Dual Adversarial Inference for Text-to-Image Synthesis

Qicheng Lao[1 2] Mohammad Havaei[1] Ahmad Pesaranghader[1 3] Francis Dutil[1]
Lisa Di Jorio[1] Thomas Fevens[2]
[1]Imagia Inc. [2]Concordia University [3]Dalhousie University
{qi_lao, fevens}@encs.concodia.ca, {mohammad, ahmad.pgh, francis.dutil, lisa}@imagia.com

Abstract

*Synthesizing images from a given text description involves engaging two types of information: the content, which includes information explicitly described in the text (*e.g.*, color, composition, etc.), and the style, which is usually not well described in the text (*e.g.*, location, quantity, size, etc.). However, in previous works, it is typically treated as a process of generating images only from the content,* i.e.*, without considering learning meaningful style representations. In this paper, we aim to learn two variables that are disentangled in the latent space, representing content and style respectively. We achieve this by augmenting current text-to-image synthesis frameworks with a dual adversarial inference mechanism. Through extensive experiments, we show that our model learns, in an unsupervised manner, style representations corresponding to certain meaningful information present in the image that are not well described in the text. The new framework also improves the quality of synthesized images when evaluated on Oxford-102, CUB and COCO datasets.*

1. Introduction

The problem of text-to-image synthesis is to generate diverse yet plausible images given a text description of the image and a general data distribution of images and matching descriptions. In recent years, generative adversarial networks (GANs) [9] have asserted themselves as perhaps the most effective architecture for image generation, along with their variant Conditional GANs [22], wherein the generator is conditioned on a vector encompassing some desired property of the generated image.

A common approach for text-to-image synthesis is to use a pre-trained text encoder to produce a text embedding from the description. This vector is used as the conditioning factor in a conditional GAN-based model. The very first GAN model for the text-to-image synthesis task [26] uses a noise vector sampled from a normal distribution to capture image style features left out of the text representation, enabling the model to generate a variety of images given a certain textual description. StackGan [32] introduces conditioning augmentation as a way to augment the text embeddings, where a text embedding can be sampled from a learned distribution representing the text embedding space. As a result, current state-of-the-art methods for text-to-image synthesis generally have two sources of randomness: one for the text embedding variability, and the other (noise z given a normal distribution) capturing image variability.

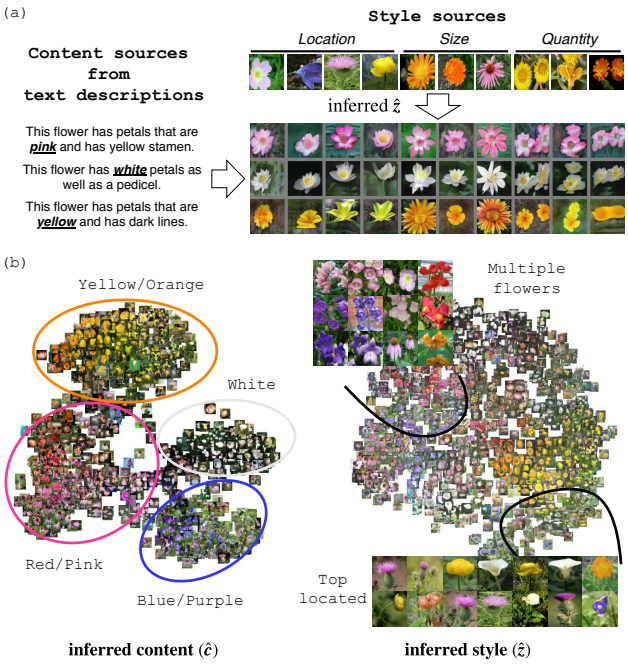

Figure 1: (a) Controlling the style (in columns) of generated images given a text description as the content (in rows). Columns 1-4 show locations (*e.g.*, left, right and top) of the content in the image; Columns 5-7 and columns 8-10 represent size and quantity of the content respectively. (b) The learned content and style features through our dual adversarial inference, visualized by t-SNE. The inferred content is clustered solely on color (one dominant factor that is described in the text), while the inferred style shows a more diffused cluster pattern, with local clusters such as multiple flowers and top-located flowers.

Having two sources of randomness is, however, only meaningful if they represent different factors of variation. Problematically, our empirical investigation of some previously published methods reveals that those two sources can overlap: due to the randomness in the text embedding, the noise vector $\boldsymbol{z}$ then does not meaningfully contribute to the variability nor the quality of generated images, and can be discarded. This is illustrated in Figure 8 and Figure 9 in the supplementary material.

In this paper we aim to learn a latent space that represents meaningful information in the context of text-to-image synthesis. To do this, we incorporate an inference mechanism that encourages the latent space to learn the distribution of the data. To capture different factors of variation, we construct the latent space through two independent random variables, representing content ('$\boldsymbol{c}$') and style ('$\boldsymbol{z}$'). Similar to previous work [26], '$\boldsymbol{c}$' encodes image content which is the information in the text description. This mostly includes color, composition, etc. On the other hand, '$\boldsymbol{z}$' encodes style which we define as all other information in the image data that is not well described in the text. This would typically include location, size, pose, and quantity of the content in the image, background, etc. This new framework allows us to better represent information found in both text and image modalities, achieving better results on Oxford-102 [23], CUB [29] and COCO [20] datasets at 64×64 resolution.

The main goal of this paper is to learn disentangled representations of style and content through an inference mechanism for text-to-image synthesis. This allows us to use not only the content information described in the text descriptions but also the desired styles when generating images. To that end, we only focus on the generation of low-resolution images (*i.e.*, 64×64). In the literature, high-resolution images are generally produced by iterative refinement of lower-resolution images and thus we consider it a different task, more closely related to generating super-resolution images.

To the best of our knowledge, this is the first time an attempt has been made to explicitly separate the learning of style and content for text-to-image synthesis. We believe that capturing these subtleties is important to learn richer representations of the data. As shown in Figure 1, by learning disentangled representations of content and style, we can generate images that respect the content information from a text source while controlling style by inferring the style information from a style source. It is worth noting that although we hope to learn the style from the image modality, the style information could possibly be connected to (or leaked into) some text instances. Despite this, the integration of the style in the model eventually depends on how well it is represented in both modalities. For example, if certain types of style information are commonly present in the text, then according to our definition, those types of information are considered as content. If only a few text instances describe that information however, then it would not be fully representative of a shared commonality among texts and therefore would not be captured as content, and whether it can be captured as style depends on how well it is represented in the image modality. On the other hand, we would also like to explore modalities other than *text* as the content in our future work using the proposed method, which may bring us closer to image-to-image translation [18] if we choose both modalities to be *image*.

The contributions of this paper are twofold: (i) we are the first to learn two variables that are disentangled for content and style in the context of text-to-image synthesis using inference; and (ii) by incorporating inference we improve on the state-of-the-art in image quality while maintaining comparable variability and visual-semantic similarity when evaluated on the Oxford-102, CUB and COCO datasets.

2. Related Work

Text-to-image synthesis methods Text-to-image synthesis has been made possible by Reed *et al*. [26], where a conditional GAN-based model is used to generate text-matching images from the text description. Zhang *et al*. [32] use a two-stage GAN to first generate low-resolution images in stage I and then improve the image quality to high-resolution in stage II. By using a hierarchically-nested GAN (HDGAN) which incorporates multiple loss functions at increasing levels of resolution, Zhang *et al*. [35] further improve the state-of-the-art on this task in an end-to-end manner. Several attempts have been made to leverage additional available information, such as object location [27], class label [5, 2], attention extracted from word features [30, 24] and text regeneration [24]. Hong *et al*. [12] propose another approach by providing the image generator with a semantic structure that is sequentially constructed with a box generator followed by a shape generator; however, their approach would not be applicable for single-object image synthesis. Compared to all previous work, our method incorporates the inference mechanism into the current framework for text-to-image synthesis, and by doing so, we explicitly force the model to simultaneously learn separate representations of content and style. Reed *et al*. [26] have also investigated the separation of content and style information. The differences are elaborated in the supplementary material.

Adversarial inference methods Various papers have explored learning representations through adversarial training. Notable mentions are BiGANs [6, 7] where a bidirectional discriminator acts on pairs $(\boldsymbol{x}, \boldsymbol{z})$ of data and generated points. While these models assume that a single random variable $\boldsymbol{z}$ encodes data representations, in this work we extend the adversarial inference to two random variables that

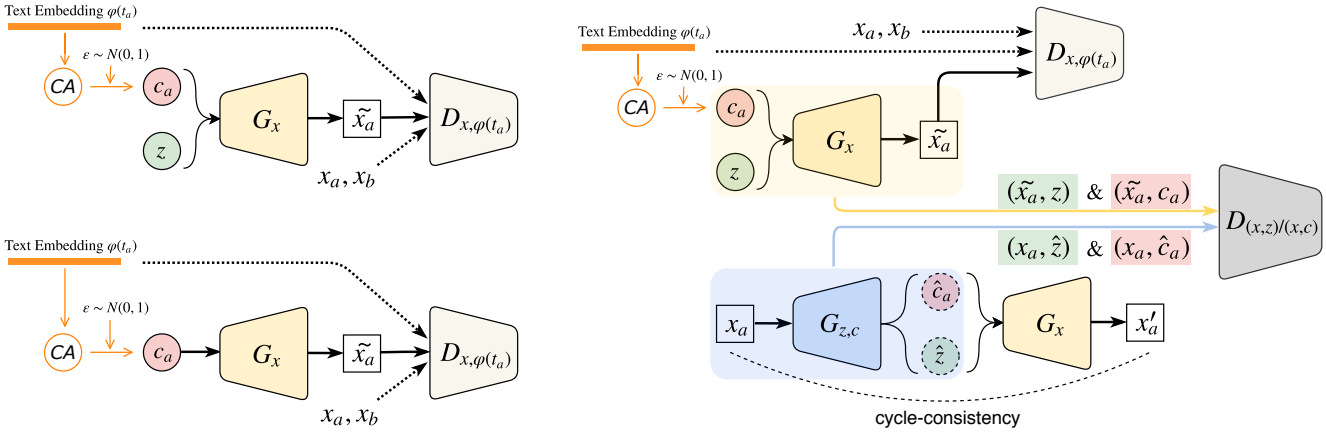

Figure 2: Overview of the current state-of-the-art methods (left top) and our proposed method (right) for text-to-image synthesis at low-resolution scale. By default, the current state-of-the-art methods adopt *conditioning augmentation* (CA), which introduces variable $\boldsymbol{c} \sim p(\boldsymbol{c}|\varphi_t)$, in addition to variable $\boldsymbol{z} \sim \mathcal{N}(0,1)$ as the inputs for the image generator G_x. The removal of $\boldsymbol{z}$ (left bottom) does not affect the model performance (*viz.* Figure 9 in supplementary material for quantitative evaluations). In our method (right), we incorporate the inference mechanism, where $G_{z,c}$ encodes both $\boldsymbol{z}$ and $\boldsymbol{c}$, and the discriminator $D_{(x,z)/(x,c)}$ distinguishes between joint pairs. For the cycle consistency, sampled $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{c}}$ are also used to reconstruct $\boldsymbol{x}'$.

are disentangled with each other. Our model is also closely related to [19], where the authors incorporate an adversarial reconstruction loss into the BiGAN framework. They show that the additional loss term results in better reconstructions and more stable training. Although Dumoulin *et al.* [7] show results for conditional image generation, in their model the conditioning factor is discrete, fully observed and not inferred through the inference model. In our model however, '$\boldsymbol{c}$' can be a continuous conditioning variable that we infer from the text and image.

Relation to InfoGAN While the matching-aware loss (Section 3.1) used in many text-to-image works can also be viewed as maximizing mutual information between the two modalities (*i.e.*, *text* and *image*), the way it is approximated is different. InfoGAN [3] uses the variational mutual information maximization technique, whereas the matching-aware loss uses the concept of matched and mismatched pairs. In addition, InfoGAN concentrates all semantic features on the latent code $\boldsymbol{c}$, which contains both content and style, whereas in this work, we only maximize mutual information on the content since we consider *text* as our content.

3. Methods

3.1. Preliminaries

We start by describing text-to-image synthesis. Let φ_t be the text embedding of a given text description associated with image $\boldsymbol{x}$. The goal of text-to-image synthesis is to generate a variety of visually-plausible images that are text-matched. Reed *et al.* [26] first propose a conditional GAN-based framework, where a generator G_x takes as input a noise vector $\boldsymbol{z}$ sampled from $p(\boldsymbol{z}) = \mathcal{N}(0,1)$ and φ_t as the conditioning factor to generate an image $\tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}, \varphi_t)$. A *matching-aware* discriminator D_{x,φ_t} is then trained to not only judge between real and fake images, but also discriminate between matched and mismatched image-text pairs. The minimax objective function for text-to-image (subscript denoted as $t2i$) framework is given as:

$$\begin{aligned}\min_G \max_D V_{t2i}(D_{x,\varphi_t}, G_x) = \\ \mathbb{E}_{(\boldsymbol{x}_a,t_a)\sim p_{\text{data}}}[\log D_{x,\varphi_t}(\boldsymbol{x}_a, \varphi_{t_a})] + \\ \frac{1}{2}\{\mathbb{E}_{(\boldsymbol{x}_a,t_b)\sim p_{\text{data}}}[\log(1 - D_{x,\varphi_t}(\boldsymbol{x}_a, \varphi_{t_b}))] + \\ \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),t_a\sim p_{\text{data}}}[\log(1 - D_{x,\varphi_t}(G_x(\boldsymbol{z}, \varphi_{t_a}), \varphi_{t_a}))]\}, \quad (1)\end{aligned}$$

where $(\boldsymbol{x}_a, t_a)$ is a matched pair and $(\boldsymbol{x}_a, t_b)$ is a mismatched pair.

To augment the text data, Zhang *et al.* [32] replace the deterministic text embedding φ_t in the generator with a latent variable $\boldsymbol{c}$, which is sampled from a learned Gaussian distribution $p(\boldsymbol{c}|\varphi_t) = \mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t))$, where μ and Σ are functions of φ_t parameterized by neural networks. For simplicity in notation, we denote $p(\boldsymbol{c}|\varphi_t)$ as $p(\boldsymbol{c})$. As a result, the objective function (1) is updated to:

$$\begin{aligned}\min_G \max_D V_{t2i}(D_{x,\varphi_t}, G_x) = \\ \mathbb{E}_{(\boldsymbol{x}_a,t_a)\sim p_{\text{data}}}[\log D_{x,\varphi_t}(\boldsymbol{x}_a, \varphi_{t_a})] + \\ \frac{1}{2}\{\mathbb{E}_{(\boldsymbol{x}_a,t_b)\sim p_{\text{data}}}[\log(1 - D_{x,\varphi_t}(\boldsymbol{x}_a, \varphi_{t_b}))] + \\ \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{c}\sim p(\boldsymbol{c}),t_a\sim p_{\text{data}}}[\log(1 - D_{x,\varphi_t}(G_x(\boldsymbol{z}, \boldsymbol{c}), \varphi_{t_a}))]\}. \quad (2)\end{aligned}$$

In addition to the matching-aware pair loss that guarantees the semantic consistency, Zhang *et al*. [35] propose another type of adversarial loss that focuses on the image fidelity (*i.e*., image loss), further updating (2) to:

$$\min_G \max_D V_{t2i}(D_x, D_{x,\varphi_t}, G_x) = \mathbb{E}_{\boldsymbol{x}_a \sim p_{\text{data}}}[\log D_x(\boldsymbol{x}_a)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{c}\sim p(\boldsymbol{c})}[\log(1 - D_x(G_x(\boldsymbol{z},\boldsymbol{c})))] + \mathbb{E}_{(\boldsymbol{x}_a,t_a)\sim p_{\text{data}}}[\log D_{x,\varphi_t}(\boldsymbol{x}_a, \varphi_{t_a})] + \frac{1}{2}\{\mathbb{E}_{(\boldsymbol{x}_a,t_b)\sim p_{\text{data}}}[\log(1 - D_{x,\varphi_t}(\boldsymbol{x}_a, \varphi_{t_b}))] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{c}\sim p(\boldsymbol{c}),t_a\sim p_{\text{data}}}[\log(1 - D_{x,\varphi_t}(G_x(\boldsymbol{z},\boldsymbol{c}), \varphi_{t_a}))]\}, \tag{3}$$

where D_x is a discriminator distinguishing between images sampled from p_{data} and those sampled from the distribution parameterized by the generator (*i.e*., p_{model}).

Consider two general probability distributions $q(\boldsymbol{x})$ and $p(\boldsymbol{z})$ over two domains $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$, where $q(\boldsymbol{x})$ represents the empirical data distribution and $p(\boldsymbol{z})$ is usually specified as a simple random distribution, *e.g*., a standard normal $\mathcal{N}(0, 1)$. Adversarial inference [6, 7] aims to match the two joint distributions $q(\boldsymbol{x}, \boldsymbol{z}) = q(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{x})$ and $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$, which in turn implies that $q(\boldsymbol{z}|\boldsymbol{x})$ matches $p(\boldsymbol{z}|\boldsymbol{x})$. To achieve this, an encoder $G_z(\boldsymbol{x}) : \hat{\boldsymbol{z}} = G_z(\boldsymbol{x}), \boldsymbol{x} \sim q(\boldsymbol{x})$ is introduced in the generation phase, in addition to the standard generator $G_x(\boldsymbol{z}) : \tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}), \boldsymbol{z} \sim p(\boldsymbol{z})$. The discriminator D is trained to distinguish joint pairs between $(\boldsymbol{x}, \hat{\boldsymbol{z}})$ and $(\tilde{\boldsymbol{x}}, \boldsymbol{z})$. The minimax objective of adversarial inference can be written as:

$$\min_G \max_D V(D, G_x, G_z) = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x}),\hat{\boldsymbol{z}}\sim q(\boldsymbol{z}|\boldsymbol{x})}[\log D(\boldsymbol{x}, \hat{\boldsymbol{z}})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z})}[\log(1 - D(\tilde{\boldsymbol{x}}, \boldsymbol{z}))]. \tag{4}$$

3.2. Dual adversarial inference

As described in Section 3.1, the current state-of-the-art methods for text-to-image synthesis can be viewed as variants of conditional GANs, where the conditioning is initially on φ_t itself [26] and later on updated to the latent variable $\boldsymbol{c}$ sampled from a distribution learned through φ_t [32, 35, 30, 24]. The generator then has two latent variables $\boldsymbol{z}$ and $\boldsymbol{c}$: $\boldsymbol{z} \sim p(\boldsymbol{z})$, $\boldsymbol{c} \sim p(\boldsymbol{c})$ (left, Figure 2). The priors can be Gaussian or non-Gaussian distributions such as the Bernoulli distribution [1]. To learn disentangled representations for style ($\boldsymbol{z}$) and content ($\boldsymbol{c}$) and to enforce the separation between these two variables, we incorporate dual adversarial inference into the current framework for text-to-image synthesis (right, Figure 2). In this dual inference process, we are interested in matching the conditional $q(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})$ to the posterior $p(\boldsymbol{z}, \boldsymbol{c}|\boldsymbol{x})$, which under the independence assumption can be factorized as follows:

$$q(\boldsymbol{z}, \boldsymbol{c} \mid \boldsymbol{x}) = q(\boldsymbol{z} \mid \boldsymbol{x})q(\boldsymbol{c} \mid \boldsymbol{x}),$$
$$p(\boldsymbol{z}, \boldsymbol{c} \mid \boldsymbol{x}) = p(\boldsymbol{z} \mid \boldsymbol{x})p(\boldsymbol{c} \mid \boldsymbol{x}).$$

This formulation allows us to match $q(\boldsymbol{z}|\boldsymbol{x})$ with $p(\boldsymbol{z}|\boldsymbol{x})$ and $q(\boldsymbol{c}|\boldsymbol{x})$ with $p(\boldsymbol{c}|\boldsymbol{x})$, respectively. Similar to previous work [7, 6], we achieve this by matching the two pairs of joint distributions:

$$q(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z}, \boldsymbol{x}),$$
$$q(\boldsymbol{c}, \boldsymbol{x}) = p(\boldsymbol{c}, \boldsymbol{x}).$$

The encoder for our dual adversarial inference then encodes both $\boldsymbol{z}$ and $\boldsymbol{c}$: $\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}} = G_{z,c}(\boldsymbol{x}), \boldsymbol{x} \sim q(\boldsymbol{x})$, while the generator decodes $\boldsymbol{z}$ and $\boldsymbol{c}$ sampled from their corresponding prior distributions into an image: $\tilde{\boldsymbol{x}} = G_x(\boldsymbol{z}, \boldsymbol{c}), \boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{c} \sim p(\boldsymbol{c})$. To compete with G_x and $G_{z,c}$, the discrimination phase also has two components: the discriminator $D_{x,z}$ is trained to discriminate $(\boldsymbol{x}, \boldsymbol{z})$ pairs sampled from either $q(\boldsymbol{x}, \boldsymbol{z})$ or $p(\boldsymbol{x}, \boldsymbol{z})$, and the discriminator $D_{x,c}$ for the discrimination of $(\boldsymbol{x}, \boldsymbol{c})$ pairs sampled from either $q(\boldsymbol{x}, \boldsymbol{c})$ or $p(\boldsymbol{x}, \boldsymbol{c})$. Given the above setting, the original adversarial inference objective (4) is updated as:

$$\min_G \max_D V_{dual}(D_{x,z}, D_{x,c}, G_x, G_{z,c}) = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x}),\hat{\boldsymbol{z}},\hat{\boldsymbol{c}}\sim q(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})}[\log D_{x,z}(\boldsymbol{x}, \hat{\boldsymbol{z}}) + \log D_{x,c}(\boldsymbol{x}, \hat{\boldsymbol{c}})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c}),\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{c}\sim p(\boldsymbol{c})}[\log(1 - D_{x,z}(\tilde{\boldsymbol{x}}, \boldsymbol{z})) + \log(1 - D_{x,c}(\tilde{\boldsymbol{x}}, \boldsymbol{c}))]. \tag{5}$$

3.3. Cycle consistency

In unsupervised learning, *cycle-consistency* refers to the ability of the model to reconstruct the original image $\boldsymbol{x}$ from its inferred latent variable $\boldsymbol{z}$. It has been reported that bidirectional adversarial inference models often have difficulties in reproducing faithful reconstructions as they do not explicitly include any reconstruction loss in the objective function [7, 6, 19]. The cycle-consistency criterion, as having been demonstrated in many previous works such as CycleGAN [36], DualGAN [31], DiscoGAN [14] and augmented CycleGAN [1], enforces a strong connection between domains (here $\boldsymbol{x}$ and $\boldsymbol{z}$) by constraining the models (*e.g*., encoder and decoder) to be consistent with one another. Li *et al*. [19] show that the integration of the cycle-consistency objective stabilizes the learning of adversarial inference, thus yielding better reconstruction results. With the above in mind, we integrate cycle-consistency in our dual adversarial inference framework in a similar fashion to [19] . More concretely, we use another discriminator $D_{x,x'}$ to distinguish between $\boldsymbol{x}$ and its reconstruction

[1] In this paper, we experiment with both Gaussian and Bernoulli distributions for $p(\boldsymbol{c})$ (More details in Section 4).

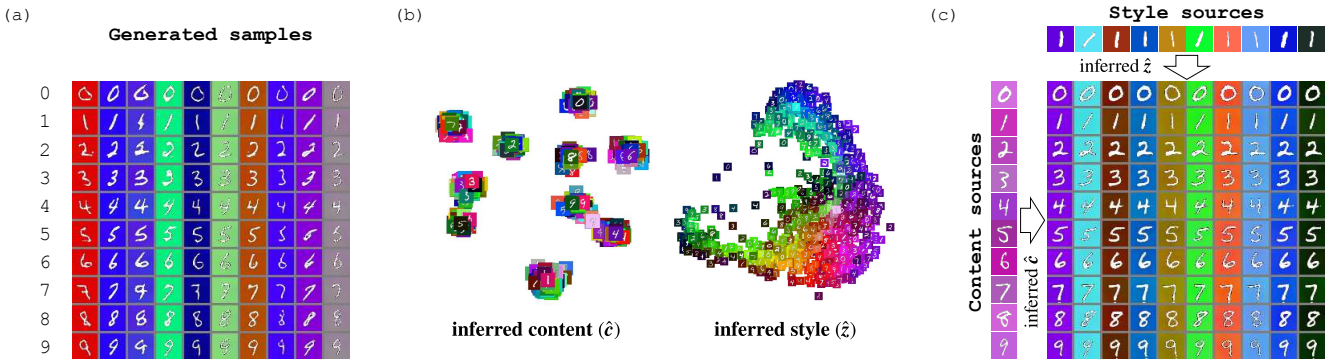

Figure 3: Disentangling content and style on MNIST-CB dataset. (a) Generated samples given digit identities as the content $\boldsymbol{c}$. Each column uses the same style $\boldsymbol{z}$ sampled from $\mathcal{N}(0, 1)$. (b) The t-SNE visualizations of inferred content $\hat{\boldsymbol{c}}$ and inferred style $\hat{\boldsymbol{z}}$. (c) Reconstructed samples using inferred content $\hat{\boldsymbol{c}}$ (in rows) and inferred style $\hat{\boldsymbol{z}}$ (in columns) from image sources.

$\boldsymbol{x}' = G_x(\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}})$, where $\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}} = G_{z,c}(\boldsymbol{x})$, by optimizing:

$$\min_G \max_D V_{cycle}(D_{x,x'}, G_x, G_{z,c}) = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log D_{x,x'}(\boldsymbol{x}, \boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x}),(\hat{\boldsymbol{z}},\hat{\boldsymbol{c}})\sim q(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})}[\log(1 - D_{x,x'}(\boldsymbol{x}, G_x(\hat{\boldsymbol{z}}, \hat{\boldsymbol{c}})))]. \quad (6)$$

We later show in an ablation study (Section 4.6) that using l_2 loss for cycle-consistency leads to blurriness in the generated images, which agrees with previous studies [17, 31].

3.4. Full objective

Taking (3), (5), (6) into account, our full objective is:

$$\begin{aligned}\min_G \max_D V_{full}(D, G) &= V_{t2i}(D_x, D_{x,\varphi_t}, G_x) \\ &+ V_{dual}(D_{x,z}, D_{x,c}, G_x, G_{z,c}) \\ &+ V_{cycle}(D_{x,x'}, G_x, G_{z,c}), \end{aligned} \quad (7)$$

where G and D are the sets of all generators and discriminators in our method: $G = \{G_x, G_{z,c}\}$ and $D = \{D_x, D_{x,\varphi_t}, D_{x,z}, D_{x,c}, D_{x,x'}\}$.

Note that in addition to the latent variable $\boldsymbol{c}$, the encoded $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{c}}$ in our method are also sampled from the inferred posterior distributions through the reparameterization trick [16], *i.e.*, $\hat{\boldsymbol{z}} \sim q(\boldsymbol{z}|\boldsymbol{x})$ and $\hat{\boldsymbol{c}} \sim q(\boldsymbol{c}|\boldsymbol{x})$. In order to encourage smooth sampling over the latent space, we regularize the posterior distributions $q(\boldsymbol{z}|\boldsymbol{x})$ and $q(\boldsymbol{c}|\boldsymbol{x})$ to match their respective priors by minimizing the KL divergence. We apply a similar regularization term to $p(\boldsymbol{c})$, *e.g.*, $\lambda D_{KL}(p(\boldsymbol{c}) \,||\, \mathcal{N}(0, 1))$ for a normal distribution prior, as done in previous text-to-image synthesis works [32, 35]. Our preliminary experiments [2] showed that without the above regularization, the training became unstable and the gradients typically explode after certain number of epochs.

[2]We also experimented with minimizing the cosine similarity between $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{c}}$, but did not observe improved performance in terms of the inception score and FID.

4. Experiments

4.1. Proof-of-concept study

To evaluate the effectiveness of our proposed dual adversarial inference on the disentanglement of content and style, we first validate our proposed method on a toy dataset: MNIST-CB [8], where we formulate the digit generation problem as a text-to-image synthesis problem by considering the digit identity as the text content. In this setup, digit font and background color represent styles learned in an unsupervised manner through adversarial inference. We add a cross-entropy regularization term to the content inference objective since our content in this case is discrete (*i.e.*, one-hot vector for digit identity). As shown in Figure 3 (a), the content and style are disentangled in the generation phase, where the generator has learned to assign the same style to different digit identities when the same $\boldsymbol{z}$ is used. More importantly, the t-SNE visualizations (Figure 3 (b)) from our inferred content and style ($\hat{\boldsymbol{c}}$ and $\hat{\boldsymbol{z}}$) indicate that our dual adversarial inference has successfully separated the information on content (digit identity) and style (font and background color). This is further validated in Figure 3 (c) where we show our model's ability to infer style and content from different image sources and fuse them to generate hybrid images, using content from one source and style from the other.

4.2. Text-to-image setup

Once validated on the toy example, we move to the original text-to-image synthesis task. We evaluate our method based on model architectures similar to HDGAN [35], one of the current state-of-the-art methods for text-to-image synthesis, making HDGAN our baseline method. The architecture designs are the same as described in [35], keeping in mind that we only consider the 64×64 resolution. Three quantitative metrics are used to evaluate our method: Inception score [28], Fréchet inception distance (FID) [10] and

Method	Inception Score			FID		
	Oxford-102	CUB	COCO	Oxford-102	CUB	COCO
GAN-INT-CLS [26]	2.66 ± 0.03	2.88 ± 0.04	7.88 ± 0.07	79.55	68.79	60.62
GAWWN [27]	—	3.10 ± 0.03	—	—	53.51	—
StackGAN [32, 33]	2.73 ± 0.03	3.02 ± 0.03	8.35 ± 0.11	43.02	35.11	33.88
HDGAN [35]	—	3.53 ± 0.03	—	—	—	—
HDGAN mean*	2.90 ± 0.03	$\mathbf{3.58 \pm 0.03}$	8.64 ± 0.37	40.02 ± 0.55	20.60 ± 0.96	29.13 ± 3.76
Ours mean*	$\mathbf{2.90 \pm 0.03}$	$\mathbf{3.58 \pm 0.05}$	$\mathbf{8.94 \pm 0.20}$	$\mathbf{37.94 \pm 0.39}$	$\mathbf{18.41 \pm 1.07}$	$\mathbf{27.07 \pm 2.55}$

* mean calculated on three experiments at five different epochs (600, 580, 560, 540, 520), or three different epochs (200, 190, 180) for COCO dataset

Table 1: Comparison of inception score and FID at 64×64 resolution scale. Higher inception score and lower FID mean better performance.

Figure 4: Examples of generated images on Oxford-102 (top), CUB (middle) and COCO (bottom) datasets.

Visual-semantic similarity [35]. It has been noticed in our experiments and also reported by others [21] that, due to the variations in the training of GAN models, it is unfair to draw a conclusion based on one single experiment that achieves the best result; therefore, in our experiments, we perform three independent experiments for each method, with averages reported as final results. More implementation, dataset and evaluation details can be found in the supplementary material.

4.3. Quantitative results

To get a global overview of how our method, the baseline method and its variants (by either fixing or removing the noise vector z) behave throughout training, we evaluate each model in 20 epoch intervals. Figure 9 (supplementary material) shows inception score (left axis) and FID (right axis) for both Oxford-102 and CUB datasets. Consistent with the qualitative results presented in Figure 8 (supplementary material), we quantitatively show that by either fixing or removing z, the baseline models retain unimpaired performance, suggesting that z has no contribution in the baseline models. However, with our proposed dual adversarial inference, the model performance is significantly improved on FID scores for both datasets (red curves, Figure 9), indicating the proposed method's ability to produce better-quality images. Table 1 summarizes the comparison of the results of our method to the baseline method and also other reported results of previous state-of-the-art methods for the 64×64 resolution task on the three benchmark datasets: Oxford-102, CUB and COCO. Our method achieves the best performance based on the mean scores for both metrics on all datasets; on the FID score, it shows a 5.2% improvement (from 40.02 to 37.94) on the Oxford-102 dataset, and a 10.6% improvement (from 20.60 to 18.41) on the CUB dataset. In addition, we also achieve comparable results on visual-semantic similarity (Table 3, supplementary material).

4.4. Qualitative results

In this subsection, we present qualitative results on text-to-image generation and interpolation analysis based on inferred content ($\hat{c}$) and inferred style ($\hat{z}$).

First, we visually compare the quality and diversity of images generated from our method against the baseline. Figure 4 shows one example for each dataset, illustrating that our method is able to generate better-quality images compared to the baseline method, which agrees with our quantitative results in Table 1. We provide more examples in the supplementary material (Section 6.8).

To make sure we are not overfitting, and to investigate whether we have learned a representative latent space, we look at interpolations of projected locations in the latent space. Interpolations also enable us to examine whether the model has indeed learned to separate style from content in an unsupervised way. To do this, we provide the trained inference model with two images: the source image and the target image, and extract their projections $\hat{z}$ and $\hat{c}$ for interpolation analysis. As shown in Figure 5, the rows correspond to reconstructed images of linear interpolations in $\hat{c}$ from source to target image and the same for $\hat{z}$ as dis-

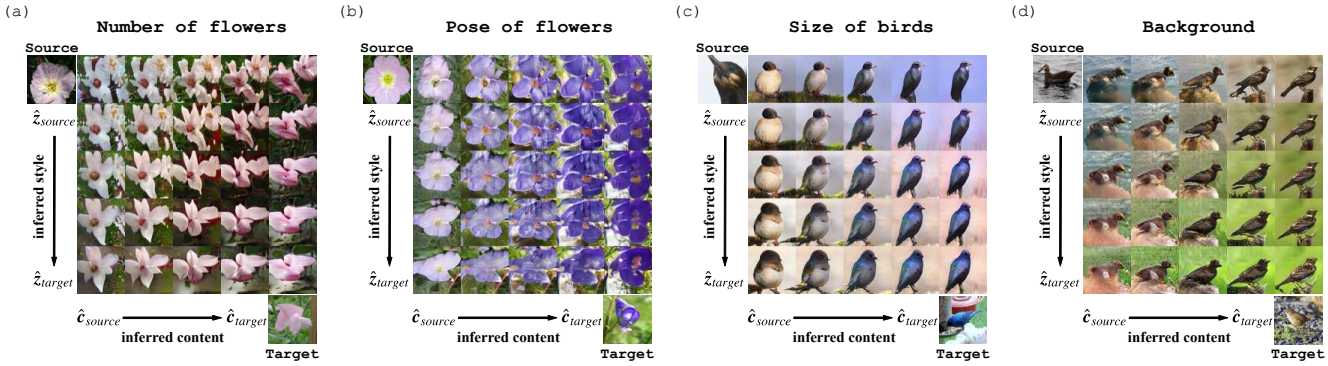

Figure 5: Examples of reconstructed images by interpolation of inferred content $\hat{c}$ and inferred style $\hat{z}$ from sources to targets. The learned style information includes: (a) quantity, (b) pose, (c) size and (d) background.

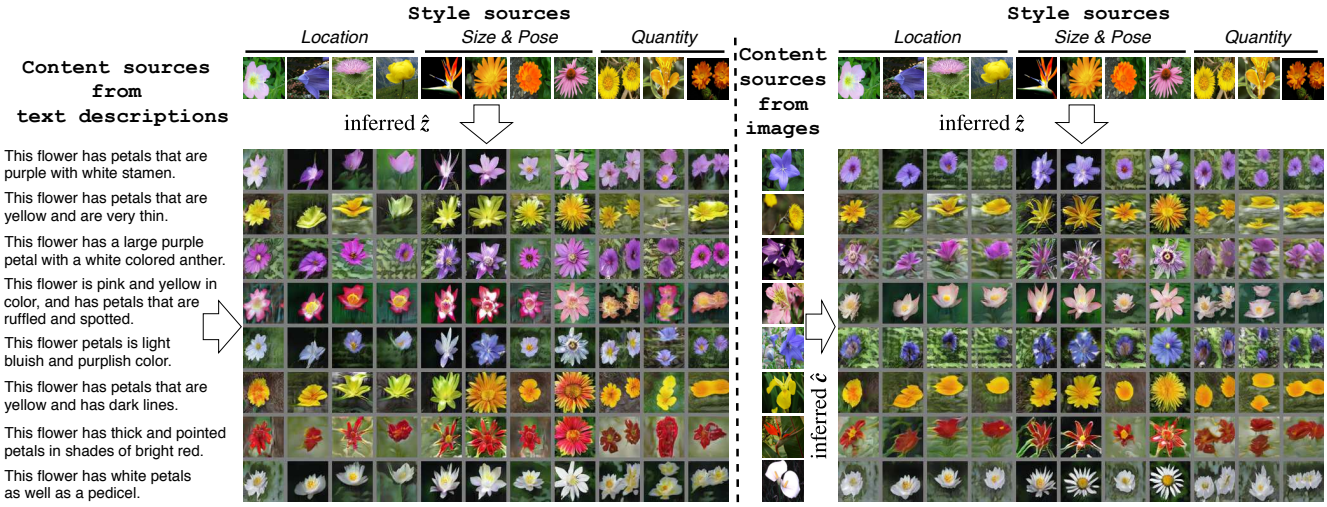

Figure 6: Disentangling content (in rows) and style (in columns) on Oxford-102 dataset by using content sources either from text descriptions (left) or images (right). More results are provided in the supplementary material (Section 6.9).

played in columns. The smooth transitions of both the content represented by $\hat{c}$ from the left to right and the style represented by $\hat{z}$ from the top to bottom indicate a good generalization of our model representing both latent spaces, and more interestingly, we find promising results showing that $\hat{z}$ is indeed controlling some meaningful style information, *e.g.*, the number and pose of flowers, the size of birds and the background (Figure 5, more examples in supplementary material).

4.5. Disentanglement constraint

Despite promising results evidenced by many such examples as shown in Figure 5, we notice that the information captured by inferred style ($\hat{z}$) is not always consistent and faithful when we use Gaussian priors for both content and style. Inspired by the theories from independent component analysis (ICA) for separating a multivariate signal into additive subcomponents [4], we use a Bernoulli distribution for the content representation to satisfy the non-Gaussian constraint. This provides us with a better disentanglement of content and style. Note that an alternative approach for ICA has also recently been explored in [13]. As shown in Figure 6 and Figure 7, our models learn to synthesize images by combining content and style information from different sources while preserving their respective properties (*e.g.*, color for the content; and location, pose, quantity, etc. for the style), which suggests the disentanglement of content and style. Note that the content information can either directly come from a text description (left, Figure 6 and Figure 7) or be inferred from an image source (right, Figure 6 and Figure 7). More examples and discussions are provided in the supplementary material (Section 6.9).

Higgins *et al.* [11] and Zhang *et al.* [34] have proposed quantitative metrics for the disentanglement analysis which involve classification of the style attributes or comparison of the distance between generated style and true style. However, in our case, the dataset does not contain any labeled attribute that can be used to evaluate a captured style. As a

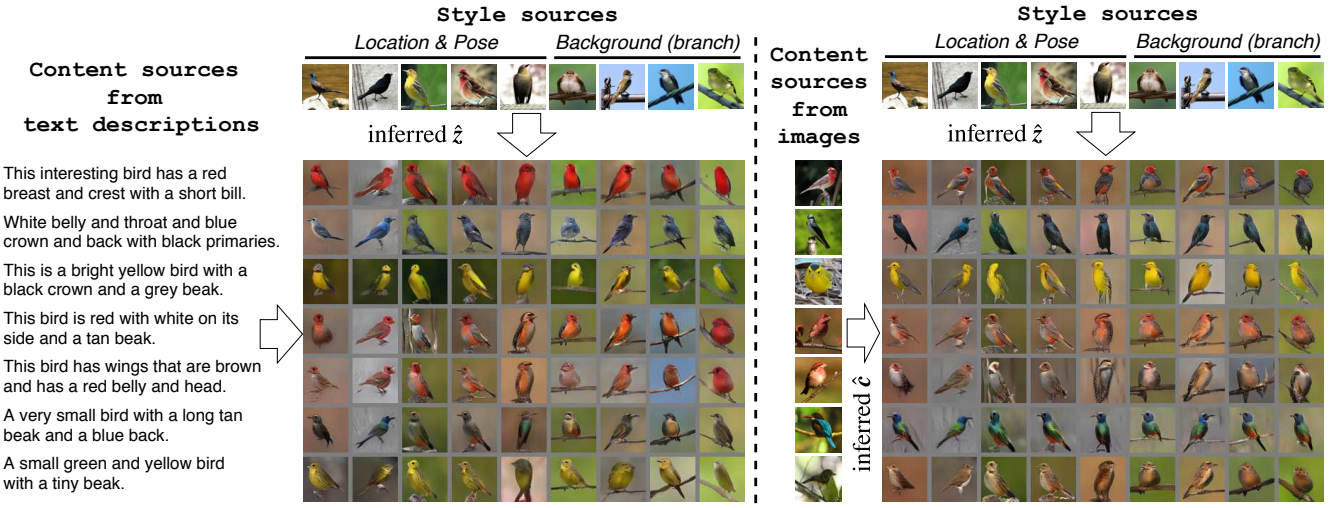

Figure 7: Disentangling content (in rows) and style (in columns) on CUB dataset by using content sources either from text descriptions (left) or images (right). More results are provided in the supplementary material (Section 6.9).

result, their proposed metrics would not be suitable in our case. One possible solution would be to artificially create a new dataset that has the same content over multiple known styles. We leave this exploration for future work.

4.6. Ablation study

In our method, we have multiple components, each of which is optimized by its corresponding objective. The previous works [26, 32, 35] for text-to-image synthesis use the discriminator D_{x,φ_t} to discriminate whether the image $\boldsymbol{x}$ matches its text embedding φ_t. However, with the integration of adversarial inference, where a new discriminator $D_{x,c}$ is designed to match the joint distribution of $(\boldsymbol{x}, \hat{\boldsymbol{c}})$ and $(\tilde{\boldsymbol{x}}, \boldsymbol{c})$, we now question whether the discriminator D_{x,φ_t} is still required, given the fact that $\boldsymbol{c}$ is learned from φ_t. To answer this question, we remove the objective $V_{t2i}(D, G)$ from our method, and as seen in Table 2, the performance on the CUB dataset significantly drops for both inception score and FID, indicating that D_{x,φ_t} is not redundant in our method by providing strong supervision over the text embeddings. Similarly, we examine the role of cycle-consistency loss in our method by removing $V_{cycle}(D, G)$ from the objective. We observe a slight drop in both inception score and FID (Table 2), suggesting that cycle-consistency can further improve the learning of adversarial inference, which is in agreement with [19]. It is also worth mentioning that our method without cycle-consistency still achieves better FID scores than the baseline method on the CUB dataset (Table 1 and Table 2), which additionally supports our proposal to integrate the inference mechanism in the current text-to-image framework. We also examine the model performance by using l_2 loss for cycle-consistency instead of the adversarial loss. The resulting degradation in quality is unexpectedly dramatic (Table 2). Figure 10 (supplementary material) shows the generated images using adversarial loss compared with those using l_2 loss, and it is clear that the latter gives blurrier images.

Method	Inception Score	FID
ours	3.58 ± 0.05	18.41 ± 1.07
ours without V_{t2i}	3.31 ± 0.04	20.65 ± 0.47
ours without V_{cycle}	3.53 ± 0.06	19.29 ± 0.90
l_2 loss for V_{cycle}	1.73 ± 0.15	149.8 ± 16.4

Table 2: Ablation study on CUB dataset. Note that the ablation on V_{dual} eventually turns into the baseline.

5. Conclusion

In this paper, we incorporate a dual adversarial inference procedure in order to learn disentangled representations of content and style in an unsupervised way, which we show improves text-to-image synthesis. It is worth noting that the content is learned both in a supervised way through the text embedding and in an unsupervised way through the adversarial inference. The style, however, is learned solely in an unsupervised manner. Despite the challenges of the task, we show promising results on interpreting what has been learned for style. With the proposed inference mechanism, our method achieves improved quality and comparable variability in generated images evaluated on Oxford-102, CUB and COCO datasets.

Acknowledgements This work was supported by Mitacs project IT11934. The authors thank Nicolas Chapados for his constructive comments, and Gabriel Chartrand, Thomas Vincent, Andew Jesson, Cecile Low-Kam and Tanya Nair for their help and review.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*, 2018. 4

[2] Miriam Cha, Youngjune L Gown, and HT Kung. Adversarial learning of semantic relevance in text to image synthesis. In *AAAI*, 2019. 2

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 3

[4] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. 7

[5] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal. Tac-gan-text conditioned auxiliary classifier generative adversarial network. In *arXiv preprint arXiv:1703.06412*, 2017. 2

[6] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017. 2, 4

[7] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017. 2, 3, 4

[8] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *NIPS*, 2018. 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[10] 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 *NIPS*, 2017. 5, 12

[11] 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. 7

[12] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, 2018. 2

[13] Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. 2019. 7

[14] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 4

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 12

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[17] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 5

[18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[19] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *NIPS*, 2017. 3, 4, 8

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 12

[21] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *NIPS*, 2018. 6

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

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2, 12

[24] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, 2019. 2, 4

[25] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 12

[26] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 1, 2, 3, 4, 6, 8, 12, 13

[27] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, 2016. 2, 6

[28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 5, 12

[29] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 2, 12

[30] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2, 4

[31] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 4, 5

[32] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 1, 2, 3, 4, 5, 6, 8, 12

[33] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. In *arXiv preprint arXiv:1710.10916*, 2017. 6

[34] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *CVPR*, 2018. 7

[35] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018. 2, 4, 5, 6, 8, 12

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A

Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017. 4