	49.60 ± 3.43	53.00 ± 7.28	$\mathbf{34.40 \pm 6.11}$	46.00 ± 1.41	$\mathbf{44.80 \pm 3.90}$
	MSGAN	$\mathbf{46.60 \pm 7.40}$	$\mathbf{51.80 \pm 2.28}$	39.40 ± 1.95	$\mathbf{41.80 \pm 3.70}$	46.80 ± 4.92
JS ↓	DCGAN	0.034 ± 0.001	0.035 ± 0.002	$\mathbf{0.025 \pm 0.002}$	0.030 ± 0.002	$\mathbf{0.033 \pm 0.001}$
	MSGAN	$\mathbf{0.031 \pm 0.001}$	$\mathbf{0.033 \pm 0.001}$	0.027 ± 0.001	$\mathbf{0.027 \pm 0.001}$	0.035 ± 0.003
		dog	frog	horse	ship	truck
NDB ↓	DCGAN	50.40 ± 4.62	52.00 ± 3.81	54.40 ± 4.04	42.80 ± 5.45	47.80 ± 4.55
	MSGAN	$\mathbf{33.80 \pm 3.27}$	$\mathbf{42.00 \pm 2.92}$	$\mathbf{47.60 \pm 5.03}$	$\mathbf{41.00 \pm 2.92}$	$\mathbf{43.80 \pm 6.61}$
JS ↓	DCGAN	0.033 ± 0.001	0.034 ± 0.002	0.035 ± 0.001	0.029 ± 0.003	0.032 ± 0.001
	MSGAN	$\mathbf{0.024 \pm 0.001}$	$\mathbf{0.030 \pm 0.002}$	$\mathbf{0.033 \pm 0.003}$	$\mathbf{0.027 \pm 0.001}$	$\mathbf{0.029 \pm 0.003}$

NDB and JSD. To measure the similarity between the distribution between real images and generated one, we adopt two bin-based metrics, NDB and JSD, proposed in [24]. These metrics evaluate the extent of mode missing of generative models. Following [24], the training samples are first clustered using K-means into different bins which can be viewed as modes of the real data distribution. Then each generated sample is assigned to the bin of its nearest neighbor. We calculate the bin-proportions of the training samples and the synthesized samples to evaluate the difference between the generated distribution and the real data distribution. NDB score and JSD of the bin-proportion are then computed to measure the mode collapse. Lower NDB score and JSD mean the generated data distribution approaches the real data distribution better by fitting more modes. Please refer to [24] for more details.

4.2. Conditioned on Class Label

We first validate the proposed method on categorical generation. In categorical generation, networks take class labels as conditional contexts to synthesize images of different categories. We apply the regularization term to the baseline framework DCGAN [22].

We conduct experiments on the CIFAR-10 [13] dataset which includes images of ten categories. Since images in the CIFAR-10 dataset are of size 32×32 and upsampling degrades the image quality, we do not compute LPIPS in this task. Table 1 and Table 2 present the results of NDB, JS, and FID. MSGAN mitigates the mode collapse issue in most classes while maintaining image quality.

Table 2: **FID results on the CIFAR-10 dataset.**

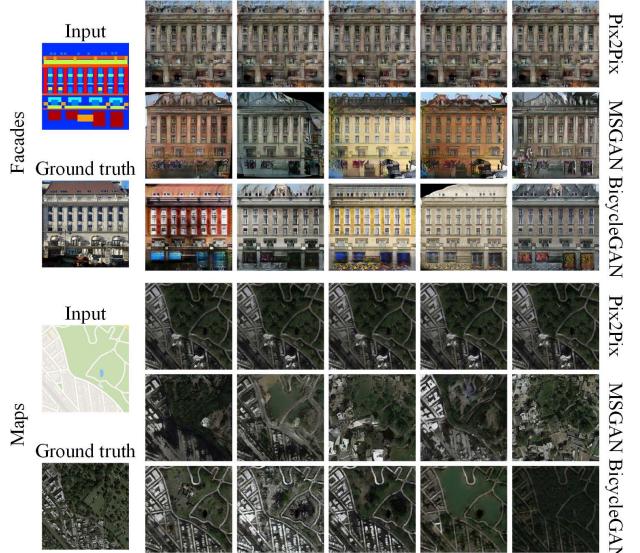| Model	DCGAN	MSGAN
FID↓	29.65 ± 0.06	$\mathbf{28.73 \pm 0.06}$

Figure 4: **Diversity comparison.** The proposed regularization term helps Pix2Pix learn more diverse results.

4.3. Conditioned on Image

Image-to-image translation aims to learn the mapping between two visual domains. Conditioned on images from the source domain, models attempt to synthesize corresponding images in the target domain. Despite the multimodal nature of the image-to-image translation task, early work [11, 34] abandons noise vectors and performs one-to-one mapping since the latent codes are easily ignored during training as shown in [11, 35]. To achieve multimodality, several recent attempts [10, 15, 35] introduce additional encoder networks and objective functions to impose a bijection constraint between the latent code space and the image space. To demonstrate the generalizability, we apply the proposed method to a unimodal model Pix2Pix [11] using paired training data and a multimodal model DRIT [15] using unpaired images.
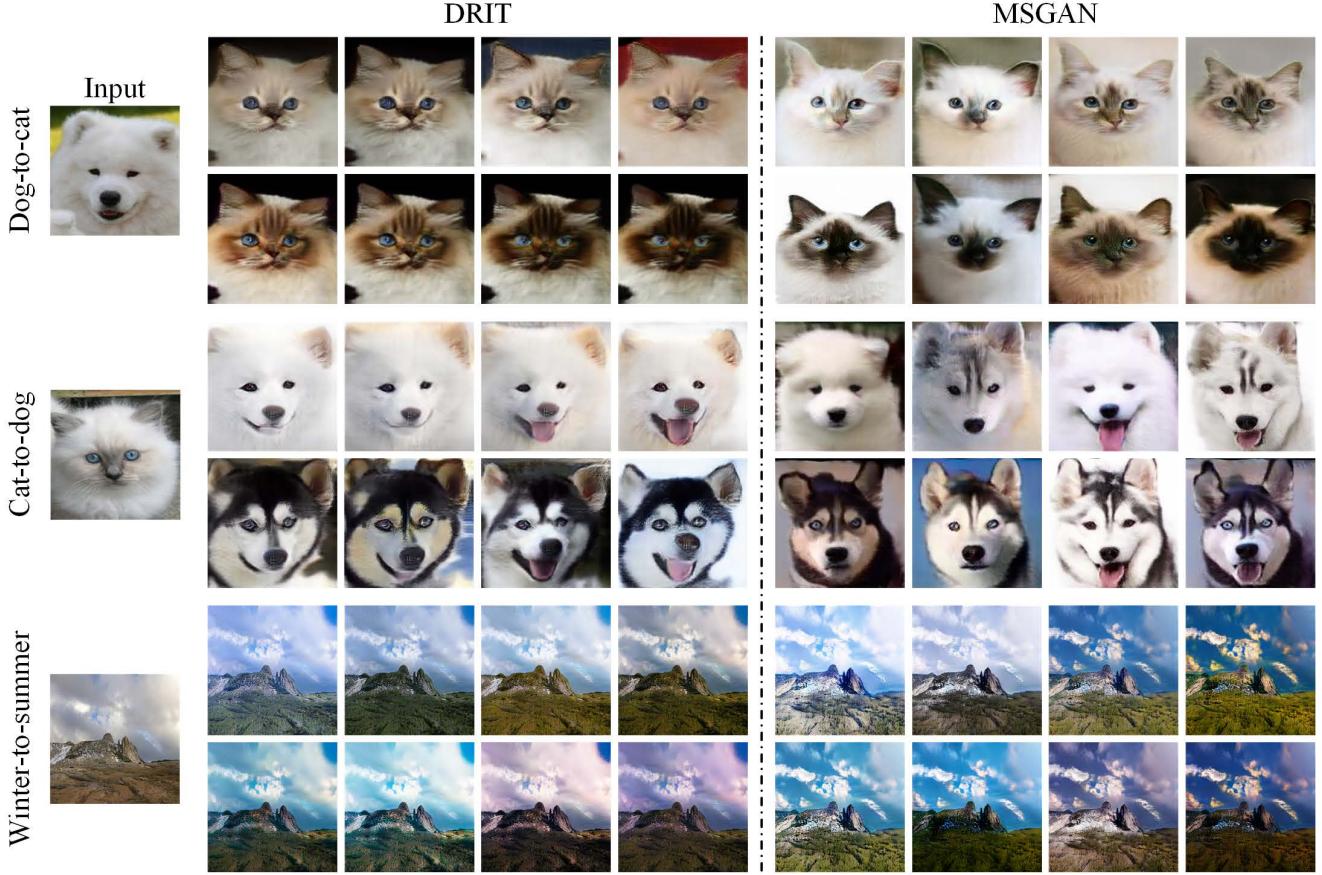
Figure 5: **Diversity comparison.** We compare MSGAN with DRIT on the dog-to-cat, cat-to-dog, and winter-to-summer translation tasks. Our model produces more diverse samples over DRIT.

Table 3: **Quantitative results on the facades and maps dataset.**

Datasets	Facades		
	Pix2Pix [11]	MSGAN	BicycleGAN [35]
FID ↓	139.19 ± 2.94	$\mathbf{92.84 \pm 1.00}$	98.85 ± 1.21
NDB↓	14.40 ± 1.82	$\mathbf{12.40 \pm 0.55}$	13.80 ± 0.45
JSD↓	0.074 ± 0.012	$\mathbf{0.038 \pm 0.004}$	0.058 ± 0.004
LPIPS↑	0.0003 ± 0.0000	$\mathbf{0.1894 \pm 0.0011}$	0.1413 ± 0.0005

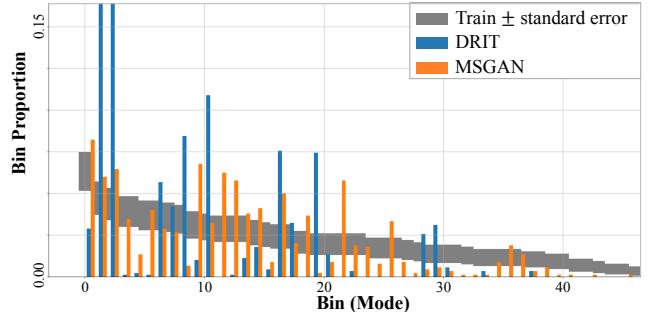| Datasets	Maps		
	Pix2Pix [11]	MSGAN	BicycleGAN [35]
FID ↓	168.99 ± 2.58	152.43 ± 2.52	$\mathbf{145.78 \pm 3.90}$
NDB↓	49.00 ± 1.00	$\mathbf{41.60 \pm 0.55}$	46.60 ± 1.34
JSD↓	0.088 ± 0.018	0.031 ± 0.003	$\mathbf{0.023 \pm 0.002}$
LPIPS↑	0.0016 ± 0.0003	$\mathbf{0.2189 \pm 0.0004}$	0.1150 ± 0.0007

Figure 6: **Visualization of the bins on dog→cat translation.** The translated results of DRIT collapse into few modes, while the generated image of MSGAN fit the real data distribution better.

4.3.1 Conditioned on Paired Images

We take Pix2Pix as the baseline model. We also compare MSGAN to BicycleGAN [35] which generates diverse im-ages with paired training images. For fair comparisons, architectures of the generator and the discriminator in all methods follow the ones in BicycleGAN [35].

We conduct experiments on the facades and maps datasets. MSGAN obtains consistent improvements on all
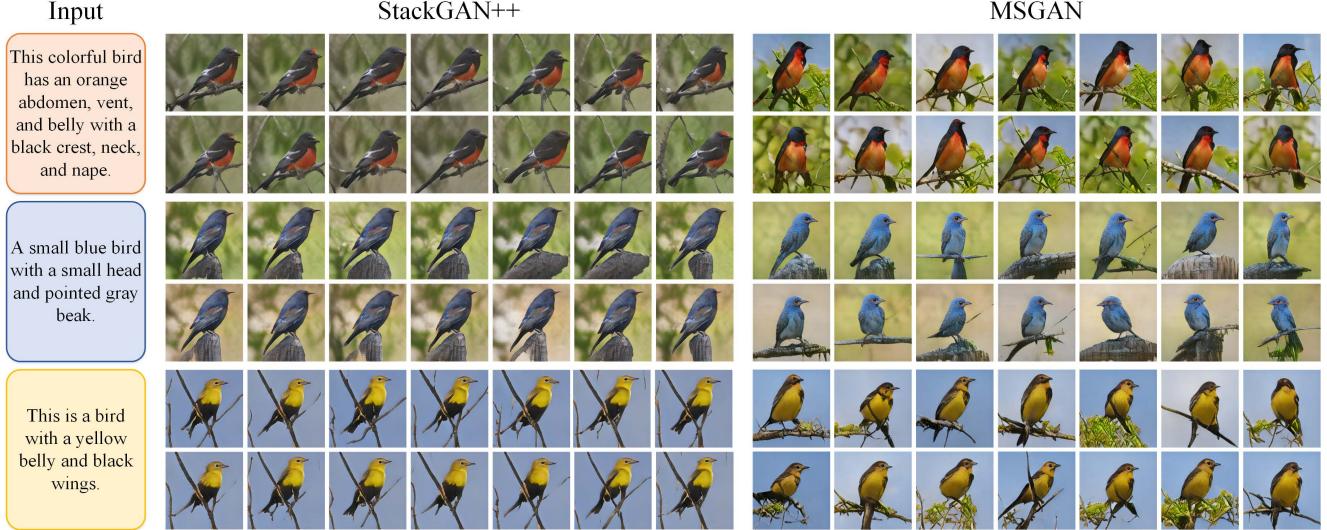
6

Figure 7: **Diversity comparison.** We show examples of StackGAN++ [32] and MSGAN on the CUB-200-2011 dataset of text-to-image synthesis. When the text code is fixed, the latent codes in MSGAN help to generate more diverse appearances and poses of birds as well as different backgrounds.

Table 4: **Quantitative results of the Yosemite (Summer⇌Winter) and the Cat⇌Dog dataset.**

Datasets	Summer2Winter		Winter2Summer	
	DRIT [15]	MSGAN	DRIT [15]	MSGAN
FID ↓	57.24 ± 2.03	$\mathbf{51.85 \pm 1.16}$	47.37 ± 3.25	$\mathbf{46.23 \pm 2.45}$
NDB↓	25.60 ± 1.14	$\mathbf{22.80 \pm 2.96}$	30.60 ± 2.97	$\mathbf{27.80 \pm 3.03}$
JSD↓	0.066 ± 0.005	$\mathbf{0.046 \pm 0.006}$	0.049 ± 0.009	$\mathbf{0.038 \pm 0.004}$
LPIPS↑	0.1150 ± 0.0003	$\mathbf{0.1468 \pm 0.0005}$	0.0965 ± 0.0004	$\mathbf{0.1183 \pm 0.0007}$
Datasets	Cat2Dog		Dog2Cat	
	DRIT [15]	MSGAN	DRIT [15]	MSGAN
FID↓	22.74 ± 0.28	$\mathbf{16.02 \pm 0.30}$	62.85 ± 0.21	$\mathbf{29.57 \pm 0.23}$
NDB↓	42.00 ± 2.12	$\mathbf{27.20 \pm 0.84}$	41.00 ± 0.71	$\mathbf{31.00 \pm 0.71}$
JSD↓	0.127 ± 0.003	$\mathbf{0.084 \pm 0.002}$	0.272 ± 0.002	$\mathbf{0.068 \pm 0.001}$
LPIPS↑	0.245 ± 0.002	$\mathbf{0.280 \pm 0.002}$	0.102 ± 0.001	$\mathbf{0.214 \pm 0.001}$

metrics over Pix2Pix. Moreover, MSGAN demonstrates comparable diversity to BicycleGAN, which applies an additional encoder network. Figure. 4 and Table. 3 demonstrate the qualitative and quantitative results, respectively.

4.3.2 Conditioned on Unpaired Images

We choose DRIT [15], one of the state-of-the-art frameworks to generate diverse images with unpaired training data, as the baseline framework. Though DRIT synthesizes diverse images in most cases, mode collapse occurs in some challenging shape-variation cases (*e.g.*, translation between cats and dogs). To demonstrate the robustness of the proposed method, we evaluate on the shape-preserving Yosemite (summer⇌winter) [34] dataset and the cat⇌dog [15] dataset that requires shape variations.

As the quantitative results exhibited in Table. 4, MSGAN performs favorably against DRIT in all metrics on both datasets. Especially in the challenging cat⇌dog dataset, MSGAN obtains substantial diversity gains. From the statistical point of view, we visualize the bin proportions of the dog-to-cat translation in Figure. 6. The graph shows the severe mode collapse issue of DRIT and the substantial improvement with the proposed regularization term. Qualitatively, Figure. 5 shows that MSGAN discovers more modes without the loss of visual quality.

Table 5: **Quantitative results on the CUB-200-2011 dataset.** We conduct experiments in two settings: 1) Conditioned on text descriptions, where every description can be mapped to different text codes. 2) Conditioned on text codes, where the text codes are fixed so that their effects are excluded.

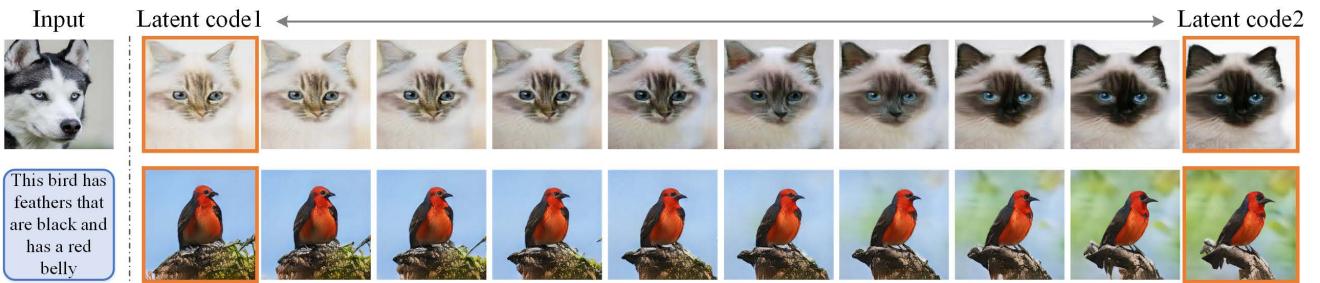|	Conditioned on text descriptions		Conditioned on text codes	
	StackGAN++ [32]	MSGAN	StackGAN++ [32]	MSGAN
FID ↓	25.99 ± 4.26	$\mathbf{25.53 \pm 1.83}$	$\mathbf{27.12 \pm 1.15}$	27.94 ± 3.10
NDB↓	38.20 ± 2.39	$\mathbf{30.60 \pm 2.51}$	39.00 ± 0.71	$\mathbf{30.60 \pm 2.41}$
JSD↓	0.092 ± 0.005	$\mathbf{0.073 \pm 0.003}$	0.102 ± 0.016	$\mathbf{0.095 \pm 0.016}$
LPIPS↑	0.362 ± 0.004	$\mathbf{0.373 \pm 0.007}$	0.156 ± 0.004	$\mathbf{0.207 \pm 0.005}$

Figure 8: **Linear interpolation between two latent codes in MSGAN.** Image synthesis results with linear-interpolation between two latent codes in the dog-to-cat translation and text-to-image synthesis.

4.4. Conditioned on Text

Text-to-image synthesis targets at generating images conditioned on text descriptions. We integrate the proposed regularization term on StackGAN++ [32] using the CUB-200-2011 [29] dataset. To improve diversity, StackGAN++ introduces a Conditioning Augmentation (CA) module that re-parameterizes text descriptions into text codes of the Gaussian distribution. Instead of applying the regularization term on the semantically meaningful text codes, we focus on exploiting the latent codes randomly sampled from the prior distribution. However, for a fair comparison, we evaluation MSGAN against StackGAN++ in two settings: 1) Perform generation without fixing text codes for text descriptions. In this case, text codes also provide variations for output images. 2) Perform generation with fixed text codes. In this setting, the effects of text codes are excluded.

Table. 5 presents quantitative comparisons between MS-GAN and StackGAN++. MSGAN improves the diversity of StackGAN++ and maintains visual quality. To better illustrate the role that latent codes play for the diversity, we show qualitative comparisons with the text codes fixed. In this setting, we do not consider the diversity resulting from CA. Figure. 7 illustrates that latent codes of StackGAN++ have minor effects on the variations of the image. On the contrary, latent codes of MSGAN contribute to various appearances and poses of birds.

4.5. Interpolation of Latent Space in MSGANs

We perform linear interpolation between two given latent codes and generate corresponding images to have a better understanding of how well MSGANs exploit the latent space. Figure. 8 shows the interpolation results on the dog-to-cat translation and the text-to-image synthesis task. In the dog-to-cat translation, we can see the coat colors and patterns varies smoothly along with the latent vectors. In the text-to-image synthesis, both orientations of birds and the appearances of footholds change gradually with the variations of the latent codes.

5. Conclusions

In this work, we present a simple but effective mode seeking regularization term on the generator to address the model collapse in cGANs. By maximizing the distance between generated images with respect to that between the corresponding latent codes, the regularization term forces the generators to explore more minor modes. The proposed regularization method can be readily integrated with existing cGANs framework without imposing training overheads and modifications of network structures. We demonstrate the generalizability of the proposed method on three different conditional generation tasks including categorical generation, image-to-image translation, and text-to-image synthesis. Both qualitative and quantitative results show that the proposed regularization term facilitates the baseline frameworks improving the diversity without sacrificing visual quality of the generated images.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 2, 3

[2] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. Mode regularized generative adversarial networks. In *ICLR*, 2017. 2, 3

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 12

[4] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. In *ICLR*, 2017. 2

[5] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. In *ICLR*, 2017. 2

[6] I. Durugkar, I. Gemp, and S. Mahadevan. Generative multi-adversarial networks. In *ICLR*, 2017. 2, 3

[7] A. Ghosh, V. Kulharia, V. Namboodiri, P. H. Torr, and P. K. Dokania. Multi-agent diverse generative adversarial networks. In *CVPR*, 2018. 2, 3

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 2, 4

[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 2, 3, 4, 5

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 4, 5, 6, 10, 11, 12

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 10

[13] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2, 5, 10, 12

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[15] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 1, 2, 3, 4, 5, 7, 10, 12

[16] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016. 1, 2

[17] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 1, 2

[18] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 2, 3

[19] X. Mao, Q. Li, H. Xie, R. YK, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 2, 3

[20] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. In *ICLR*, 2017. 2, 3

[21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[22] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2, 5, 10, 12

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 1, 2

[24] E. Richardson and Y. Weiss. On GANs and GMMs. In *NIPS*, 2018. 2, 5, 10

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. 2015. 10

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, 2016. 1, 2, 3

[27] A. Srivastava, L. Valkoz, C. Russell, M. U. Gutmann, and C. Sutton. VEEGAN: Reducing mode collapse in GANs using implicit variational learning. In *NIPS*, 2017. 2, 3

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 4

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 8, 12

[30] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 1, 11

[31] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee. Diversity-sensitive conditional generative adversarial networks. In *ICLR*, 2019. 3

[32] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 2018. 1, 2, 4, 7, 8, 10, 12

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 4

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2, 5, 7, 10, 12

[35] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017. 1, 2, 3, 4, 5, 6, 11

Appendix A. Implementation Details

Table 9 summarizes the datasets and baseline models used on various tasks. For all of the baseline methods, we incorporate the original objective functions with the proposed regularization term. Note that we remain the original network architecture design and use the default setting of hyper-parameters for the training.

DCGAN. Since the images in the CIFAR-10 [13] dataset are of size 32×32, we modify the structure of the generator and discriminator in DCGAN [22], as shown in Table 10. We use the batch size of 32, learning rate of 0.0002 and Adam [12] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to train both the baseline and MSGAN network.

Pix2Pix. We adopt the generator and discriminator in BicycleGAN [34] to build the Pix2Pix [11] model. Same as BicycleGAN, we use a U-Net network [25] for the generator, and inject the latent codes **z** into every layer of the generator. The architecture of the discriminator is a two-scale PatchGAN network [11]. For the training, both Pix2Pix and MSGAN framework use the same hyper-parameters as the officially released version [1].

DRIT. DRIT [15] involves two stages of image-to-image translations in the training process. We only apply the mode seeking regularization term to generators in the first stage, which is modified on the officially released code [2].

StackGAN++. StackGAN++ [32] is a tree-like structure with multiple generators and discriminators. We use the output images from the last generator and input latent codes to calculate the mode seeking regularization term. The implementation is based on the officially released code [3].

Appendix B. Evaluation Details

We employ the official implementation of FID [4], NDB and JSD [5], and LPIPS [6]. For NDB and JSD, we use the K-means method on training samples to obtain the clusters. Then the generated samples are assigned to the nearest cluster to compute the bin proportions. As suggested by the author of [24], there are at least 10 training samples for each cluster. Therefore, we cluster the number of bins $K \approx N_{\text{train}}/20$ in all tasks, where N_{train} denotes the number of training samples for computing the clusters. We have verified that the performance is consistent within a large range of K. For evaluation, we randomly generate

N images for a given conditional context on various tasks. We conduct five independent trials and report the mean and standard derivation based on the result of each trial. More evaluation details of one trial are presented as follows.

Conditioned on Class Label. We randomly generate $N = 5000$ images for each class label. We use all the training samples and the generated samples to compute FID. For NDB and JSD, we employ the training samples in each class to calculate $K = 250$ clusters.

Conditioned on Image. We randomly generate $N = 50$ images for each input image in the test set. For LPIPS, we randomly select 50 pairs of the 50 images of each context in the test set to compute LPIPS and average all the values for this trial. Then, we randomly choose 100 input images and their corresponding generated images to form 5000 generated samples. We use the 5000 generated samples and all samples in training set to compute FID. For NDB and JSD, we employ all the training samples for clustering and choose $K = 20$ bins for facades, and $K = 50$ bins for other datasets.

Conditioned on Text. We randomly select 200 sentences and generate $N = 10$ images for each sentence, which forms 2000 generated samples. Then, we randomly select $N_{\text{train}} = 2000$ samples for computing FID, and clustering them into $K = 100$ bins for NDB and JSD. For LPIPS, we randomly choose 10 pairs for each sentence and average the values of all the pairs for this trial.

Appendix C. Ablation Study on the Regularization Term

C.1. The Weighting Parameter λ_{ms}

To analyze the influence of the regularization term, we conduct an ablation study by varying the weighting parameter λ_{ms} on image-to-image translation task using the facades dataset. Table 6 presents the quantitative results with diverse λ_{ms}. It can be observed that increasing λ_{ms} improves the diversity of the generated images. Nevertheless, as the weighting parameter λ_{ms} becomes larger than a threshold value (1.0), the training becomes unstable, which yields low quality, and even low diversity synthesized images. As a result, we empirically set the weighting parameter $\lambda_{\text{ms}} = 1.0$ for all experiments.

Table 6: **Quantitative results with different λ_{ms} on the facades dataset.**

λ_{ms}	0	0.1	0.5	1	3
FID ↓	134.95	126.31	88.41	92.45	338.45
LPIPS↑	0.0003	0.0155	0.0929	**0.1888**	0.1393

---

[1] https://github.com/junyanz/BicycleGAN/
[2] https://github.com/HsinYingLee/DRIT
[3] https://github.com/hanzhanggit/StackGAN-v2
[4] https://github.com/bioinf-jku/TTUR
[5] https://github.com/eitanrich/gans-n-gmms
[6] https://github.com/richzhang/PerceptualSimilarity

C.2. The Design Choice of the Distance Metric

We have explored other design choice of the distance metric. We conduct experiments using discriminator feature distance in our regularization term in a way similar to feature matching loss [30],

$$\mathcal{L}_{ms} = \frac{\frac{1}{L}\sum_{l=1}^{L}\|D^l(G(\mathbf{c}, \mathbf{z}_2)) - D^l(G(\mathbf{c}, \mathbf{z}_1))\|_1}{\|\mathbf{z}_2 - \mathbf{z}_1\|_1}, \quad (5)$$

where D^l denotes the lth layer of the discriminator. We apply it to Pix2Pix on the facades dataset. Table. 7 shows that MSGAN using feature distance also obtains improvement over Pix2Pix. However, MSGAN using L_1 distance has higher diversity. Therefore, we employ MSGAN using L_1 distance for all experiments.

Table 7: **Quantitative results on the facades dataset.**

	Pix2Pix [11]	MSGAN-L_1	MSGAN-FD
FID $\downarrow$	139.19 ± 2.94	$\mathbf{92.84 \pm 1.00}$	100.16 ± 3.14
NDB $\downarrow$	14.40 ± 1.82	12.40 ± 0.55	$\mathbf{11.80 \pm 1.48}$
JSD $\downarrow$	0.074 ± 0.012	$\mathbf{0.038 \pm 0.004}$	0.072 ± 0.014
LPIPS $\uparrow$	0.0003 ± 0.0000	$\mathbf{0.1894 \pm 0.0011}$	0.0565 ± 0.0003

Appendix D. Computational Overheads

We compare MSGAN with Pix2Pix, BicycleGAN in terms of training time, memory consumption, and model parameters on an NVIDIA TITAN X GPU. Table. 8 shows that our method incurs marginal overheads. However, BicycleGAN requires longer time per iteration and larger memory with an additional encoder and another discriminator network.

Table 8: **Comparisons of computational overheads on the facades dataset.**

Model	Time (s)	Memory (MB)	Parameters (M)
Pix2Pix [11]	0.122	1738	58.254
MSGAN	0.122	1739	58.254
BicycleGAN [35]	0.192	2083	64.303

Appendix E. Additional Results

We present more results of categorical generation, image-to-image translation, and text-to-image synthesis in Figure 10, Figure 11, Figure 12, Figure 13, Figure 14, and Figure 15, respectively.

Table 9: **Statistics of different generation tasks.** We summarize the number of training and testing images in each generation task. The baseline model used for each task is also provided.

Context	Class Label		Paired Images				Unpaired Images								Text	
Dataset	CIFAR-10 [13]		Facades [3]		Maps [11]		Yosemite [34]				Cat ⇌ Dog [15]				CUB-200-2011 [29]	
							Summer		Winter		Cat		Dog			
	train	test	train	test	train	test	train	test	train	test	train	test	train	test	train	test
Samples	50,000	10,000	400	206	1,096	1,098	1,069	309	770	238	771	100	1,264	100	8,855	2,933
Baseline	DCGAN [22]		Pix2Pix [11]				DRIT [15]								StackGAN++ [32]	

Table 10: **The architecture of the generator and discriminator of DCGAN.** We employ the following abbreviation: N= Number of filters, K= Kernel size, S= Stride size, P= Padding size. "Conv", "Dconv","BN" denote the convolutional layer, transposed convolutional layer and batch normalization, respectively.

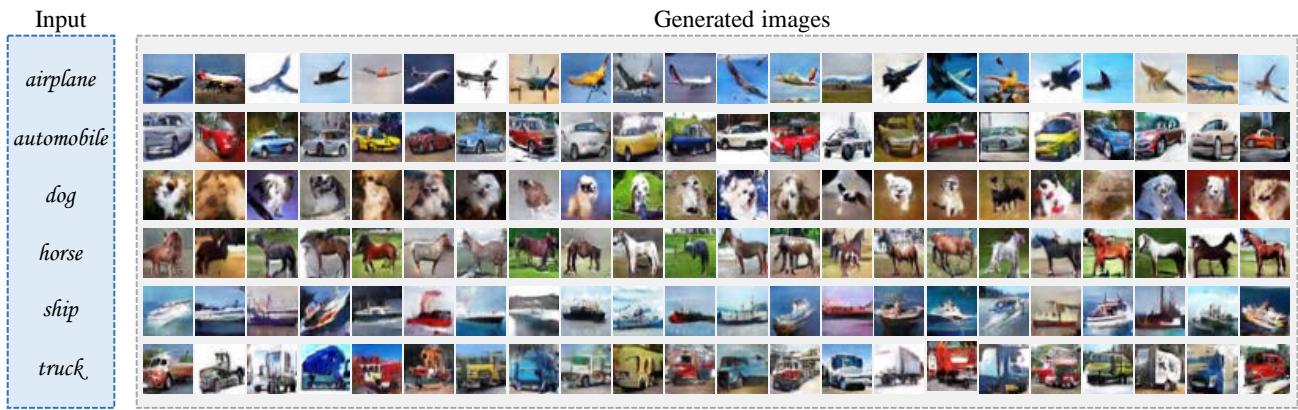| Layer	Generator	Discriminator
1	Dconv(N512-K4-S1-P0), BN, Relu	Conv(N128-K4-S2-P1), Leaky-Relu
2	Dconv(N256-K4-S2-P1), BN, Relu	Conv(N256-K4-S2-P1), BN, Leaky-Relu
3	Dconv(N128-K4-S2-P1), BN, Relu	Conv(N512-K4-S2-P1), BN, Leaky-Relu
4	Dconv(N3-K4-S2-P1), Tanh	Conv(N1-K4-S1-P0), Sigmoid

Figure 9: **More categorical generation results of CIFAR-10.** We show the results of DCGAN [22] with the proposed mode seeking regularization term on categorical generation task.
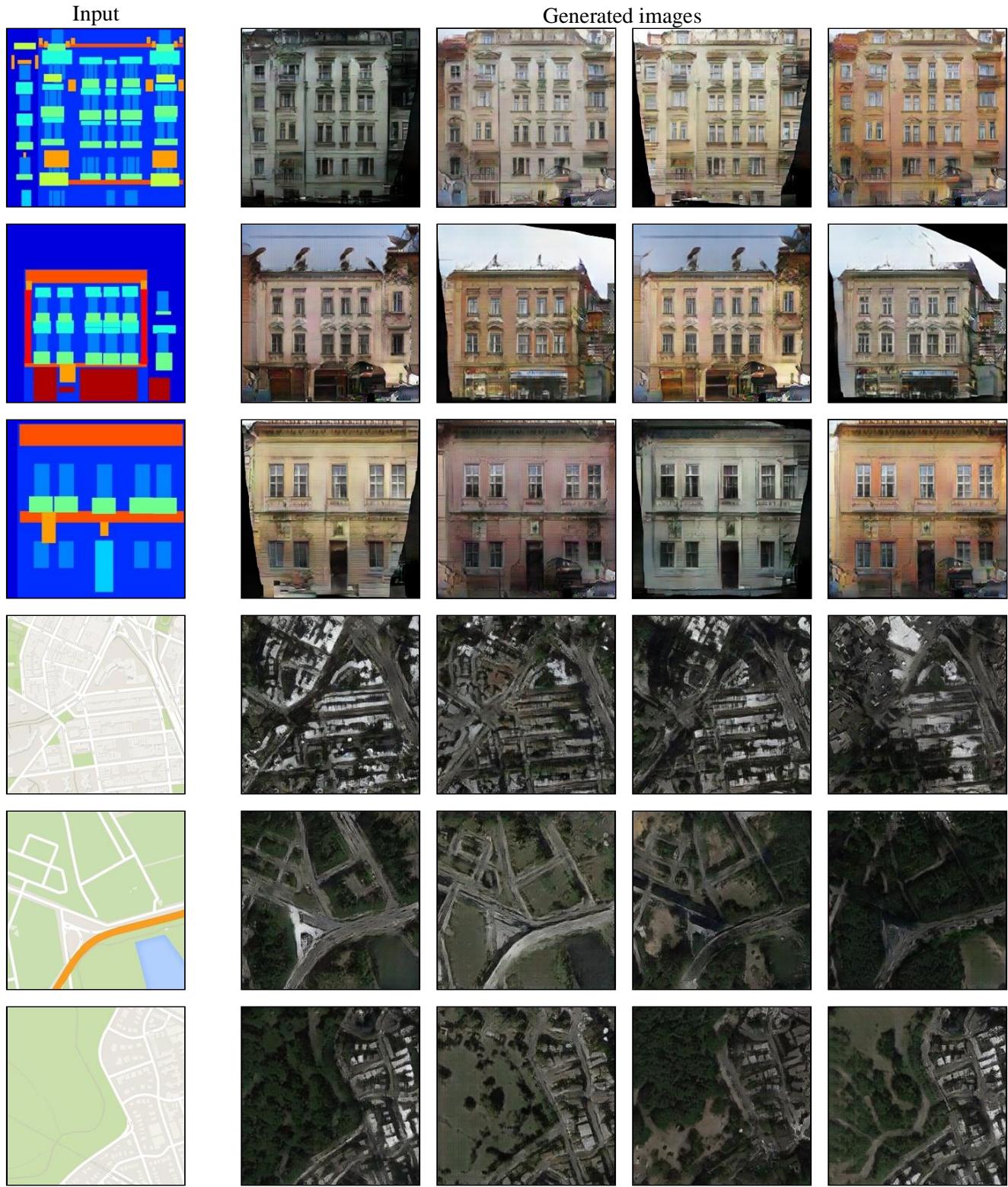
Figure 10: **More image-to-image translation results of facades and maps.** Top three rows: facades, bottom three rows: maps.
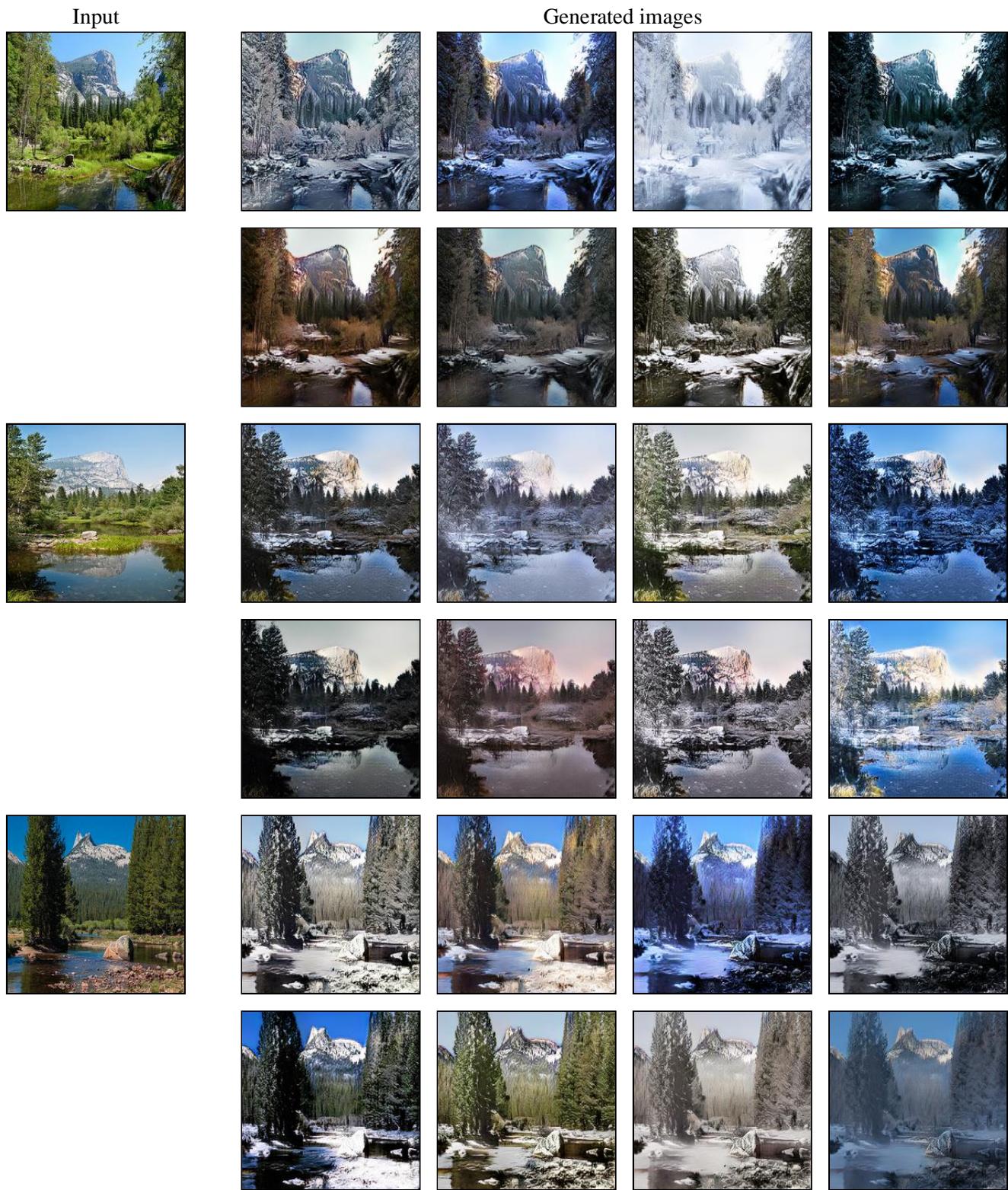
Input   Generated images

Figure 11: **More image-to-image translation results of Yosemite, Summer→Winter.**

Input

Generated images
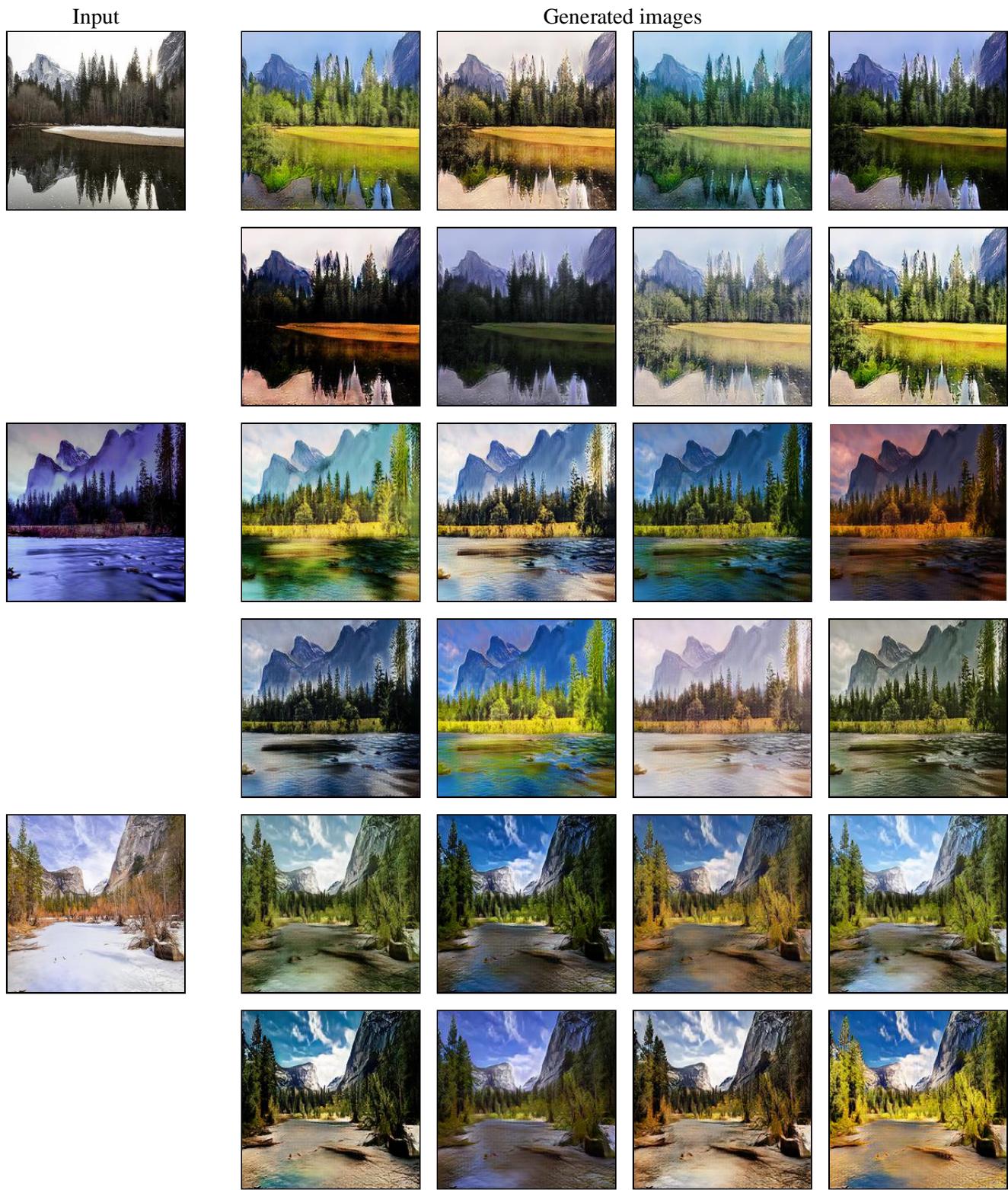
Figure 12: **More image-to-image translation results of Yosemite, Winter→Summer.**

15

Input

Generated images
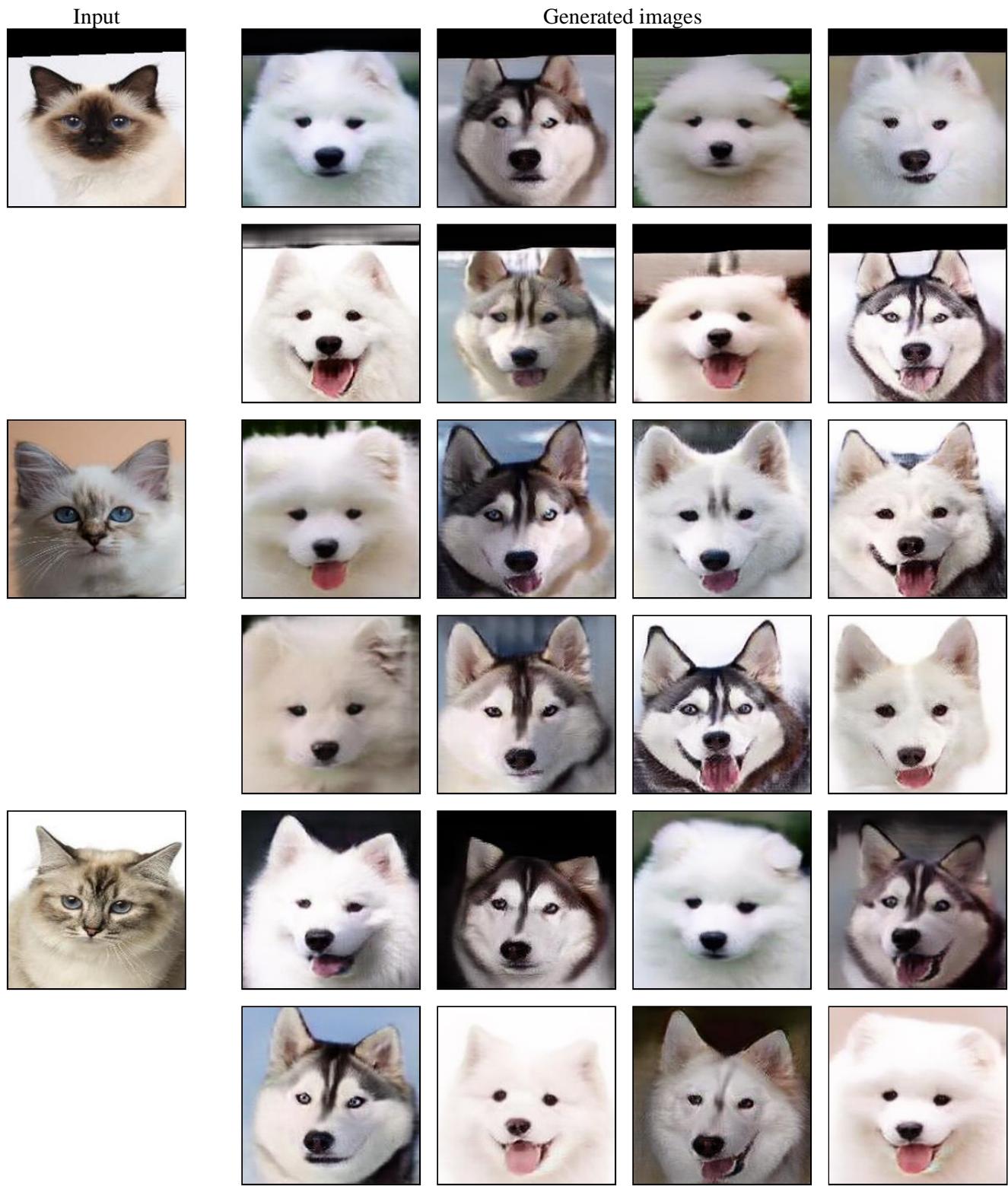
Figure 13: **More image-to-image translation results of Cat→Dog.**

Input Generated images
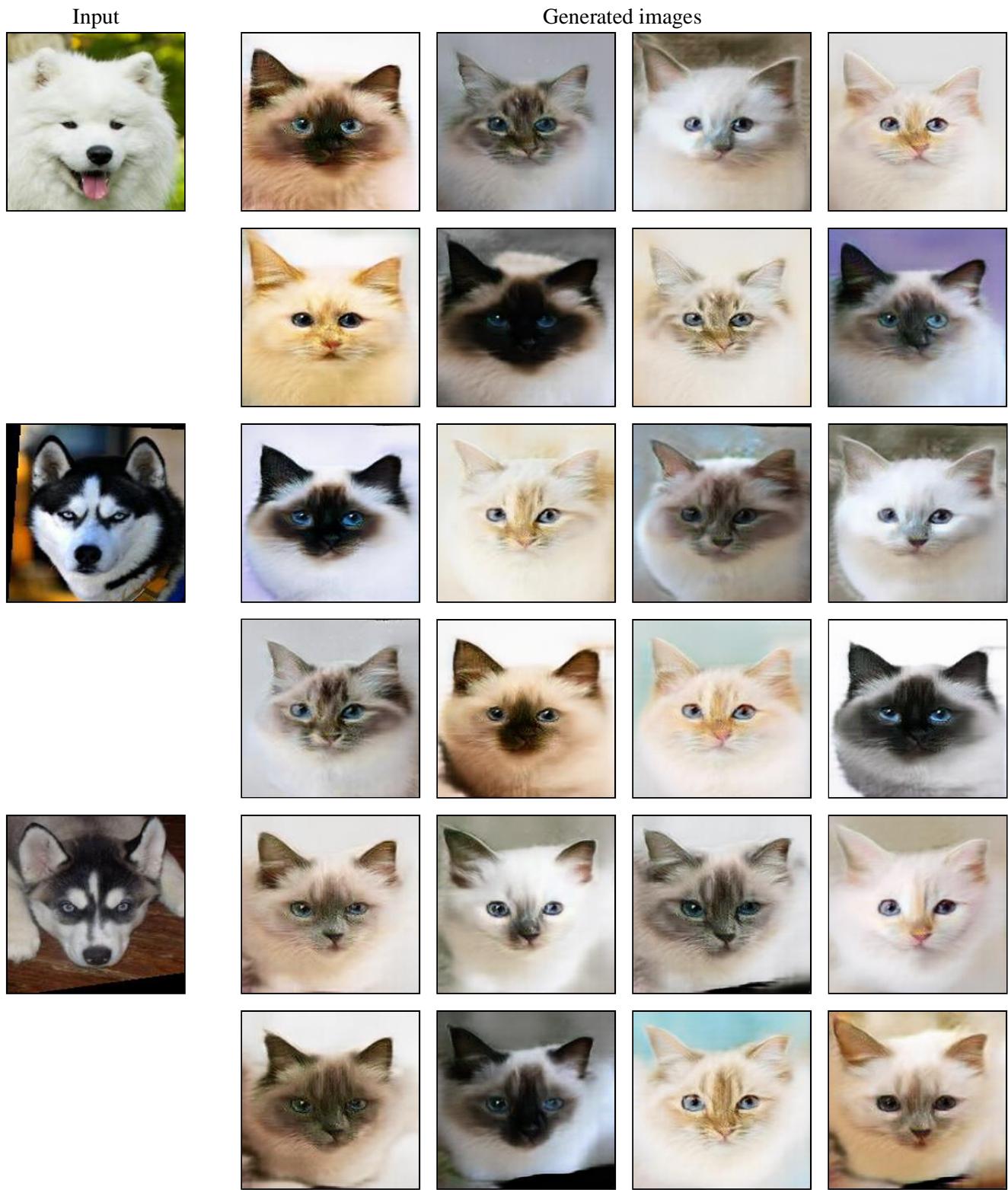
Figure 14: **More image-to-image translation results of Dog→Cat.**

Input

Generated images

This chubby bird has a small fat bill and a red belly.

A small bird with a blue nape, white breast and body with a short bill.

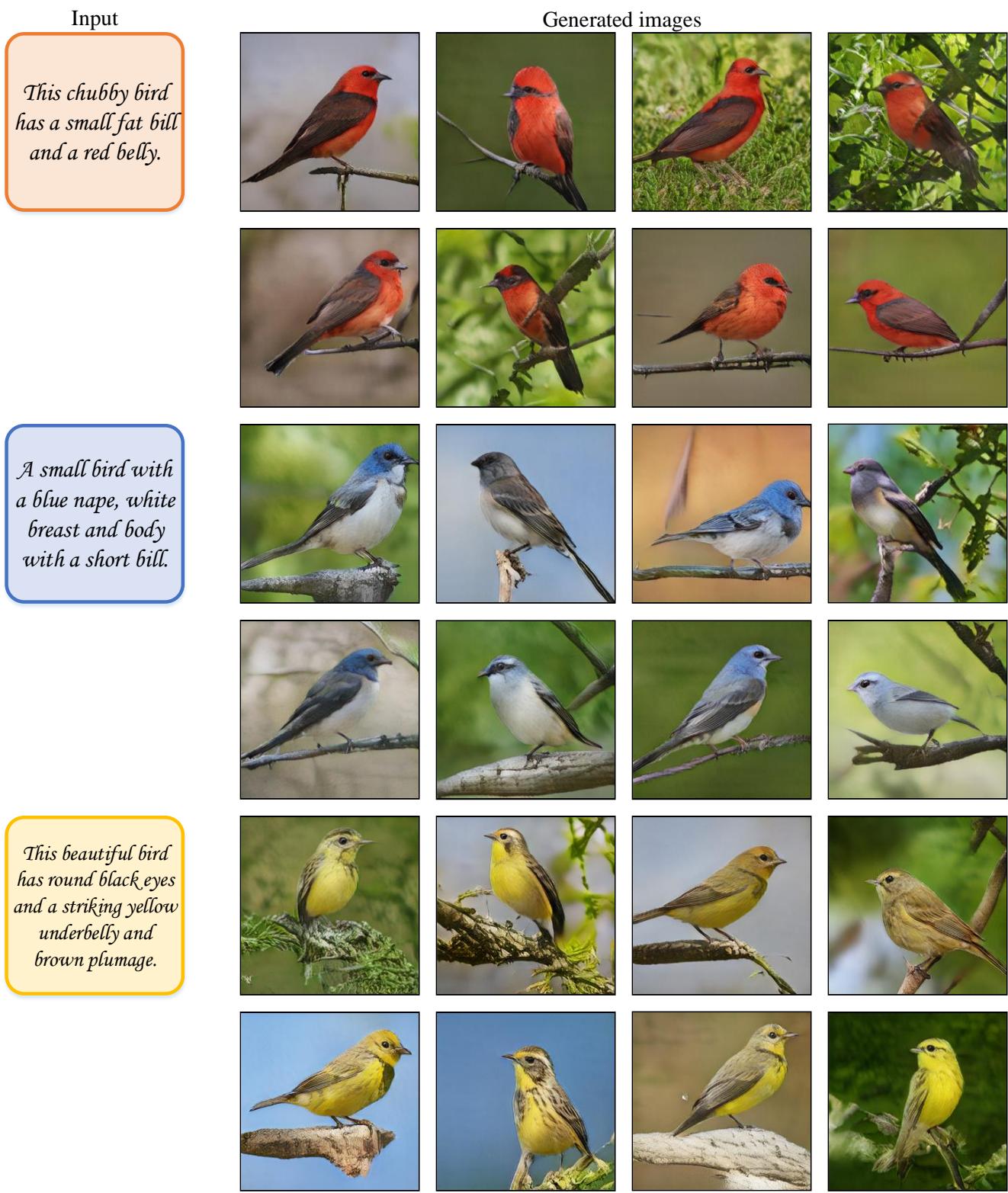*This beautiful bird has round black eyes and a striking yellow underbelly and brown plumage.*

Figure 15: **More text-to-image synthesis results of CUB-200-2011.**