Scaling up GANs for Text-to-Image Synthesis

Minguk Kang[1,3] Jun-Yan Zhu[2] Richard Zhang[3]
Jaesik Park[1] Eli Shechtman[3] Sylvain Paris[3] Taesung Park[3]

[1]POSTECH [2]Carnegie Mellon University [3]Adobe Research

Abstract

The recent success of text-to-image synthesis has taken the world by storm and captured the general public's imagination. From a technical standpoint, it also marked a drastic change in the favored architecture to design generative image models. GANs used to be the de facto choice, with techniques like StyleGAN. With DALL·E 2, autoregressive and diffusion models became the new standard for large-scale generative models overnight. This rapid shift raises a fundamental question: can we scale up GANs to benefit from large datasets like LAION? We find that naïvely increasing the capacity of the StyleGAN architecture quickly becomes unstable. We introduce GigaGAN, a new GAN architecture that far exceeds this limit, demonstrating GANs as a viable option for text-to-image synthesis. GigaGAN offers three major advantages. First, it is orders of magnitude faster at inference time, taking only 0.13 seconds to synthesize a 512px image. Second, it can synthesize high-resolution images, for example, 16-megapixel images in 3.66 seconds. Finally, GigaGAN supports various latent space editing applications such as latent interpolation, style mixing, and vector arithmetic operations.

1. Introduction

Recently released models, such as DALL·E 2 [53], Imagen [59], Parti [73], and Stable Diffusion [58], have ushered in a new era of image generation, achieving unprecedented levels of image quality and model flexibility. The now-dominant paradigms, diffusion models and autoregressive models, both rely on iterative inference. This is a double-edged sword, as iterative methods enable stable training with simple objectives but incur a high computational cost during inference.

Contrast this with Generative Adversarial Networks (GANs) [5,17,33,51], which generate images through a single forward pass and thus inherently efficient. While such models dominated the previous "era" of generative modeling, scaling them requires careful tuning of the network architectures and training considerations due to instabilities in the training procedure. As such, GANs have excelled at modeling single or multiple object classes, but scaling to complex datasets, much less an open world, has remained challenging. As a result, ultra-large models, data, and compute resources are now dedicated to diffusion and autoregressive models. In this work, we ask – *can GANs continue to be scaled up and potentially benefit from such resources, or have they plateaued? What prevents them from further scaling, and can we overcome these barriers?*

We first experiment with StyleGAN2 [34] and observe that simply scaling the backbone causes unstable training. We identify several key issues and propose techniques to stabilize the training while increasing the model capacity. First, we effectively scale the generator's capacity by retaining a bank of filters and taking a sample-specific linear combination. We also adapt several techniques commonly used in the diffusion context and confirm that they bring similar benefits to GANs. For instance, interleaving both self-attention (image-only) and cross-attention (image-text) with the convolutional layers improves performance.

Furthermore, we reintroduce multi-scale training, finding a new scheme that improves image-text alignment and low-frequency details of generated outputs. Multi-scale training allows the GAN-based generator to use parameters in low-resolution blocks more effectively, leading to better image-text alignment and image quality. After careful tuning, we achieve stable and scalable training of a one-billion-parameter GAN (GigaGAN) on large-scale datasets, such as LAION2B-en [63]. Our results are shown in Figure 1.

In addition, our method uses a multi-stage approach [12, 76]. We first generate at 64×64 and then upsample to 512×512. These two networks are modular and robust enough to be used in a plug-and-play fashion. We show that our text-conditioned GAN-based upsampling network can be used as an efficient, higher-quality upsampler for a base diffusion model such as DALL·E 2, despite never having seen diffusion images at training time (Figures 1).

Together, these advances enable our GigaGAN to go far beyond previous GANs: 36× larger than StyleGAN2 [34] and 6× larger than StyleGAN-XL [62] and

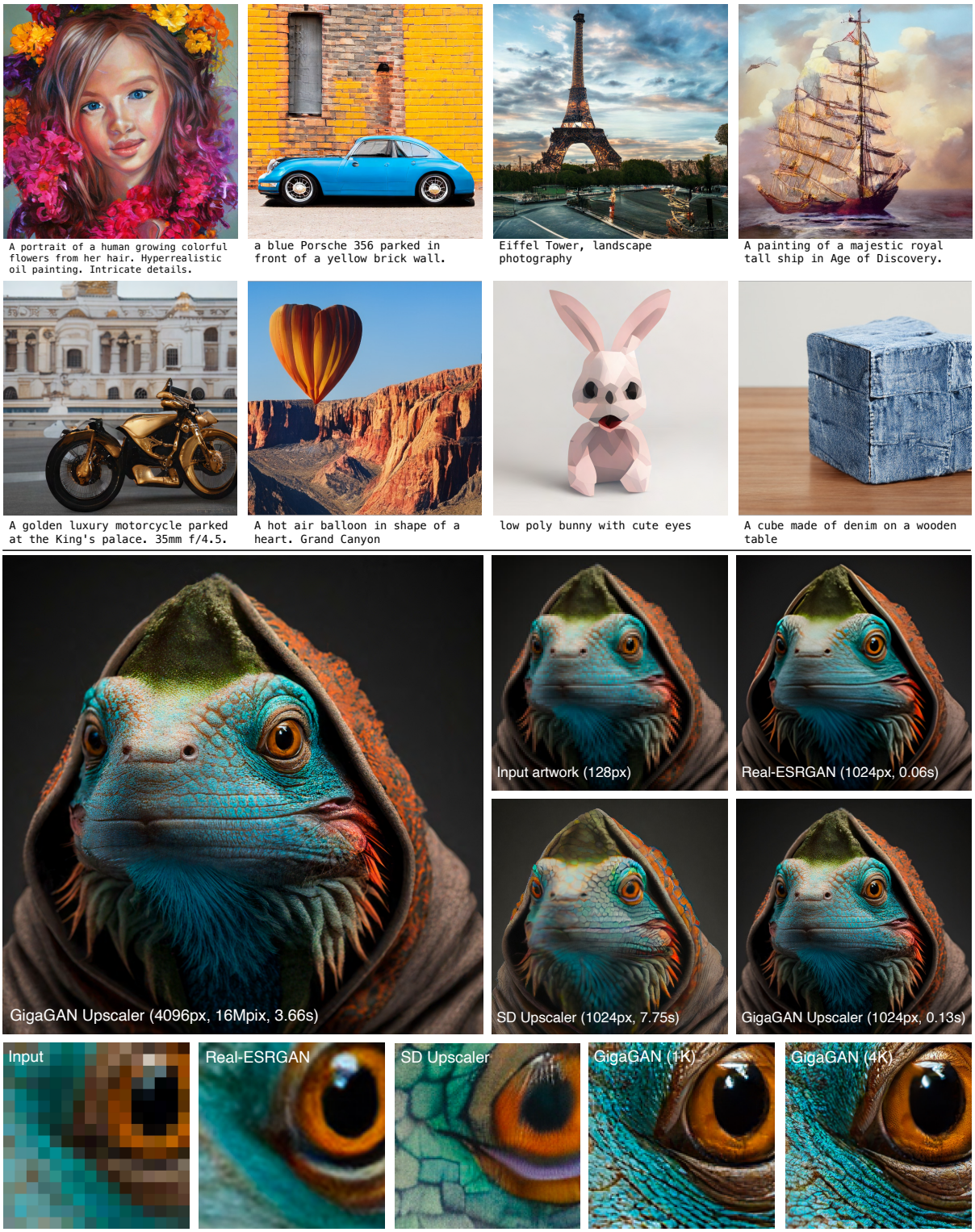

Figure 1. Our model, GigaGAN, shows GAN frameworks can also be scaled up for general text-to-image synthesis and super-resolution tasks, generating a 512px output at an interactive speed of 0.13s, and 4096px within 3.7s. Selected examples at 2K resolution (text-to-image synthesis) and 1k or 4k resolutions (super-resolution) are shown. For the super-resolution task, we use the caption of "Portrait of a colored iguana dressed in a hoodie." and compare our model with the text-conditioned upscaler of Stable Diffusion [57] and unconditional Real-ESRGAN [26]. Please zoom in for more details. See our arXiv paper and website for more uncurated comparisons.

XMC-GAN [75]. While our 1B parameter count is still lower than the largest synthesis models, such as DALL·E 2 (5.5B), and Parti (20B), we have not yet observed a quality saturation regarding the model size. GigaGAN achieves a zero-shot FID of 9.09 on COCO2014, lower than the FID of DALL·E 2, Parti-750M, and Stable Diffusion.

Furthermore, GigaGAN has three major practical advantages compared to diffusion and autoregressive models. First, it is orders of magnitude faster, generating a 512px image in 0.13 seconds. Second, it can synthesize ultra high-res images at 4k resolution in 3.66 seconds. Third, it is endowed with a controllable, latent vector space that lends itself to well-studied controllable image synthesis applications, such as prompt mixing (Figure 4), style mixing (Figure 5), and prompt interpolation (Figures A7 and A8).

In summary, our model is the first GAN-based method that successfully trains a billion-scale model on billions of real-world complex Internet images. This suggests that GANs are still a viable option for text-to-image synthesis and should be considered for future aggressive scaling. Please visit our website for additional results.

2. Related Works

Text-to-image synthesis. Generating a realistic image given a text description, explored by early works [42, 85], is a challenging task. A common approach is text-conditional GANs [55, 56, 67, 71, 76, 83] on specific domains [68] and datasets with a closed-world assumption [41]. With the development of diffusion models [13, 21], autoregressive (AR) transformers [10], and large-scale language encoders [50, 52], text-to-image synthesis has shown remarkable improvement on an open-world of arbitrary text descriptions. GLIDE [46], DALL·E 2 [53], and Imagen [59] are representative diffusion models that show realistic outputs with the aid of a pretrained language encoder [50, 52]. AR models, such as DALL·E [54], Make-A-Scene [16], CogView [14, 15], and Parti [73] also achieve amazing results. While these models exhibit unprecedented image synthesis ability, they require time-consuming iterative processes to achieve high-quality image sampling.

GAN-based image synthesis. GANs [17] have been one of the primary families of generative models for natural image synthesis. As the sampling quality and diversity of GANs improve [31–34, 36, 51, 60], GANs have been deployed to various computer vision and graphics applications, such as text-to-image synthesis [55], image-to-image translation [23, 27, 37, 47, 48, 82], and image editing [1, 6, 49, 81]. Notably, StyleGAN-family models [32, 34] have shown impressive ability in image synthesis tasks for single-category domains [1, 25, 49, 70, 84]. Other works have explored class-conditional GANs [5, 29, 62, 74, 79] on datasets with a fixed set of object categories.

In this paper, we change the data regimes from single- or multi-categories datasets to extremely data-rich situations. We make the first expedition toward training a large-scale GAN for text-to-image generation on a vast amount of web-crawled text and image pairs, such as LAION2B-en [63] and COYO-700M [7]. Existing GAN-based text-to-image synthesis models [39, 55, 67, 71, 75, 76, 83] are trained on relatively small datasets, such as CUB-200 (12k training pairs), MSCOCO (82k) and LN-OpenImages (507k). Also, those models are evaluated on associated validation datasets, which have not been validated to perform large-scale text-image synthesis like diffusion or AR models.

Concurrent with our method, StyleGAN-T [61] and GALIP [66] share similar goals and make complementary insights to ours.

3. Method

We train a generator $G(\mathbf{z}, \mathbf{c})$ to predict an image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ given a latent code $\mathbf{z} \sim \mathcal{N}(0, 1) \in \mathbb{R}^{128}$ and text-condition $\mathbf{c}$. We use a discriminator $D(\mathbf{x}, \mathbf{c})$ to judge the realism of the fake image, as compared to a sample from the training database $\mathcal{D}$, which contains image-text pairs.

Although GANs [5, 31, 33] can successfully generate realistic images on single- and multi-category datasets [11, 33, 72], open-ended text-conditioned synthesis on Internet images remains challenging. We hypothesize that the current limitation stems from its reliance on convolutional layers. That is, the same convolution filters are challenged to model the general image synthesis function for all text conditioning across all locations of the image. In this light, we seek to inject more expressivity into our parameterization by dynamically selecting convolution filters based on the input conditioning and by capturing long-range dependence via the attention mechanism.

Below, we discuss our key contributions to making ConvNets more expressive (Section 3.1), followed by our designs for the generator (Section 3.2) and discriminator (Section 3.3). Lastly, we introduce a new, fast GAN-based upsampler model that can improve the inference quality and speed of our method and diffusion models such as Imagen [59] and DALL·E 2 [53] (Section 3.4).

3.1. Modeling complex contextual interaction

Baseline StyleGAN generator. We base our architecture off the conditional version of StyleGAN2 [34], comprised of two networks $G = \widetilde{G} \circ M$. The mapping network $\mathbf{w} = M(\mathbf{z}, \mathbf{c})$ maps the inputs into a "style" vector $\mathbf{w}$, which modulates a series of upsampling convolutional layers in the synthesis network $\widetilde{G}(\mathbf{w})$ to map a learned constant tensor to an output image $\mathbf{x}$. Convolution is the main engine to generate all output pixels, with the $\mathbf{w}$ vector as the only source of information to model conditioning.

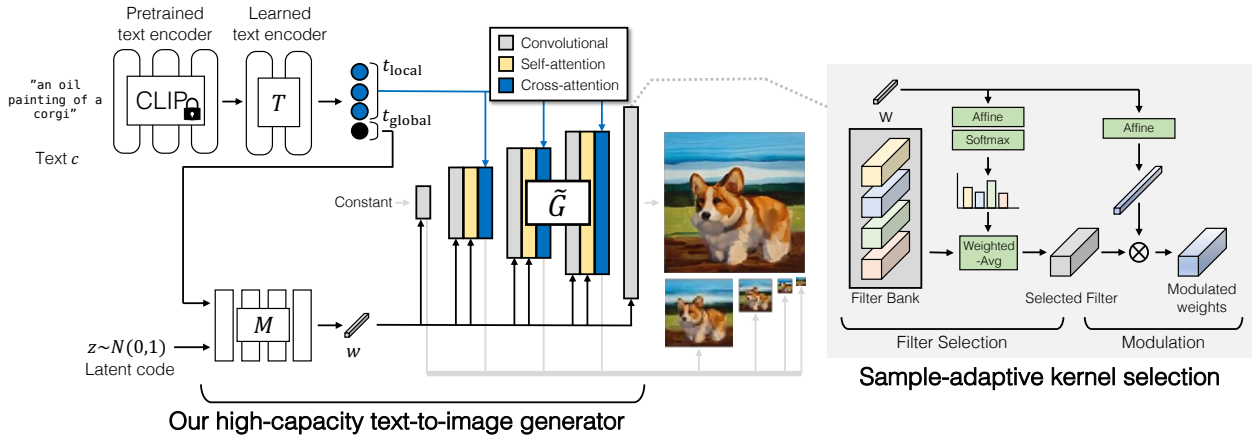

Figure 2. **GigaGAN's text-to-image generator.** First, we extract text embeddings using a pretrained CLIP model and a learned encoder T. The local text descriptors are fed to the generator using cross-attention. The global text descriptor, along with a latent code $\mathbf{z}$, is fed to a mapping network M to produce style code $\mathbf{w}$. The style code modulates the main generator using our style-adaptive kernel selection, shown on the right. The generator outputs an image pyramid by converting the intermediate features into RGB images. To achieve higher capacity, we use multiple attention and convolution layers at each scale. We also use a separate upsampler, that is not shown in this diagram.

Sample-adaptive kernel selection. To handle the highly diverse distribution of internet images, we aim to increase the capacity of convolution kernels. However, increasing the width of the convolution layers becomes too demanding, as the same operation is repeated across all locations.

We propose an efficient way to enhance the expressivity of convolutional kernels by creating them on-the-fly based on the text conditioning, as illustrated in Figure 2 (right). In this scheme, we instantiate a bank of N filters $\{\mathbf{K}_i \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}} \times K \times K}\}_{i=1}^{N}$, instead of one, that takes a feature $\mathbf{f} \in \mathbb{R}^{C_{\text{in}}}$ at each layer. The style vector $\mathbf{w} \in \mathbb{R}^{d}$ then goes through an affine layer $[W_{\text{filter}}, b_{\text{filter}}] \in \mathbb{R}^{(d+1)\times N}$ to predict a set of weights to average across the filters, to produce an aggregated filter $\mathbf{K} \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}} \times K \times K}$.

$$\mathbf{K} = \sum_{i=1}^{N} \mathbf{K}_i \cdot \text{softmax}\left(W_{\text{filter}}^{\top}\mathbf{w} + b_{\text{filter}}\right)_i \qquad (1)$$

The filter is then used in the regular convolution pipeline of StyleGAN2, with the second affine layer $[W_{\text{mod}}, b_{\text{mod}}] \in \mathbb{R}^{(d+1)\times C_{\text{in}}}$ for weight (de-)modulation [34].

$$g_{\text{adaconv}}(\mathbf{f}, \mathbf{w}) = \left((W_{\text{mod}}^{\top}\mathbf{w} + b_{\text{mod}}) \otimes \mathbf{K}\right) * \mathbf{f}, \qquad (2)$$

where $\otimes$ and $*$ represent (de-)modulation and convolution.

At a high level, the softmax-based weighting can be viewed as a differentiable filter selection process based on input conditioning. Furthermore, since the filter selection process is performed only once at each layer, the selection process is much faster than the actual convolution, decoupling compute complexity from the resolution. Our method shares a spirit with dynamic convolutions [19, 28, 65, 69] in that the convolution filters dynamically change per sample, but differs in that we explicitly instantiate a larger filter bank and select weights based on a separate pathway conditional on the $\mathbf{w}$-space of StyleGAN.

Interleaving attention with convolution. Since the convolutional filter operates within its receptive field, it cannot contextualize itself in relationship to distant parts of the images. One way to incorporate such long-range relationships is using attention layers $g_{\text{attention}}$. While recent diffusion-based models [13, 22, 58] have commonly adopted attention mechanisms, StyleGAN architectures are predominantly convolutional with the notable exceptions such as BigGAN [5], GANformer [24], and ViTGAN [38].

We aim to improve the performance of StyleGAN by integrating attention layers with the convolutional backbone. However, simply adding attention layers to StyleGAN often results in training collapse, possibly because the dot-product self-attention is not Lipschitz, as pointed out by Kim et al. [35]. As the Lipschitz continuity of discriminators has played a critical role in stable training [2, 18, 43, 44], we use the L2-distance instead of the dot product as the attention logits to promote Lipschitz continuity [35], similar to ViTGAN [38].

To further improve performance, we find it crucial to match the architectural details of StyleGAN, such as equalized learning rate [31] and weight initialization from a unit normal distribution. We scale down the L2 distance logits to roughly match the unit normal distribution at initialization and reduce the residual gain from the attention layers. We further improve stability by tying the key and query matrix [38], and applying weight decay.

In the synthesis network $\widetilde{G}$, the attention layers are interleaved with each convolutional block, leveraging the style vector $\mathbf{w}$ as an additional token. At each attention block, we add a separate cross-attention mechanism $g_{\text{cross-attention}}$ to attend to individual word embeddings [3]. We use each input feature tensor as the query, and the text embeddings as the key and value of the attention mechanism.

3.2. Generator design

Text and latent-code conditioning. First, we extract the text embedding from the prompt. Previous works [54, 59] have shown that leveraging a strong language model is essential for producing strong results. To do so, we tokenize the input prompt (after padding it to $C = 77$ words, following best practices [54, 59]) to produce conditioning vector $\mathbf{c} \in \mathbb{R}^{C\times 1024}$, and take the features from the penultimate layer [59] of a frozen CLIP feature extractor [50]. To allow for additional flexibility, we apply additional attention layers T on top to process the word embeddings before passing them to the MLP-based mapping network. This results in text embedding $\mathbf{t} = T(\mathcal{E}_{\text{txt}}(\mathbf{c})) \in \mathbb{R}^{C\times 1024}$. Each component $\mathbf{t}_i$ of $\mathbf{t}$ captures the embedding of the i^{th} word in the sentence. We refer to them as $\mathbf{t}_{\text{local}} = \mathbf{t}_{\{1:C\}\setminus\text{EOT}} \in \mathbb{R}^{(C-1)\times 1024}$. The EOT ("end of text") component of $\mathbf{t}$ aggregates global information, and is called $\mathbf{t}_{\text{global}} \in \mathbb{R}^{1024}$. We process this global text descriptor, along with the latent code $\mathbf{z} \sim \mathcal{N}(0, 1)$, via an MLP mapping network to extract the style $\mathbf{w} = M(\mathbf{z}, \mathbf{t}_{\text{global}})$.

$$\begin{aligned}(\mathbf{t}_{\text{local}}, \mathbf{t}_{\text{global}}) &= T(\mathcal{E}_{\text{txt}}(\mathbf{c})),\\ \mathbf{w} &= M(\mathbf{z}, \mathbf{t}_{\text{global}}).\end{aligned} \tag{3}$$

Different from the original StyleGAN, we use both the text-based style code $\mathbf{w}$ to modulate the synthesis network $\widetilde{G}$ and the word embeddings $\mathbf{t}_{\text{local}}$ as features for cross-attention.

$$\mathbf{x} = \widetilde{G}(\mathbf{w}, \mathbf{t}_{\text{local}}). \tag{4}$$

Similar to earlier works [42, 53, 59], the text-image alignment visually improves with cross-attention.

Synthesis network. Our synthesis network consists of a series of upsampling convolutional layers, with each layer enhanced with the adaptive kernel selection (Equation 1) and followed by our attention layers.

$$\mathbf{f}_{\ell+1} = g^{\ell}_{\text{xa}}(g^{\ell}_{\text{attn}}(g^{\ell}_{\text{adaconv}}(\mathbf{f}_{\ell}, \mathbf{w}), \mathbf{w}), \mathbf{t}_{\text{local}}), \tag{5}$$

where g^{ℓ}_{xa}, g^{ℓ}_{attn}, and $g^{\ell}_{\text{adaconv}}$ denote the l-th layer of cross-attention, self-attention, and weight (de-)modulation layers. We find it beneficial to increase the depth of the network by adding more blocks at each layer. In addition, our generator outputs a multi-scale image pyramid with $L = 5$ levels, instead of a single image at the highest resolution, similar to MSG-GAN [30] and AnycostGAN [40]. We refer to the pyramid as $\{\mathbf{x}_i\}_{i=0}^{L-1} = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_4\}$, with spatial resolutions $\{S_i\}_{i=0}^{L-1} = \{64, 32, 16, 8, 4\}$, respectively. The base level $\mathbf{x}_0$ is the output image $\mathbf{x}$. Each image of the pyramid is independently used to compute the GAN loss, as discussed in Section 3.3. We follow the findings of StyleGAN-XL [62] and turn off the style mixing and path length regularization [34]. We include more training details in our arXiv version.

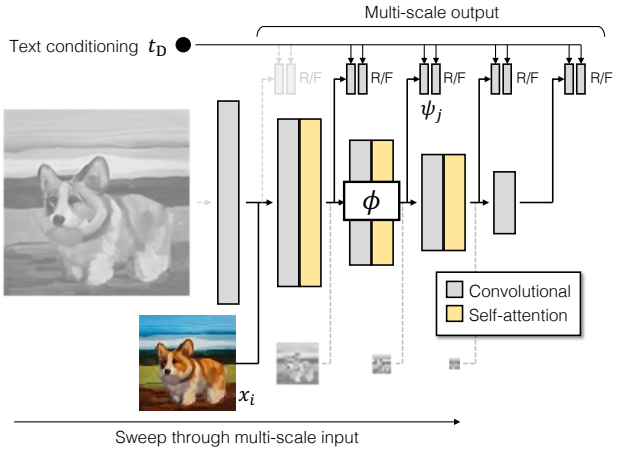

Figure 3. **Our discriminator** consists of two branches for processing the image and the text conditioning t_D. The text branch processes the text similar to the generator (Figure 2). The image branch receives an image pyramid and makes independent predictions for each image scale. Moreover, the predictions are made at all subsequent scales of the downsampling layers, making it a *multi-scale input, multi-scale output* (MS-I/O) discriminator.

3.3. Discriminator design

As shown in Figure 3, our discriminator consists of separate branches for processing text with the function $\mathbf{t}_D$ and images with function ϕ. The prediction of real vs. fake is made by comparing the features from the two branches using function ψ. We introduce a new way of making predictions on multiple scales. Finally, we use additional CLIP and Vision-Aided GAN losses [36] to improve stability.

Text conditioning. First, to incorporate conditioning into discriminators, we extract text descriptor $\mathbf{t}_{\text{D}}$ from text $\mathbf{c}$. Similar to the generator, we apply a pretrained text encoder, such as CLIP [50], followed by a few learnable attention layers. In this case, we only use the global descriptor.

Multiscale image processing. We observe that the early, low-resolution layers of the generator become inactive, using small dynamic ranges irrespective of the provided prompts. StyleGAN2 [34] also observes this phenomenon, concluding that the network relies on the high-resolution layers, as the model size increases. As recovering performance in low frequencies, which contains complex structure information, is crucial, we redesign the model architecture to provide training signals across multiple scales.

Recall the generator produces a pyramid $\{\mathbf{x}_i\}_{i=0}^{L-1}$, with the full image $\mathbf{x}_0$ at the pyramid base. MSG-GAN [30] improves performance by making a prediction on the entire pyramid at once, enforcing consistency across scales. However, in our large-scale setting, this harms stability, as this limits the generator from making adjustments to its initial low-res output.

Instead, we process each level of the pyramid *independently*. As shown in Figure 3, each level $\mathbf{x}_i$ makes a real/fake prediction at multiple scales $i < j \leq L$. For example, the full $\mathbf{x}_0$ makes predictions at $L = 5$ scales, the next level $\mathbf{x}_1$ makes predictions at 4 scales, and so on. In total, our discriminator produces $\frac{L(L-1)}{2}$ predictions, supervising multi-scale generations at multiple scales.

To extract features at different scales, we define a feature extractor $\phi_{i \to j} : \mathbb{R}^{X_i \times X_i \times 3} \to \mathbb{R}^{X_j^D \times X_j^D \times C_j}$. Practically, each sub-network $\phi_{i \to j}$ is a subset of full $\phi \triangleq \phi_{0 \to L}$, with $i > 0$ indicating late entry and $j < L$ indicating early exit. Each layer in ϕ consists of self-attention, followed by convolution with a stride 2. The final layer flattens the spatial extent into a 1×1 tensor, producing output resolutions at $\{X_j^D\} = \{32, 16, 8, 4, 1\}$. This allows us to inject lower-resolution images on pyramid into intermediate layers [31]. As we use a shared feature extractor across different levels and most of the added predictions are made at low resolutions, the increased computation overhead is manageable.

Multi-scale input, multi-scale output adversarial loss. In total, our training objective consists of discriminator losses, along with our proposed matching loss, to encourage the discriminator to take into account the conditioning:

$$\mathcal{V}_{\text{MS-I/O}}(G, D) = \sum_{i=0}^{L-1} \sum_{j=1}^{L} \mathcal{V}_{\text{GAN}}(G_i, D_{ij}) + \mathcal{V}_{\text{match}}(G_i, D_{ij}), \tag{6}$$

where $\mathcal{V}_{\text{GAN}}$ is the standard, non-saturating GAN loss [17]. To compute the discriminator output, we train predictor ψ, which uses text feature $\mathbf{t}_{\text{D}}$ to modulate image features $\phi(\mathbf{x})$:

$$D_{ij}(\mathbf{x}, \mathbf{c}) = \psi_j(\phi_{i \to j}(\mathbf{x}_i), \mathbf{t}_D) + \text{Conv}_{1\times1}(\phi_{i \to j}(\mathbf{x}_i)), \tag{7}$$

where ψ_j is implemented as a 4-layer 1×1 modulated convolution, and $\text{Conv}_{1\times1}$ is added as a skip connection to explicitly maintain an unconditional prediction branch [45].

Matching-aware loss. The previous GAN terms measure how closely the image $\mathbf{x}$ matches the conditioning $\mathbf{c}$, as well as how realistic $\mathbf{x}$ looks, irrespective of conditioning. However, during early training, when artifacts are obvious, the discriminator tends to make a decision independent of conditioning and hesitates to account for the conditioning.

To enforce the discriminator to incorporate conditioning, we match $\mathbf{x}$ with a random, independently sampled condition $\hat{\mathbf{c}}$, and present them as a fake pair:

$$\begin{aligned} \mathcal{V}_{\text{match}} = \mathbb{E}_{\mathbf{x},\mathbf{c},\hat{\mathbf{c}}} \big[&\log(1 + \exp(D(\mathbf{x}, \hat{\mathbf{c}}))) \\ &+ \log(1 + \exp(D(G(\mathbf{c}), \hat{\mathbf{c}}))) \big], \end{aligned} \tag{8}$$

where $(\mathbf{x}, c)$ and $\hat{\mathbf{c}}$ are separately sampled from p_{data}. This loss has previously been explored in text-to-image GAN works [55, 76], except we find that enforcing the Matching-aware loss on generated images from G, as well real images $\mathbf{x}$, leads to clear gains in performance (Table A2).

CLIP contrastive loss. We further leverage off-the-shelf pretrained models as a loss function [36, 60, 64]. In particular, we enforce the generator to produce outputs that are identifiable by the pre-trained CLIP image and text encoders [50], $\mathcal{E}_{\text{img}}$ and $\mathcal{E}_{\text{txt}}$, in the contrastive cross-entropy loss that was used to train them originally.

$$\mathcal{L}_{\text{CLIP}} = \mathbb{E}_{\{\mathbf{c}_n\}} \Big[-\log \frac{\exp(\mathcal{E}_{\text{img}}(G(\mathbf{c}_0))^\top \mathcal{E}_{\text{txt}}(\mathbf{c}_0))}{\sum_n \exp(\mathcal{E}_{\text{img}}(G(\mathbf{c}_0))^\top \mathcal{E}_{\text{txt}}(\mathbf{c}_n))} \Big], \tag{9}$$

where $\{\mathbf{c}_n\} = \{\mathbf{c}_0, \dots\}$ are sampled captions from the training data.

Vision-Aided adversarial loss. Lastly, we build an additional discriminator that uses the CLIP model as a backbone, known as Vision-Aided GAN [36]. We freeze the CLIP image encoder, extract features from the intermediate layers, and process them through a simple network with 3×3 conv layers to make real/fake predictions.

We also incorporate conditioning through modulation, as in Equation 7. To stabilize training, we also add a fixed random projection layer, as proposed by Projected GAN [60]. We refer to this as $\mathcal{L}_{\text{Vision}}(G)$ (omitting the learnable discriminator parameters for clarity).

Our final objective is $\mathcal{V}(G, D) = \mathcal{V}_{\text{MS-I/O}}(G, D) + \mathcal{L}_{\text{CLIP}}(G) + \mathcal{L}_{\text{Vision}}(G)$, with weighting between the terms specified in Table A1.

3.4. GAN-based upsampler

Furthermore, GigaGAN framework can be easily extended to train a text-conditioned super-resolution model, capable of upsampling the outputs of the base GigaGAN generator to obtain high-resolution images at 512px or 4k resolution. By training our pipeline in two separate stages, we can afford a higher capacity 64px base model within the same computational resources.

In the upsampler, the synthesis network is rearranged to an asymmetric U-Net architecture, which processes the 64px input through 3 downsampling residual blocks, followed by 6 upsampling residual blocks with attention layers to produce the 512px image. There exist skip connections at the same resolution, similar to CoModGAN [78]. The model is trained with the same losses as the base model, as well as the LPIPS Perceptual Loss [77] with respect to the ground truth high-resolution image. Vision-Aided GAN is not used for the upsampler. During training and inference time, we apply moderate Gaussian noise augmentation to reduce the gap between real and GAN-generated images.

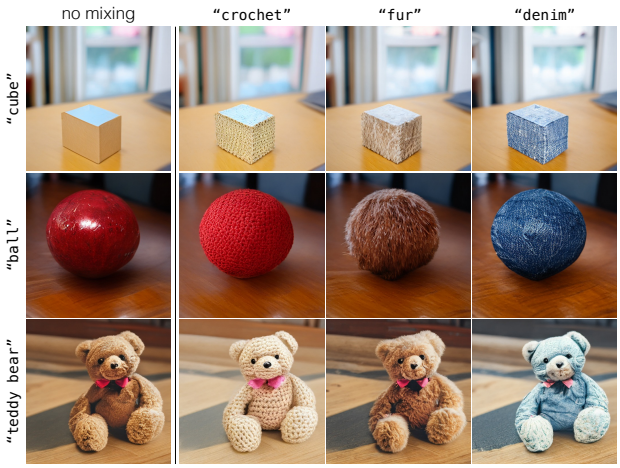

Figure 4. **Prompt mixing**. GigaGAN can directly control the style with text prompts. Here we generate three outputs using the prompts "a X on tabletop", shown in the "no mixing" column. Then we re-compute the text embeddings $\mathbf{t}$ and the style codes $\mathbf{w}$ using the new prompts "a X with the texture of Y on tabletop", such as "a cube with the texture of crochet on tabletop", and apply them to the second half layers of the generator, achieving layout-preserving fine style control. Cross-attention mechanism automatically localizes the style to the object of interest.

Our GigaGAN framework becomes particularly effective for the super-resolution task compared to the diffusion-based models, which cannot afford as many sampling steps as the base model at high resolution. The LPIPS regression loss also provides a stable learning signal. We believe that our GAN upsampler can serve as a drop-in replacement for the super-resolution stage of other generative models.

4. Experiments

Systematic, controlled evaluation of large-scale text-to-image synthesis tasks is difficult, as most existing models are not publicly available. Training a new model from scratch would be prohibitively costly, even if the training code were available. Still, we compare our model to recent text-to-image models, such as Imagen [59], Latent Diffusion Models (LDM) [58], Stable Diffusion [57], and Parti [73], based on the available information, while acknowledging considerable differences in the training dataset, number of iterations, batch size, and model size. In addition to text-to-image results, we evaluate our model on ImageNet class-conditional generation and unconditional super-resolution in our arXiv, for an apples-to-apples comparison with other methods at a more controlled setting.

For quantitative evaluation, we mainly use the Fréchet Inception Distance (FID) [20] for measuring the realism of the output distribution and the CLIP score for evaluating the image-text alignment. All our models are trained and evaluated on A100 GPUs. For more information about training and evaluation of GigaGAN, please refer to the arXiv version.

Figure 5. **Style mixing**. Our GAN-based architecture retains a disentangled latent space, enabling us to blend the coarse style of one sample with the fine style of another. All outputs are generated with the prompt "A Toy sport sedan, CG art." The corresponding latent codes are spliced together to produce a style-swapping grid.

4.1. Text-to-image synthesis

We proceed to train a larger model by increasing the capacity of the base generator and upsampler to 652.5M and 359.1M, respectively. This results in an unprecedented size of GAN model, with a total parameter count of 1.0B. Table 1 compares the performance of our end-to-end pipeline to various text-to-image generative models [4, 9, 46, 53, 54, 57–59, 73, 80]. Note that there exist differences in the training dataset, the pretrained text encoders, and even image resolutions. For example, GigaGAN initially synthesizes 512px images, which are resized to 256px before evaluation.

Table 1 shows that GigaGAN exhibits a lower FID than DALL·E 2 [53], Stable Diffusion [57], and Parti-750M [73]. While our model can be optimized to better match the feature distribution of real images than existing models, the quality of the generated images is not necessarily better (see our arXiv for more samples). We acknowledge that this may represent a corner case of zero-shot FID on COCO2014 dataset and suggest that further research on a better evaluation metric is necessary to improve text-to-image models. Nonetheless, we emphasize that GigaGAN is the first GAN model capable of synthesizing promising images from arbitrary text prompts and exhibits competitive zero-shot FID with other text-to-image models.

4.2. Super-resolution for large-scale image synthesis

We separately evaluate the performance of the GigaGAN upsampler in a text-conditional upsampling task. We combine the Stable Diffusion [57] 4x Upscaler and 2x Latent Upscaler to establish an 8x upscaling model (SD Upscaler).

Table 1. **Comparison to recent text-to-image models**. Model size, total images seen during training, COCO FID-30k, and inference speed of text-image models. $*$ denotes that the model has been evaluated by us. GigaGAN achieves a lower FID than DALL·E 2 [53], Stable Diffusion [57], and Parti-750M [73], while being much faster than competitive methods. GigaGAN and SD-v1.5 require 4,783 and 6,250 A100 GPU days, and Imagen and Parti need approximately 4,755 and 320 TPUv4 days for training.

	Model	Type	# Param.	# Images	FID-30k ↓	Inf. time
256	DALL·E [54]	Diff	12.0B	1.54B	27.50	-
	GLIDE [46]	Diff	5.0B	5.94B	12.24	15.0s
	LDM [58]	Diff	1.5B	0.27B	12.63	9.4s
	DALL·E 2 [53]	Diff	5.5B	5.63B	10.39	-
	Imagen [59]	Diff	3.0B	15.36B	7.27	9.1s
	eDiff-I [4]	Diff	9.1B	11.47B	6.95	32.0s
	Parti-750M [73]	AR	750M	3.69B	10.71	-
	Parti-3B [73]	AR	3.0B	3.69B	8.10	6.4s
	Parti-20B [73]	AR	20.0B	3.69B	7.23	-
	LAFITE [80]	GAN	75M	-	26.94	0.02s
512	SD-v1.5* [57]	Diff	0.9B	3.16B	9.62	2.9s
	Muse-3B [9]	AR	3.0B	0.51B	7.88	1.3s
	GigaGAN	GAN	1.0B	0.98B	9.09	0.13s

Table 2. **Text-conditioned 128→1024 super-resolution** on random 10K LAION samples, compared against unconditional Real-ESRGAN [26] and Stable Diffusion Upscaler [57]. GigaGAN enjoys the fast speed of a GAN-based model while achieving better FID, patch-FID [8], CLIP score, and LPIPS [77].

Model	# Param.	Inf. time	FID-10k ↓	pFID ↓	CLIP ↑	LPIPS↓
Real-ESRGAN [26]	17M	0.06s	8.60	22.8	0.314	0.363
SD Upscaler [57]	846M	7.75s	9.39	41.3	0.316	0.523
GigaGAN	693M	0.13s	1.54	8.90	0.322	0.274

We also use the unconditional Real-ESRGAN [26] as another baseline. Table 2 measures the performance of the upsampler on random 10K images from the LAION dataset and shows that our GigaGAN upsampler significantly outperforms the other upsamplers in realism scores (FID and patch-FID [8]), text alignment (CLIP score) and closeness to the ground truth (LPIPS [77]).

4.3. Controllable image synthesis

In Figure 4, we show that the disentangled style manipulation can be controlled via text inputs. In detail, we can compute the text embedding $\mathbf{t}$ and style code $\mathbf{w}$ using different prompts and apply them to different layers of the generator. This way, we gain not only the coarse and fine style disentanglement but also an intuitive prompt-based maneuver in the style space.

StyleGANs are known to possess a linear latent space useful for image manipulation, called the $\mathcal{W}$-space. Likewise, we perform coarse and fine-grained style swapping using style vectors $\mathbf{w}$. Similar to the $\mathcal{W}$-space of StyleGAN, Figure 5 illustrates that GigaGAN maintains a disentangled $\mathcal{W}$-space, suggesting existing latent manipulation techniques of StyleGAN can transfer to GigaGAN. Furthermore, our model possesses another latent space of text embedding $\mathbf{t} = [\mathbf{t}_{\text{local}}, \mathbf{t}_{\text{global}}]$ prior to $\mathcal{W}$, and we explore its potential for image synthesis.

Figure 6. **Failure cases.** Our outputs with the same prompts as DALL·E 2. Each column conditions on "a teddy bear on a skateboard in Times Square", "a Vibrant portrait painting of Salvador Dali with a robotic half face", and "A close up of a handpalm with leaves growing from it". Compared to production-grade models such as DALL·E 2, our model exhibits limitations in realism and compositionality. See our website for uncurated comparisons.

5. Discussion and Limitations

Our experiments provide a conclusive answer about the scalability of GANs: our new architecture can scale up to model sizes that enable text-to-image synthesis. However, the visual quality of our results is not yet comparable to production-grade models like DALL·E 2. Figure 6 shows several instances where our method fails to produce high-quality results when compared to DALL·E 2, in terms of photorealism and text-to-image alignment for the same input prompts used in their paper.

Nevertheless, we have tested capacities well beyond what is possible with a naïve approach and achieved competitive results with autoregressive and diffusion models trained with similar resources while being orders of magnitude faster and enabling latent interpolation and stylization. Our GigaGAN architecture opens up a whole new design space for large-scale generative models and brings back key editing capabilities that became challenging with the transition to autoregressive and diffusion models. We expect our performance to improve with larger models.

Acknowledgments. We thank Simon Niklaus, Alexandru Chiculita, and Markus Woodson for building the distributed training pipeline. We thank Nupur Kumari, Gaurav Parmar, Bill Peebles, Phillip Isola, Alyosha Efros, and Joonghyuk Shin for their helpful comments. We also want to thank Chenlin Meng, Chitwan Saharia, and Jiahui Yu for answering many questions about their fantastic work. We thank Kevin Duarte for discussions regarding upsampling beyond 4K. Part of this work was done while Minguk Kang was an intern at Adobe Research. Minguk Kang and Jaesik Park were supported by IITP grant funded by the government of South Korea (MSIT) (POSTECH GSAI: 2019-0-01906 and Image restoration: 2021-0-00537).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, 2017. 4

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 7, 8

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 3, 4

[6] Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. Neural Photo Editing with Introspective Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[7] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 3

[8] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision (ECCV)*, 2022. 8

[9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 7, 8

[10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning (ICML)*. PMLR, 2020. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[12] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 28, 2015. 1

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3, 4

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3

[15] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 3

[16] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 1, 3, 6

[18] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 4

[19] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. In *International Conference on Learning Representations (ICLR)*, 2017. 4

[20] 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 *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 7

[21] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[22] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research*, pages 47:1–47:33, 2022. 4

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[24] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *International Conference on Machine Learning (ICML)*, 2021. 4

[25] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering Interpretable GAN Controls. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[26] intao Wang and Liangbin Xie and Chao Dong and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *IEEE International Conference on Computer Vision (ICCV) Workshop*, 2021. 2, 8

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[28] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 29, 2016. 4

[29] Minguk Kang, Woohyeon Shim, Minsu Cho, and Jaesik Park. Rebooting ACGAN: Auxiliary Classifier GANs with Stable Training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3

[30] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7799–7808, 2020. 5

[31] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 4, 6

[32] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 1, 3

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 1, 3, 4, 5

[35] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning (ICML)*, 2021. 4

[36] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5, 6

[37] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[38] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. ViTGAN: Training GANs with vision transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[39] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[40] Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14986–14996, 2021. 5

[41] 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 *European Conference on Computer Vision (ECCV)*, 2014. 3

[42] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating Images from Captions with Attention. In *International Conference on Learning Representations (ICLR)*, 2016. 3, 5

[43] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, 2018. 4

[44] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2018. 4

[45] Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. In *International Conference on Learning Representations (ICLR)*, 2018. 6

[46] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2022. 3, 7, 8

[47] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[48] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[49] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3, 5, 6

[51] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 3

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 2020. 3

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3, 5, 7, 8

[54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021. 3, 5, 7, 8

[55] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, 2016. 3, 6

[56] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016. 3

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable Diffusion. https://github.com/CompVis/stable-diffusion. Accessed: 2022-11-06. 2, 7, 8

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 4, 7, 8

[59] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3, 5, 7, 8

[60] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs Converge Faster. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3, 6

[61] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 3

[62] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 1, 3, 5

[63] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1, 3

[64] Diana Sungatullina, Egor Zakharov, Dmitry Ulyanov, and Victor Lempitsky. Image manipulation with perceptual discriminators. In *European Conference on Computer Vision (ECCV)*, 2018. 6

[65] Md Mehrab Tanjim. DynamicRec: a dynamic convolutional network for next item recommendation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020. 4

[66] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. *arXiv preprint arXiv:2301.12959*, 2023. 3

[67] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[68] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 3

[69] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay Less Attention with Lightweight and Dynamic Convolutions. In *International Conference on Learning Representations (ICLR)*, 2018. 4

[70] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 3

[71] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[72] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3

[73] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 3, 7, 8

[74] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-Attention Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, pages 7354–7363, 2019. 3

[75] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[76] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 3, 6

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6, 8

[78] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2021. 6

[79] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv 2006.10738*, 2020. 3

[80] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8

[81] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[82] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[83] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[84] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 3

[85] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. A text-to-picture synthesis system for augmenting communication. In *The AAAI Conference on Artificial Intelligence*, 2007. 3