Multimodal Conditional Image Synthesis with Product-of-Experts GANs

Xun Huang Arun Mallya Ting-Chun Wang Ming-Yu Liu

NVIDIA

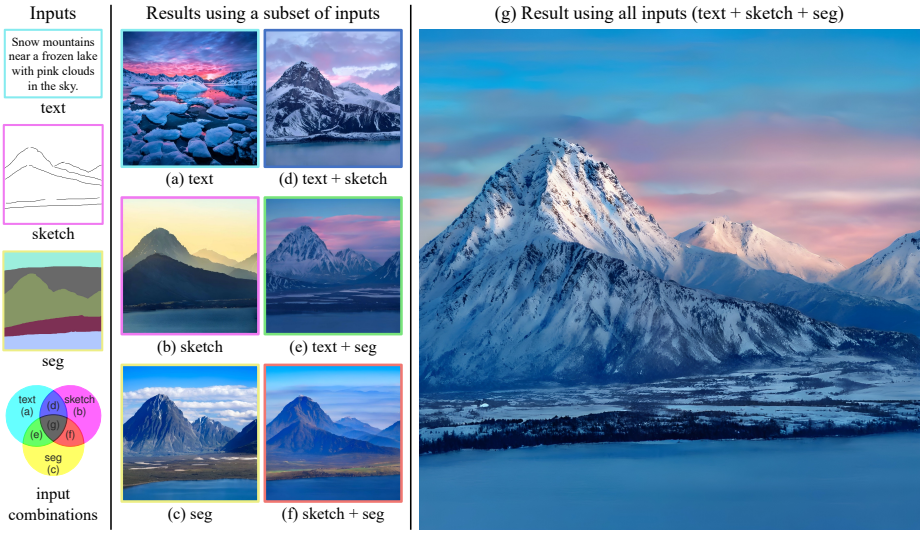


Fig. 1: Given conditional inputs in multiple modalities (the left column), our approach can synthesize images that satisfy all input conditions (the right column, (g)) or an arbitrary subset of input conditions (the middle column, (a)–(f)) with *a single model*.

Abstract. Existing conditional image synthesis frameworks generate images based on user inputs in a single modality, such as text, segmentation, or sketch. They do not allow users to simultaneously use inputs in multiple modalities to control the image synthesis output. This reduces their practicality as multimodal inputs are more expressive and complement each other. To address this limitation, we propose the Product-of-Experts Generative Adversarial Networks (PoE-GAN) framework, which can synthesize images conditioned on multiple input modalities or any subset of them, even the empty set. We achieve this capability with a single trained model. PoE-GAN consists of a product-of-experts generator and a multimodal multiscale projection discriminator. Through our carefully designed training scheme, PoE-GAN learns to synthesize images with high quality and diversity. Besides advancing the state of the art in multimodal conditional image synthesis, PoE-GAN also outperforms the best existing unimodal conditional image synthesis approaches when tested in the unimodal setting. The project website is available at this link.

Keywords: Image synthesis, multimodal learning, GAN.

1 Introduction

Conditional image synthesis allows users to use their creative inputs to control the output of image synthesis methods. It has found applications in many content creation tools. Over the years, a variety of input modalities have been studied, mostly based on conditional GANs [11,19,34,27]. To this end, we have various single modality-to-image models. When the input modality is text, we have the *text-to-image* model [41,62,58,67,51,60,40]. When the input modality is a segmentation mask, we have the *segmentation-to-image* model [19,54,36,29,45,8]. When the input modality is a sketch, we have the *sketch-to-image* model [44,6,10,4].

However, different input modalities are best suited for conveying different types of conditioning information. For example, as seen in the first column of Fig. 1, segmentation makes it easy to define the coarse layout of semantic classes in an image—the relative locations of sky, cloud, mountain, and water regions. Sketch allows us to specify the structure and details within the same semantic region, such as individual mountain ridges. On the other hand, text is well-suited for modifying and describing objects or regions in the image, which cannot be achieved via segmentation or sketch, *e.g.*, '*frozen* lake' and '*pink* clouds' in Fig. 1. Despite this synergy among modalities, prior work has considered image generation conditioned on each modality as a distinct task and studied it in isolation. Existing models thus fail to utilize complementary information available in different modalities. Clearly, a conditional generative model that can combine input information from all available modalities would be of immense value.

Even though the benefits are enormous, the task of conditional image synthesis with multiple input modalities poses several challenges. First, it is unclear how to combine multiple modalities with different dimensions and structures in a single framework. Second, from a practical standpoint, the generator needs to handle missing modalities since it is cumbersome to ask users to provide every single modality all the time. This means that the generator should work well even when only a subset of modalities are provided. Lastly, conditional GANs are known to be susceptible to mode collapse [19,34], wherein the generator produces identical images when conditioned on the same inputs. This makes it difficult for the generator to produce diverse output images that capture the full conditional distribution when conditioned on an arbitrary set of modalities.

We present Product-of-Experts Generative Adversarial Networks (PoE-GAN), a framework that can generate images conditioned on any subset of the input modalities presented during training, as illustrated in Fig. 1 (a)-(g). This framework provides users unprecedented control, allowing them to specify exactly what they want using multiple complementary input modalities. When users provide no inputs, it falls back to an unconditional GAN model [11,39,32,20,2,22,23,21]. One key ingredient of our framework is a novel product-of-experts generator that can effectively fuse multimodal user inputs and handle missing modalities (Sec. 3.1). A novel hierarchical and multiscale latent representation leads to better usage of the structure in spatial modalities, such as segmentation and sketch (Sec. 3.2). Our model is trained with a multimodal projection discriminator (Sec. 3.4) together with contrastive losses for better input-output alignment. In addition, we

adopt modality dropout for additional robustness to missing inputs (Sec. 3.5). Extensive experiment results show that PoE-GAN outperforms prior work in both multimodal and unimodal settings (Sec. 4), including state-of-the-art approaches specifically designed for a single modality. We also show that PoE-GAN can generate diverse images when conditioned on the same inputs.

2 Related Work

Image Synthesis. Our network architecture design is inspired by previous work in unconditional image synthesis. Our decoder employs some techniques proposed in StyleGAN [22] such as global modulation. Our latent space is constructed in a way similar to hierarchical variational auto-encoders (VAEs) [48,30,52,7]. While hierarchical VAEs encode the image itself to the latent space, our network encodes conditional information from different modalities into a unified latent space. Our discriminator design is inspired by the projection discriminator [33] and multiscale discriminators [54,29], which we extend to our multimodal setting.

Multimodal Image Synthesis. Prior work [50,53,56,46,49,25] has explored learning the joint distribution of multiple modalities using VAEs [24,42]. Some of them [53,56,25] use a product-of-experts inference network to approximate the posterior distribution. This is conceptually similar to how our generator combines information from multiple modalities. While their goal is to estimate the complete joint distribution, we focus on learning the image distribution conditioned on other modalities. Besides, our framework is based on GANs rather than VAEs and we perform experiments on high-resolution and large-scale datasets, unlike the above work. Recently, Xia *et al.* [57] propose a GAN-based multimodal image synthesis method named TediGAN. Their method relies on a pretrained unconditional generator. However, such a generator is difficult to train on a complex dataset such as MS-COCO [26]. Concurrently, Zhang *et al.* [65] propose a multimodal image synthesis method based on VQGAN. The way they combine different modalities is similar to our baseline using concatenation and modality dropout (Sec. 4.2). We will show that our product-of-experts generator design significantly improves upon this baseline. Another parallel work by Gafni *et al.* [9] propose a VQGAN model that is conditioned on both text and segmentation. They assume both modalities are always present at inference time and cannot deal with missing modalities.

3 Product-of-experts GANs

Given a dataset of images x paired with M different input modalities $(y_1, y_2, ..., y_M)$, our goal is to train a single generative model that learns to capture the image distribution conditioned on an arbitrary subset of possible modalities $p(x|\mathcal{Y}), \forall \mathcal{Y} \subseteq \{y_1, y_2, ..., y_M\}$. In this paper, we consider four different modalities including text,

semantic segmentation, sketch, and style reference. Note that our framework is general and can easily incorporate additional modalities.

Learning image distributions conditioned on *any subset* of M modalities is challenging because it requires a single generator to simultaneously model 2^M distributions. Of particular note, the generator needs to capture the unconditional image distribution $p(x)$ when $\mathcal{Y}$ is an empty set, and the unimodal conditional distributions $p(x|y_i), \forall i \in \{1, 2, ..., M\}$, such as the image distribution conditioned on text alone. These settings have been popular and widely studied in isolation, and we aim to bring them all under a unified framework.

3.1 Product-of-experts modeling

Our generator consists of a decoder G that deterministically maps a latent code z to an output image x, and a set of encoders that estimate the latent distribution $p(z|\mathcal{Y})$ conditioned on a set of modalities $\mathcal{Y}$. The conditional image distribution $p(x|\mathcal{Y})$ is implicitly defined as $x = G(z), z \sim p(z|\mathcal{Y})$. A naive approach would require us to train 2^M different encoder networks, one for each possible combination of modalities. This is highly parameter-inefficient and does not scale to a large number of modalities. Fortunately, if we assume all modalities $(y_1, ..., y_M)$ are conditionally independent given the image (x or equivalently z), i.e., $p(y_1, ..., y_M|z) = \prod_{i=1}^{M} p(y_i|z)$[1], we can prove that the distribution $p(z|\mathcal{Y})$ is proportional to a product of distributions:

$$p(z|\mathcal{Y}) = \frac{p(\mathcal{Y}|z)p(z)}{p(\mathcal{Y})} = \frac{p(z)}{p(\mathcal{Y})} \prod_{y_i \in \mathcal{Y}} p(y_i|z) = \frac{p(z)}{p(\mathcal{Y})} \prod_{y_i \in \mathcal{Y}} \frac{p(z|y_i)p(y_i)}{p(z)}$$
$$= \frac{\prod_{y_i \in \mathcal{Y}} p(z|y_i)}{(p(z))^{|\mathcal{Y}|-1}} \cdot \frac{\prod_{y_i \in \mathcal{Y}} p(y_i)}{p(\mathcal{Y})} \propto \frac{\prod_{y_i \in \mathcal{Y}} p(z|y_i)}{(p(z))^{|\mathcal{Y}|-1}} = p(z) \prod_{y_i \in \mathcal{Y}} \tilde{q}(z|y_i)\,, \tag{1}$$

where $\tilde{q}(z|y_i) \equiv \frac{p(z|y_i)}{p(z)}$. Dividing it by the normalization constant, we have

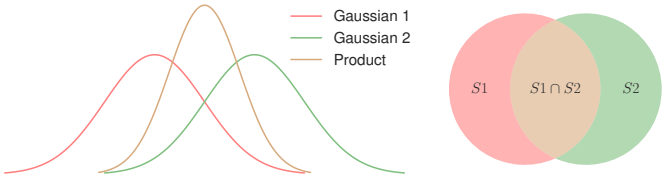
$$p(z|\mathcal{Y}) \propto p(z) \prod_{y_i \in \mathcal{Y}} q(z|y_i), q(z|y_i) = \frac{\tilde{q}(z|y_i)}{\int \tilde{q}(z|y_i)dz}\,, \tag{2}$$

where $q(z|y_i)$ is a latent distribution only dependent on a single modality y_i and $p(z)$ is the unconditional prior distribution. As a result, we can reduce the number of encoders from 2^M to M, with each encoder estimating the distribution $q(z|y_i)$ from a single modality[2]. This idea of combining several distributions ("*experts*") by multiplying them has been previously referred to as product-of-experts [16].

Figs. 2a and 2b show that the product of distributions is intuitively analogous to the intersection of sets. The product distribution only has a high density in

[1] The conditional independence assumption is sound in our setting since an image alone contains sufficient information to infer a modality independent of other modalities. For example, given an image, we do not need its caption to infer its segmentation.

[2] With a slight abuse of notation, we will use $q(z|y_i)$ (and similarly $p(z|\mathcal{Y})$) to denote both the "true" distribution and the estimated distribution produced by our network.



(a) Product of distributions. (b) Intersection of sets.

Fig. 2: The product of distributions (a) is analogous to the intersection of sets (b).

regions where all distributions have a relatively high density. Also, the product distribution is always narrower (of lower entropy) than the individual distributions, just like the intersection of sets is always smaller than the individual set. While each set poses a *hard* constraint, each individual distribution in a product represents a *soft* constraint, which is more amenable to neural network learning. In the multimodal conditional image synthesis setting, the model samples images from the prior $p(z)$ when no modalities are given. Each additional modality y_i specifies a set of images that satisfy a certain constraint and we model that by multiplying the prior with an additional distribution $q(z|y_i)$.

3.2 Multiscale and hierarchical latent space

Some of the modalities we consider (*e.g.*, sketch, segmentation) are two-dimensional and naturally contain information at multiple scales. Therefore, we devise a hierarchical latent space with latent variables at different resolutions. This allows us to directly pass information from each resolution of the encoder to the corresponding resolution of the latent space, so that the high-resolution control signals can be better preserved. Mathematically, our latent code is partitioned into groups $z = (z^0, z^1, ..., z^N)$ where $z^0 \in \mathbb{R}^{c_0}$ is a feature vector and $z^k \in \mathbb{R}^{c_k \times r_k \times r_k}, 1 \leq k \leq N$ are feature maps of increasing resolutions ($r_{k+1} = 2r_k, r_1 = 4, r_N$ is the image resolution). We can therefore decompose the prior $p(z)$ into $\prod_{k=0}^{N} p(z^k|z^{<k})$ and the experts $q(z|y_i)$ into $\prod_{k=0}^{N} q(z^k|z^{<k}, y_i)$, where $z^{<k}$ denotes $(z^0, z^1, ..., z^{k-1})$. Following Eq. (2), we assume the conditional latent distribution at each resolution is a product-of-experts given by

$$p(z^k|z^{<k}, \mathcal{Y}) \propto p(z^k|z^{<k}) \prod_{y_i \in \mathcal{Y}} q(z^k|z^{<k}, y_i)\,, \tag{3}$$

where $p(z^k|z^{<k}) = \mathcal{N}\left(\mu_0^k, \sigma_0^k\right)$ and $q(z^k|z^{<k}, y_i) = \mathcal{N}\left(\mu_i^k, \sigma_i^k\right)$ are independent Gaussian distributions with mean and standard deviation parameterized by a neural network.[3] It can be shown [55] that the product of Gaussian experts is

[3] Except for $p(z^0)$, which is simply a standard Gaussian distribution.

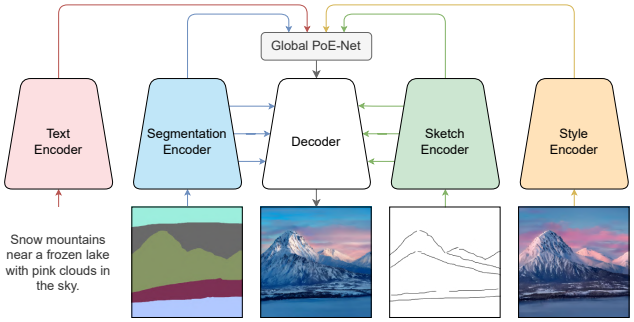
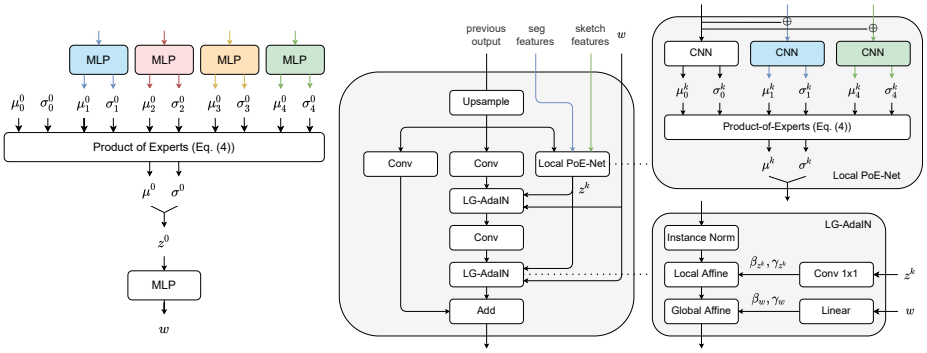


Fig. 3: An overview of our generator. The architecture of Global PoE-Net and decoder are detailed in Fig. 4a and Fig. 4b respectively. The architecture of modality encoders are described in Supplementary Material B.



(a) Global PoE-Net (b) A residual block in our decoder.

Fig. 4: (a) Global PoE-Net. We sample a latent feature vector z^0 using product-of-experts (Eq. (4) in Sec. 3.2), which is then processed by an MLP to output a feature vector w. (b) A residual block in our decoder. Local PoE-Net samples a latent feature map z^k using product-of-experts. Here $\oplus$ denotes concatenation. LG-AdaIN uses w and z^k to modulate the feature activations in the residual branch.

also a Gaussian $p(z^k|z^{<k}, \mathcal{Y}) = \mathcal{N}(\mu^k, \sigma^k)$, with

$$\mu^k = \frac{\frac{\mu_0^k}{(\sigma_0^k)^2} + \sum_i \frac{\mu_i^k}{(\sigma_i^k)^2}}{\frac{1}{(\sigma_0^k)^2} + \sum_i \frac{1}{(\sigma_i^k)^2}}, \quad \sigma^k = \frac{1}{\frac{1}{(\sigma_0^k)^2} + \sum_i \frac{1}{(\sigma_i^k)^2}}. \tag{4}$$

3.3 Generator architecture

Figure 3 shows an overview of our generator architecture. We encode each modality into a feature vector which is then aggregated in *Global PoE-Net*. We use convolutional networks with input skip connections to encode segmentation and sketch maps, a residual network to encode style images, and CLIP [38] to

encode text. Details of all modality encoders are given in Supplementary Material B. The decoder generates the image using the output of Global PoE-Net and skip connections from the segmentation and sketch encoders.

In Global PoE-Net (Fig. 4a), we predict a Gaussian $q(z^0|y_i) = \mathcal{N}\left(\mu_i^0, \sigma_i^0\right)$ from the feature vector of each modality using an MLP. We then compute the product of Gaussians including the prior $p(z^0) = \mathcal{N}(\mu_0^0, \sigma_0^0) = \mathcal{N}(0, I)$ and sample z^0 from the product distribution. An MLP further convert z^0 to the vector w.

The decoder mainly consists of a stack of residual blocks[4] [14], each of which is shown in Fig. 4b. *Local PoE-Net* samples the latent feature map z^k at the current resolution from the product of $p(z^k|z^{<k}) = \mathcal{N}\left(\mu_0^k, \sigma_0^k\right)$ and $q(z^k|z^{<k}, y_i) = \mathcal{N}\left(\mu_i^k, \sigma_i^k\right), \forall y_i \in \mathcal{Y}$, where $\left(\mu_0^k, \sigma_0^k\right)$ is computed from the output of the last layer and $\left(\mu_i^k, \sigma_i^k\right)$ is computed by concatenating the output of the last layer and the skip connection from the corresponding modality. Note that only modalities that have skip connections (segmentation and sketch, i.e. $i = 1, 4$) contribute to the computation. Other modalities (text and style reference) only provide global information but not local details. The latent feature map z^k produced by *Local PoE-Net* and the feature vector w produced by *Global PoE-Net* are fed to our local-global adaptive instance normalization (LG-AdaIN) layer,

$$\text{LG-AdaIN}(h^k, z^k, w) = \gamma_w \left(\gamma_{z^k} \frac{h^k - \mu(h^k)}{\sigma(h^k)} + \beta_{z^k} \right) + \beta_w \,, \tag{5}$$

where h^k is a feature map in the residual branch after convolution, $\mu(h^k)$ and $\sigma(h^k)$ are channel-wise mean and standard deviation. β_w, γ_w are feature vectors computed from w, while β_{z^k}, γ_{z^k} are feature maps computed from z^k. The LG-AdaIN layer can be viewed as a combination of AdaIN [17] and SPADE [36] that takes both a global feature vector and a spatially-varying feature map to modulate the activations.

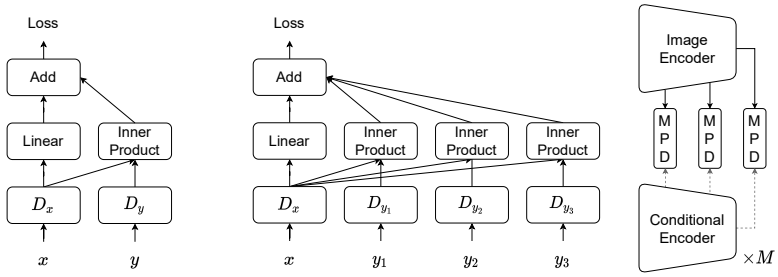
3.4 Multiscale multimodal projection discriminator

Our discriminator receives the image x and a set of conditions $\mathcal{Y}$ as inputs and produces a score $D(x, \mathcal{Y}) = \text{sigmoid}(f(x, \mathcal{Y}))$ indicating the realness of x given $\mathcal{Y}$. Under the GAN objective [11], the optimal solution of f is

$$f^*(x, \mathcal{Y}) = \underbrace{\log \frac{q(x)}{p(x)}}_{\text{unconditional term}} + \sum_{y_i \in \mathcal{Y}} \underbrace{\log \frac{q(y_i|x)}{p(y_i|x)}}_{\text{conditional term}} \,, \tag{6}$$

if we assume conditional independence of different modalities given x. The projection discriminator (PD) [33] proposes to use the inner product to estimate the conditional term. This implementation restricts the conditional term to be relatively simple, which imposes a good inductive bias that leads to strong empirical results. We propose a multimodal projection discriminator (MPD) that

[4] Except for the first layer that convolves a constant feature map and the last layer that convolves the previous output to synthesize the output.



(a) Projection discriminator (PD) (b) Multimodal PD (MPD) (c) Multiscale MPD (MMPD)

Fig. 5: Comparison between the standard projection discriminator and our proposed multiscale multimodal projection discriminator.

generalizes PD to our multimodal setting. As shown in Fig. 5a, the original PD first encodes both the image and the conditional input into a shared latent space. It then uses a linear layer to estimate the unconditional term from the image embedding and uses the inner product between the image embedding and the conditional embedding to estimate the conditional term. The unconditional term and the conditional term are summed to obtain the final discriminator logits. In our multimodal scenario, we simply encode each observed modality and add its inner product with the image embedding to the final loss (Fig. 5b)

$$f(x, \mathcal{Y}) = \text{Linear}(D_x(x)) + \sum_{y_i \in \mathcal{Y}} D_{y_i}^T(y_i) D_x(x)\,. \tag{7}$$

For spatial modalities such as segmentation and sketch, it is more effective to enforce their alignment with the image in multiple scales [29]. As shown in Fig. 5c, we encode the image and spatial modalities into feature maps of different resolutions and compute the MPD loss at each resolution. We compute a loss value at each location and resolution, and obtain the final loss by averaging first across locations then across resolutions. The resulting discriminator is named as the multiscale multimodal projection discriminator (MMPD) and detailed in Supplementary Material B.

3.5 Losses and training procedure

Latent regularization. Under the PoE assumption (Eq. (2)), the marginalized conditional latent distribution should match the unconditional prior:

$$\int p(z|y_i)p(y_i)dy_i = p(z|\varnothing) = p(z)\,. \tag{8}$$

To this end, we minimize the Kullback-Leibler (KL) divergence from the prior distribution $p(z)$ to the conditional latent distribution $p(z|y_i)$ at every resolution

$$\mathcal{L}_{\text{KL}} = \sum_{y_i \in \mathcal{Y}} \omega_i \sum_k \omega^k \mathbb{E}_{p(z^{<k}|y_i)} \left[D_{\text{KL}}(p(z^k|z^{<k}, y_i) || p(z^k|z^{<k})) \right], \tag{9}$$

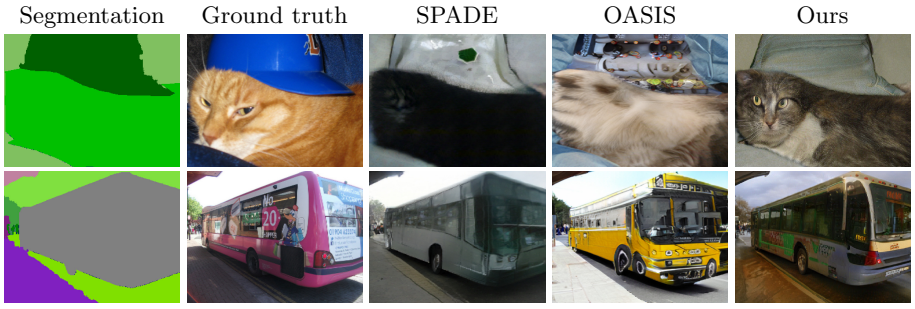


Fig. 6: Visual comparison of segmentation-to-image synthesis on MS-COCO 2017.

Fig. 7: Visual comparison of text-to-image synthesis on MS-COCO 2017.

where ω^k is a resolution-dependent rebalancing weight and ω_i is a modality-specific loss weight. We describe both weights in detail in Supplementary Material B.

The KL loss also reduces conditional mode collapse since it encourages the conditional latent distribution to be close to the prior and therefore have high entropy. From the perspective of information bottleneck [1], the KL loss encourages each modality to only provide the minimum information necessary to specify the conditional image distribution.

Contrastive losses. The contrastive loss has been widely adopted in representation learning [5,13] and more recently in image synthesis [35,61,28,12]. Given a batch of paired vectors $(\mathbf{u}, \mathbf{v}) = \{(u_i, v_i), i = 1, 2, ..., N\}$, the symmetric cross-entropy loss [64,38] maximizes the similarity of the vectors in a pair while keeping non-paired vectors apart

$$\mathcal{L}^{\text{ce}}(\mathbf{u}, \mathbf{v}) = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(\cos(u_i, v_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(u_i, v_j)/\tau)} - \frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(\cos(u_i, v_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(u_j, v_i)/\tau)}, \tag{10}$$

where τ is a temperature hyper-parameter. We use two kinds of pairs to construct two loss terms: the image contrastive loss and the conditional contrastive loss.

The *image contrastive loss* maximizes the similarity between a real image x and a fake image $\tilde{x}$ synthesized given the corresponding conditional inputs:

$$\mathcal{L}_{cx} = \mathcal{L}^{\text{ce}}(E_{\text{vgg}}(\mathbf{x}), E_{\text{vgg}}(\tilde{\mathbf{x}}))\,, \tag{11}$$

where E_{vgg} is a pretrained VGG [47] encoder. This loss serves a similar purpose to the widely used perceptual loss but has been found to perform better [35,61].

The *conditional contrastive loss* aims to better align images with the corresponding conditions. Specifically, the discriminator is trained to maximize the similarity between its embedding of a real image $\mathbf{x}$ and the conditional input $\mathbf{y}_i$.

$$\mathcal{L}^D_{cy} = \sum_{i=1}^{M} \mathcal{L}^{\text{ce}}(D_x(\mathbf{x}), D_{y_i}(\mathbf{y}_i))\,, \tag{12}$$

where D_x and D_{y_i} are two modules in the discriminator that extract features from x and y_i, respectively, as shown in Eq. (7) and Fig. 5b. The generator is trained with the same loss, but using the generated image $\tilde{\mathbf{x}}$ instead of the real image to compute the discriminator embedding,

$$\mathcal{L}^G_{cy} = \sum_{i=1}^{M} \mathcal{L}^{\text{ce}}(D_x(\tilde{\mathbf{x}}), D_{y_i}(\mathbf{y}_i))\,. \tag{13}$$

In practice, we only use the conditional contrastive loss for text since it consumes too much GPU memory to use the conditional contrastive loss for the other modalities, especially when the image resolution and batch size are large. A similar image-text contrastive loss is used in XMC-GAN [61], where they use a non-symmetric cross-entropy loss that only includes the first term in Eq. (10).

Full training objective. In summary, the generator loss $\mathcal{L}^G$ and the discriminator loss $\mathcal{L}^D$ can be written as

$$\mathcal{L}^G = \mathcal{L}^G_{\text{GAN}} + \mathcal{L}_{\text{KL}} + \lambda_1\mathcal{L}_{cx} + \lambda_2\mathcal{L}^G_{cy},\ \ \mathcal{L}^D = \mathcal{L}^D_{\text{GAN}} + \lambda_2\mathcal{L}^D_{cy} + \lambda_3\mathcal{L}_{\text{GP}}\,, \tag{14}$$

where $\mathcal{L}^G_{\text{GAN}}$ and $\mathcal{L}^D_{\text{GAN}}$ are non-saturated GAN losses [11], $\mathcal{L}_{\text{GP}}$ is the R_1 gradient penalty loss [31], and $\lambda_1, \lambda_2, \lambda_3$ are weights associated with the loss terms.

Modality dropout. By design, our generator, discriminator, and loss terms are able to handle missing modalities. We also find that randomly dropping out some input modalities before each training iteration further improves the robustness of the generator towards missing modalities at test time.

4 Experiments

We evaluate the proposed approach on several datasets, including MM-CelebA-HQ [57], MS-COCO 2017 [26] with COCO-Stuff annotations [3], and a proprietary dataset of landscape images. Images are labeled with all input modalities obtained from either manual annotation or pseudo-labeling methods. More details about datasets and the pseudo-labeling procedure are in Supplementary Material A.

Table 1: FID Comparison on MM-CelebA-HQ (1024×1024). We evaluate models conditioned on different modalities (from left to right: no conditions, text, segmentation, sketch, and all three modalities). The best scores are highlighted in bold

	Uncond	Text	Seg	Sketch	All
StyleGAN2 [23]	11.7	—	—	—	—
SPADE-Seg [36]	—	—	48.6	—	—
pSp-Seg [43]	—	—	44.1	—	—
SPADE-Sketch [36]	—	—	—	33.0	—
pSp-Sketch [43]	—	—	—	45.8	—
TediGAN [57]	—	38.4	45.1	45.1	45.1
PoE-GAN (Ours)	**10.5**	**10.1**	**9.9**	**9.9**	**8.3**

Table 2: Comparison on MS-COCO 2017 (256×256) using FID. We evaluate models conditioned on different modalities (from left to right: no conditions, text, segmentation, sketch, and all three modalities). The best scores are highlighted in bold

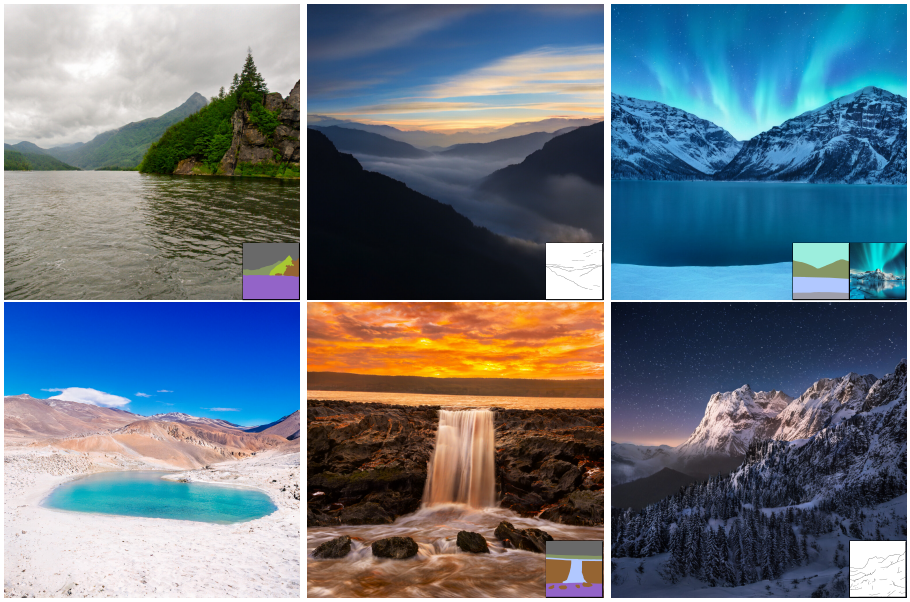
	Uncond	Text	Seg	Sketch	All
StyleGAN2 [23]	43.6	—	—	—	—
DF-GAN [51]	—	45.2	—	—	—
DM-GAN + CL [60]	—	29.9	—	—	—
SPADE-Seg [36]	—	—	22.1	—	—
VQGAN [8]	—	—	21.6	—	—
OASIS [45]	—	—	19.2	—	—
SPADE-Sketch [36]	—	—	—	63.7	—
PoE-GAN (Ours)	**43.4**	**20.5**	**15.8**	**25.5**	**13.6**

4.1 Main results

We compare PoE-GAN with a recent multimodal image synthesis method named TediGAN [57] and also with state-of-the-art approaches specifically designed for each modality. For text-to-image, we compare with DF-GAN [51] and DM-GAN + CL [60] on MS-COCO. Since the original models are trained on the 2014 split, we retrain their models on the 2017 split using the official code. For segmentation-to-image synthesis, we compare with SPADE [36], VQGAN [8], OASIS [45], and pSp [43]. For sketch-to-image synthesis, we compare with SPADE [36] and pSp [43]. We additionally compare with StyleGAN2 [23] in the unconditional setting. We use Clean-FID [37] for benchmarking due to its reported benefits over previous implementations of FID [15].[5]

Results on MM-CelebA-HQ and MS-COCO are summarized in Tab. 1 and Tab. 2, respectively. PoE-GAN obtains a much lower FID than TediGAN in all settings on MM-CelebA-HQ. In Supplementary Material C.3, we compare PoE-GAN with TediGAN in more detail and show that PoE-GAN is faster and more general than TediGAN. When conditioned on a single modality, PoE-GAN surprisingly outperforms the state-of-the-art method designed specifically for that modality on both datasets, although PoE-GAN is trained for a more general purpose. We note that PoE-GAN and TediGAN are trained on multiple modalities while other baselines are trained on an individual modality or unconditionally (StyleGAN2).

[5] As a result, the baseline scores differ slightly from those in the original papers.



A lake in the desert with mountains at a distance. *A waterfall in sunset.* *A mountain with pine trees in a starry winter night.*

Fig. 8: Examples of multimodal conditional image synthesis results produced by PoE-GAN trained on the 1024×1024 landscape dataset. We show the segmentation/sketch/style inputs on the bottom right of the generated images for the results in the first row. The results in the second row additionally leverage text inputs, which are shown below the corresponding generated images. Please zoom in for details.

In Supplementary Material C.4, we further show that PoE-GAN trained on a single modality always outperforms the multimodal-trained PoE-GAN when evaluated on that modality. This shows that the improvement of PoE-GAN over state-of-the-art unimodal image synthesis methods comes from our architecture and training scheme rather than additional annotations. In Figs. 6 and 7, we qualitatively compare PoE-GAN with previous segmentation-to-image and text-to-image methods on MS-COCO. We find that PoE-GAN produces images of much better quality and can synthesize realistic objects with complex structures, such as cats and stop signs. More qualitative comparisons are included in Supplementary Material C.5.

Multimodal generation examples. In Fig. 8, we show example images generated by our PoE-GAN using multiple input modalities on the landscape dataset. Our model is able to synthesize a wide range of landscapes in high resolution with photo-realistic details. More results are included in Supplementary Material C.5, where we additionally show that PoE-GAN can generate diverse images when given the same conditional inputs.

Table 3: Ablation study on MM-CelebA-HQ (256×256). The best scores are highlighted in bold and the second best ones are underlined

	Uncond	Text		Segmentation		Sketch		All	
Methods	FID↓	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
(a) Concatenation + dropout	29.3	26.4	0.33	19.4	0.22	9.2	0.11	**7.7**	0.11
(b) Ours w/o KL loss	29.1	26.6	0.35	18.1	0.21	**9.1**	0.10	**7.7**	0.12
(c) Ours w/o modality dropout	30.8	31.6	0.50	21.0	0.39	28.0	0.34	9.5	0.30
(d) Ours w/o MMPD	21.5	20.8	0.48	18.3	0.40	16.4	0.36	16.2	0.34
(e) Ours w/o image contrastive	15.4	14.5	0.55	13.5	**0.46**	10.2	**0.44**	9.5	**0.42**
(f) Ours w/o text contrastive	15.8	15.0	0.56	13.1	0.40	10.0	0.39	8.9	0.38
(g) Ours	**14.9**	**13.7**	**0.58**	**12.9**	0.43	9.9	0.37	8.5	0.35

Table 4: Ablation study on MS-COCO 2017 (64×64). The best scores are highlighted in bold and the second best ones are underlined

	Uncond	Text		Segmentation		Sketch		All	
Methods	FID↓	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
(a) Concatenation + dropout	59.1	30.7	0.40	20.4	0.16	36.1	0.27	**16.6**	0.12
(b) Ours w/o KL loss	59.3	30.5	0.39	21.5	0.16	33.0	0.27	16.9	0.12
(c) Ours w/o modality dropout	86.2	87.8	0.58	19.9	0.44	85.1	0.55	18.7	0.47
(d) Ours w/o MMPD	43.1	40.2	0.64	21.7	0.46	45.5	0.57	21.1	0.42
(e) Ours w/o image contrastive	**25.7**	**21.3**	0.64	18.0	**0.50**	37.9	**0.61**	18.5	**0.54**
(f) Ours w/o text contrastive	27.6	26.0	**0.66**	17.4	0.46	33.5	0.55	17.9	0.43
(g) Ours	26.6	22.2	0.65	**17.1**	0.47	**30.2**	0.58	17.1	0.44

4.2 Ablation studies

In Tabs. 3 and 4, we analyze the importance of different components of PoE-GAN. We use LPIPS [63] as an additional metric to evaluate the diversity of images conditioned on the same input. Specifically, we randomly sample two output images conditioned on the same input and report the average LPIPS distance between the two outputs. A higher LPIPS score indicates more diverse outputs.

First, we compare our product-of-experts generator (row (g)) with a baseline that simply concatenates the embedding of all modalities, while performing modality dropout (missing modality embeddings set to zero). As seen in row (a), this baseline only works well when all modalities are available and its FID significantly drops when some modalities are missing. Further, the output images have low diversity as indicated by the LPIPS score. This is not surprising as previous work has shown that conditional GANs are prone to mode collapse [66,18,59].

Row (b) of Tabs. 3 and 4 shows that the KL loss is important for training our model. Without it, our model suffers from low sample diversity and lack of robustness towards missing modalities, similar to the concatenation baseline described above. The variances of individual experts become near zero without the KL loss. The latent code z^k then becomes a deterministic weighted average of the mean of each expert, which is equivalent to concatenating all modality embeddings and projecting it with a linear layer. This explains why our model without the KL loss behaves similarly to the concatenation baseline. Row (c) shows that our modality dropout scheme is important for handling missing

Table 5: User study on text-to-image synthesis. Each column shows the percentage of users that prefer the image generated by our model over the baseline

	DF-GAN [51]	DM-GAN + CL [60]
Ours vs.	82.1%	72.9%

Table 6: User study on segmentation-to-image synthesis. Each column shows the percentage of users that prefer the image generated from our model over the baseline

	SPADE [36]	VQGAN [8]	OASIS [45]
Ours vs.	69%	66.7%	64.9%

modalities. Without it, the model tends to overly rely on the most informative modality, such as segmentation in MS-COCO.

To evaluate the proposed multiscale multimodal discriminator architecture, we replace MMPD with a discriminator that receives concatenated images and all conditional inputs. Row (d) shows that MMPD is much more effective than such a concatenation-based discriminator in all settings.

Finally in rows (e) and (f), we show that contrastive losses are useful but not essential. The image contrastive loss slightly improves FID in most settings, while the text contrastive loss improves FID for text-to-image synthesis.

4.3 User study

We conduct a user study to compare PoE-GAN with state-of-the-art text-to-image and segmentation-to-image synthesis methods on MS-COCO. We show users two images generated by different algorithms from the same conditional input and ask them which one is more realistic. As shown in Tab. 5 and Tab. 6, the majority of users prefer PoE-GAN over the baseline methods.

5 Conclusion

We introduce a multimodal conditional image synthesis model based on product-of-experts and show its effectiveness for converting an arbitrary subset of input modalities to an image satisfying all conditions. While empirically superior than the prior multimodal synthesis work, it also outperforms state-of-the-art unimodal conditional image synthesis approaches when conditioned on a single modality.

Acknowledgements. We thank Jan Kautz, David Luebke, Tero Karras, Timo Aila, and Zinan Lin for their feedback on the manuscript. We thank Daniel Gifford and Andrea Gagliano on their help on data collection.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: ICLR (2017) 9
2. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019) 2
3. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and stuff classes in context. In: CVPR (2018) 11
4. Chen, S.Y., Su, W., Gao, L., Xia, S., Fu, H.: DeepFaceDrawing: Deep generation of face images from sketches. ACM Transactions on Graphics (TOG) (2020) 2
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 9
6. Chen, W., Hays, J.: SketchyGAN: Towards diverse and realistic sketch to image synthesis. In: CVPR (2018) 2
7. Child, R.: Very deep VAEs generalize autoregressive models and can outperform them on images. In: ICLR (2021) 3
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021) 2, 11, 14
9. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: CVPR (2022) 3
10. Ghosh, A., Zhang, R., Dokania, P.K., Wang, O., Efros, A.A., Torr, P.H., Shechtman, E.: Interactive sketch & fill: Multiclass sketch-to-image translation. In: ICCV (2019) 2
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014) 2, 7, 10
12. Han, J., Shoeiby, M., Petersson, L., Armin, M.A.: Dual contrastive learning for unsupervised image-to-image translation. In: CVPR (2021) 9
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 9
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 6
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 12
16. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural computation (2002) 4
17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017) 7
18. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018) 13
19. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) 2
20. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018) 2
21. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS (2021) 2
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 2, 3

23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR (2020) 2, 11, 12
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014) 3
25. Kutuzova, S., Krause, O., McCloskey, D., Nielsen, M., Igel, C.: Multimodal variational autoencoders for semi-supervised learning: In defense of product-of-experts. arXiv preprint arXiv:2101.07240 (2021) 3
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 3, 11
27. Liu, M.Y., Huang, X., Yu, J., Wang, T.C., Mallya, A.: Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE (2021) 2
28. Liu, R., Ge, Y., Choi, C.L., Wang, X., Li, H.: DivCo: Diverse conditional image synthesis via contrastive generative adversarial network. In: CVPR (2021) 9
29. Liu, X., Yin, G., Shao, J., Wang, X., Li, H.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: NeurIPS (2019) 2, 3, 8
30. Maaløe, L., Fraccaro, M., Liévin, V., Winther, O.: BIVA: A very deep hierarchy of latent variables for generative modeling. NeurIPS (2019) 3
31. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for GANs do actually converge? In: ICML (2018) 10
32. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018) 2
33. Miyato, T., Koyama, M.: cGANs with projection discriminator. In: ICLR (2018) 3, 7
34. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: ICML (2017) 2
35. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV. Springer (2020) 9, 10
36. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 2, 7, 11, 14
37. Parmar, G., Zhang, R., Zhu, J.Y.: On buggy resizing libraries and surprising subtleties in FID calculation. arXiv preprint arXiv:2104.11222 (2021) 12
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 6, 9
39. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: CVPR (2016) 2
40. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021) 2
41. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016) 2
42. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML (2014) 3
43. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a StyleGAN encoder for image-to-image translation. In: CVPR (2021) 11, 12
44. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. In: CVPR (2017) 2
45. Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. In: ICLR (2020) 2, 11, 14

46. Shi, Y., Siddharth, N., Paige, B., Torr, P.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: NeurIPS (2019) 3
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 10
48. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. In: NeurIPS (2016) 3
49. Sutter, T.M., Daunhawer, I., Vogt, J.E.: Generalized multimodal ELBO. In: ICLR (2020) 3
50. Suzuki, M., Nakayama, K., Matsuo, Y.: Joint multimodal learning with deep generative models. In: ICLR workshop (2017) 3
51. Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.Y., Wu, F., Bao, B.: DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. arXiv preprint arXiv:2008.05865 (2020) 2, 11, 14
52. Vahdat, A., Kautz, J.: NVAE: A deep hierarchical variational autoencoder. In: NeurIPS (2020) 3
53. Vedantam, R., Fischer, I., Huang, J., Murphy, K.: Generative models of visually grounded imagination. In: ICLR (2018) 3
54. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: CVPR (2018) 2, 3
55. Williams, C.K., Agakov, F.V., Felderhof, S.N.: Products of gaussians. In: NeurIPS (2001) 5
56. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. In: NeurIPS (2018) 3
57. Xia, W., Yang, Y., Xue, J.H., Wu, B.: TediGAN: Text-guided diverse face image generation and manipulation. In: CVPR (2021) 3, 11
58. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018) 2
59. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks. In: ICLR (2019) 13
60. Ye, H., Yang, X., Takac, M., Sunderraman, R., Ji, S.: Improving text-to-image synthesis using contrastive learning. arXiv preprint arXiv:2107.02423 (2021) 2, 11, 14
61. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: CVPR (2021) 9, 10
62. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017) 2
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 12
64. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020) 9
65. Zhang, Z., Ma, J., Zhou, C., Men, R., Li, Z., Ding, M., Tang, J., Zhou, J., Yang, H.: M6-UFC: Unifying multi-modal controls for conditional image synthesis. In: NeurIPS (2021) 3
66. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NeurIPS (2017) 13
67. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR (2019) 2