# Dual Contrastive Loss and Attention for GANs

Ning Yu[1,2] Guilin Liu[3] Aysegul Dundar[3,4]
Andrew Tao[3] Bryan Catanzaro[3] Larry Davis[1] Mario Fritz[5]
[1]University of Maryland [2]Max Planck Institute for Informatics [3]NVIDIA
[4]Bilkent University [5]CISPA Helmholtz Center for Information Security
{ningyu,lsdavis}@umd.edu
{guilinl,adundar,atao,bcatanzaro}@nvidia.com
fritz@cispa.saarland

## Abstract

*Generative Adversarial Networks (GANs) produce impressive results on unconditional image generation when powered with large-scale image datasets. Yet generated images are still easy to spot especially on datasets with high variance (e.g. bedroom, church). In this paper, we propose various improvements to further push the boundaries in image generation. Specifically, we propose a novel dual contrastive loss and show that, with this loss, discriminator learns more generalized and distinguishable representations to incentivize generation. In addition, we revisit attention and extensively experiment with different attention blocks in the generator. We find attention to be still an important module for successful image generation even though it was not used in the recent state-of-the-art models. Lastly, we study different attention architectures in the discriminator, and propose a reference attention mechanism. By combining the strengths of these remedies, we improve the compelling state-of-the-art Fréchet Inception Distance (FID) by at least 17.5% on several benchmark datasets. We obtain even more significant improvements on compositional synthetic scenes (up to 47.5% in FID).*

## 1. Introduction

Photorealistic image generation has increasingly become reality, benefiting from the invention of generative adversarial networks (GANs) [24] and its successive breakthroughs [67, 3, 25, 60, 5, 41, 42, 43]. The progress is mainly driven by large-scale datasets [18, 57, 91, 38, 54, 42], architectural tuning [10, 98, 42, 43, 69], and loss designs [58, 3, 25, 60, 39, 101, 105, 96, 40, 106, 36]. GAN techniques have been popularized into extensive computer vision applications, including but not limited to image translation [35, 107, 108, 54, 33, 82, 64, 20, 63], post-processing [46, 71, 44, 45, 77, 62, 102], image manipulation [13, 14, 70, 1, 4, 80], texture synthesis [94, 53, 59], image inpainting [34, 52, 92, 93], and text-to-image generation [68, 99, 100, 74].

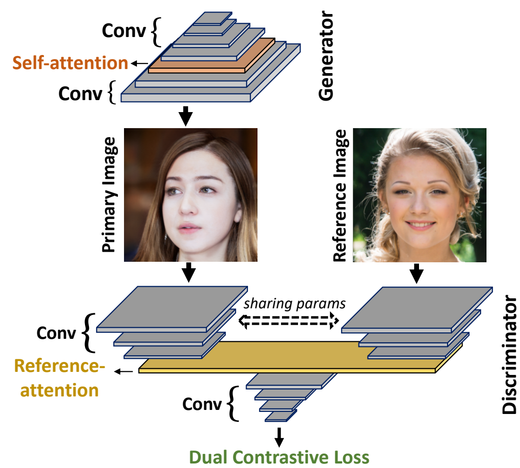


Figure 1. The diagram of our GAN framework using three key components: self-attention in the generator, reference-attention in the discriminator, and a novel dual contrastive loss. Technical diagrams are in Fig. 2 and 4.

Yet, behind the seemingly saturated performance of the state-of-the-art StyleGAN2 [43], there still persists open issues of GANs that make generated images surprisingly obvious to spot [95, 81, 21, 28]. Hence, it is still necessary to revisit the fundamental generation power when other concurrent deep learning techniques keep advancing and creating space for GAN improvements.

We investigate methods to improve GANs in two dimensions. In the first dimension, we work on the loss function. As the discriminator aims to model the intractable real data distribution via a workaround of real/fake binary classification, a more effective discriminator can back-propagate more meaningful signals for the generator to compete against.

However, the feature representations of discriminators are often not generalized enough to incentivize the adversarially evolving generator and are prone to forgetting previous tasks [11] or previous data modes [72, 49]. This often leads to the generated samples with discontinued semantic structures [51, 98] or the generated distribution with mode collapse [72, 96]. To mitigate this issue, we propose to synergize generative modeling with the advancements in contrastive learning [61, 8]. In this direction, for the first time, we replace the logistic loss of StyleGAN2 with a newly designed dual contrastive loss.

In the second dimension, we revisit the architecture of both generator and discriminator networks. Specifically, many GAN-based image generators rely on convolutional layers to encode features. In such design, long-range dependencies across pixels (e.g., large-size semantically correlated layouts) can only be formulated with a deep stack of convolutional layers. This, however, does not favor the stability of GAN training because of the challenge to coordinate multiple layers desirably. The minimax formulation and the alternating gradient ascent-descent in the GAN framework further exacerbate such instability. To circumvent this issue, attention mechanisms that support long-range modeling across image regions are incorporated into GAN models [98, 5]. After that, however, StyleGAN2 claimed the state of the art with a novel architectural design without any attention mechanisms. Therefore, it turns not clear whether attention still improves results, which of the popular attention mechanisms [37, 85, 83, 103] improves the most, and in return of how many additional parameters. To answer these questions, we extensively study the role of attention in the current state-of-the-art generator, and during this study improve the results significantly.

In the discriminator, we again explore the role of attention as shown in Fig. 1. We design a novel reference attention mechanism in the discriminator where we allow two irrelevant images as the inputs at the same time: one input is sampled from real data as a reference, and the other input is switched between a real sample and a generated sample. The two inputs are encoded through two Siamese branches [6, 15, 73, 97] and fused by a reference-attention module. In this way, we achieve to guide real/fake classification under the attention of the real world. Contributions are summarized as follow:

- We propose a novel dual contrastive loss in adversarial training that generalizes representation to more effectively distinguish between real and fake, and further incentivize the image generation quality.
- We investigate variants of the attention mechanism in GAN architecture to mitigate the local and stationary issues of convolutions.
- We design a novel reference-attention discriminator architecture that benefits limited-scale datasets.
- We redefine the state of the art by improving FID scores by at least 17.5% on several large-scale benchmark datasets. We also achieve more realistic generation on the CLEVR dataset [38] which poses different challenges from the other datasets: compositional scenes with occlusions, shadows, reflections, and mirror surfaces. It comes with 47.5% FID improvement.

## 2. Related work

**Generative adversarial networks (GANs).** Since the invention of GANs [24], there have been rapid progress to achieve photorealistic image generation [67, 3, 25, 25, 60, 5, 41, 42, 43]. Significant improvements are obtained by careful architectural designs for generators [10, 98, 42, 43, 69], discriminators [82, 56] and new regularization techniques [58, 3, 25, 60, 101, 105, 96, 40, 106, 36]. Architectural evolution in generators started from a multi-layer perceptron (MLP) [24] and moved to deep convolutional neural networks (DCNN) [67], to models with residual blocks [60], and recently style-based [42, 43] and attention-based [98, 5] models. Similarly, discriminators evolved from MLP to DCNN [67], however, their design has not been studied as aggressively. In this paper, we propose changes in both generators and discriminators, and for the loss function.

**Contrastive learning.** Contrastive learning targets a transformation of inputs into an embedding where associated signals are brought together, and they are distanced from the other samples in the dataset [26, 76, 8, 9]. The same intuition behind contrastive learning has also been the base of Siamese networks [6, 15, 73, 97]. Contrastive learning is shown to be an effective tool for unsupervised learning [61, 27, 87], conditional image synthesis [63, 40, 106], and domain adaptation [23]. In this work, we study its effectiveness when it is closely coupled with the adversarial training framework and replaces the conventional adversarial loss for unconditional image generation. It is orthogonal to [40, 106, 36, 47] where their contrastive losses serve only as an incremental auxiliary to the conventional adversarial loss, apply to the generator rather than the discriminator, and/or require expensive class annotations or augmentation for generation.

**Attention models.** Attention models have dominated the language modeling [78, 86, 17, 19, 89], and became popular among various computer vision problems from image recognition [16, 79, 31, 32, 104, 109, 30, 85] to image captioning [88, 90, 7] to video prediction [37, 83]. They are proposed in various forms: spatial attention that reweights the convolution activations [98, 83, 12], in different channels [79, 31, 32], or a combination of them [7, 84, 22]. Attention models with their reweighting mechanisms provide a possibility for long-range modeling across distant image regions. As attention models outperform others in various computer vision tasks, researchers were quick to incorporate them into unconditional image generation [10, 98, 65, 5], semantic-based image generation [56, 75], and text-guided image manipulation models [48, 66]. Even though attention models have already benefited the image generation tasks, we believe the results can be further improved by empowering the state-of-the-art image synthesis models [43] (attention not involved) with the most recent achievements in the attention modules [103]. In addition, we design a novel reference-attention architecture for the discriminator and show a further boost on limited-scale datasets.

# 3. Approach

Our improvements for GANs include a novel dual contrastive loss and variants of the attention mechanisms. For each improvement, we organize the context in a combination between method formulation and experimental investigation. After validating our optimal configuration, we compare it to the state of the art in Section 4.

## 3.1. Dual contrastive loss

Adversarial training relies on the discriminator's ability on real vs. fake classification. As in other classification tasks, discriminators are also prone to overfitting when the dataset size is limited [2]. On larger datasets, on the other hand, there is no study showing that discriminators overfit but we hypothesize that adversarial training can still benefit from novel loss functions which encourage the distinguishability power of the discriminator representations for their real vs. fake classification task.

We put another lens on the representation power of the discriminator by incentivizing generation via contrastive learning. Contrastive learning associates data points and their positive examples and disassociates the other points within the dataset which are referred to as negative examples. It is recently re-popularized by various unsupervised learning works [26, 61, 76, 8, 9] and generation works [63, 40, 106]. Among these works, contrastive learning is used as an auxiliary task. For example in image to image translation task, a translator learns to output a zebra image given a horse image via adversarial loss and in addition learns to align the input horse image and the generated zebra image via contrastive loss function [63]. Contrastive loss in that work is utilized such that given a patch showing the legs of an output zebra should be strongly associated with the corresponding legs of the input horse, more so than the other patches randomly extracted from the horse image.

In this work, different from the previous ones, we do not use contrastive learning as an auxiliary task but directly couple it in the main adversarial training by a novel loss function formulation. We, to the best of our knowledge, for the first time train an unconditional GAN by solely relying on contrastive learning. As shown in Fig. 2 Right Case I, our contrastive loss function aims at teaching the discriminator to disassociate a single real image against a batch of generated images. Dually in Case II, the discriminator learns to disassociate a single generated image against a batch of real images. The generator adversarially learns to minimize such dual contrasts. Mathematically, we derive this loss function by extending the binary classification used in [24, 43] to a noise contrastive estimation framework [61], a one-against-a-batch classification in the softmax cross-entropy formulation. The novel formulation is as follows:

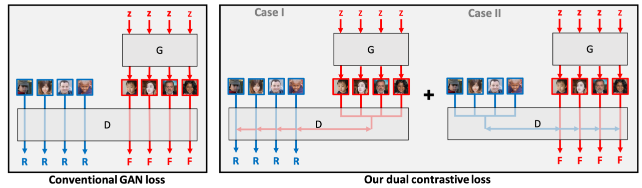


Figure 2. Comparisons between the diagram of conventional GAN loss and diagram of our dual contrastive loss. Our contrastive loss in Case I aims at teaching the discriminator to disassociate a single real image (R) against a batch of generated images (F). Dually in Case II, the discriminator learns to disassociate a single generated image against a batch of real images.

In Case I:

$$L_{real}^{contr}(G,D) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log \frac{e^{D(\mathbf{x})}}{e^{D(\mathbf{x})} + \sum\limits_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} e^{D(G(\mathbf{z}))}}\right]$$
$$= -\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log\left(1 + \sum_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} e^{D(G(\mathbf{z}))-D(\mathbf{x})}\right)\right] \quad (1)$$

In Case II:

$$L_{fake}^{contr}(G,D) = \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\log \frac{e^{-D(G(\mathbf{z}))}}{e^{-D(G(\mathbf{z}))} + \sum\limits_{\mathbf{x}\sim p(\mathbf{x})} e^{-D(\mathbf{x})}}\right]$$
$$= -\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\log\left(1 + \sum_{\mathbf{x}\sim p(\mathbf{x})} e^{D(G(\mathbf{z}))-D(\mathbf{x})}\right)\right] \quad (2)$$

Comparing between Eq. 1 and 2, the duality is formulated by switching the order of real/fake sampling while keeping the other calculation unchanged. Comparing to the logistic loss [24, 43], contrastive loss enriches the softplus formulation $\log(1+e^{D(\cdot)})$ with a batch of inner terms and using discriminator logit contrasts between real and fake samples. Finally, our adversarial objective is:

$$\min_G \max_D \; L_{real}^{contr}(G,D) + L_{fake}^{contr}(G,D) \quad (3)$$

**Investigation on loss designs.** We extensively validate the effectiveness of dual contrastive loss compared to other

| | FFHQ | Bedroom | Church | Horse | CLEVR |
|---|---|---|---|---|---|
| Non-saturating [24] (default) | 4.86 | 4.01 | 4.54 | 3.91 | 9.62 |
| Saturating [24] | 5.16 | 4.26 | 4.80 | 5.90 | 10.46 |
| Wasserstein [25] | 7.99 | 6.05 | 6.28 | 7.23 | **5.82** |
| Hinge [50] | 4.14 | 4.92 | 4.39 | 5.27 | 14.87 |
| Dual contrastive (**ours**) | **3.98** | **3.86** | **3.73** | **3.70** | 6.06 |

Table 1. Comparisons in FID among different GAN losses. Based on StyleGAN2 config E backbone, it shows our contrastive loss outperforms a variety of other losses on four out of five large-scale datasets. Wasserstein loss is better than ours on CLEVR, but are the worst on the other datasets.

| Loss | FFHQ | Bedroom | Church | Horse | CLEVR |
|---|---|---|---|---|---|
| Non-saturating [24] (default) | 245. | 332. | 517. | 1285. | 199. |
| Dual contrastive (**ours**) | **377.** | **580.** | **856.** | **1645.** | **513.** |

Table 2. Comparisons in FDDF between StyleGAN2 default loss and our loss. A larger value is more desirable, indicating the learned discriminator features are more distinguishable between real and fake.

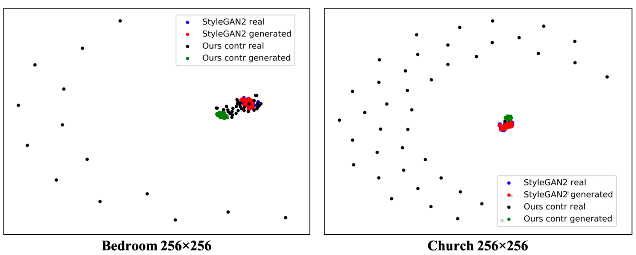


Figure 3. The tSNE plots for the distributions of discriminator features. The distinguishability of features based on our contrastive loss is much more significant than that based on the default non-saturating loss in StyleGAN2 baseline. Our loss learns to associate fake features to a "core" clique (green) while pushing real features in the wild outwards as "satellites" (black). The baseline loss fails to differentiate features from the two sources (red v.s. blue) with a clear margin.

loss functions as presented in Table 1. We replace the loss used in StyleGAN2 [43], non-saturating default loss, with other popular GAN losses while keeping all the other parameters the same. As shown in Table 1, dual contrastive loss is the only loss that significantly improves upon the default loss of StyleGAN2 consistently on all the five datasets. Wasserstein loss is better than ours on CLEVR dataset, but is the worst among all the loss functions on the other datasets. We reason the success of the dual loss to its formulation that explicitly learns an unbiased representation between real and generated distributions.

**The distinguishability of contrastive representation.** Motivated by the consistent improvement from our dual contrastive loss, we delve deeper to investigate if and by how much our contrastive representation is more distinguishable than the original discriminator representation. We measure the representation distinguishability by the Fréchet distance of the discriminator features in the last layer (FDDF) between 50K real and generated samples. A larger value indicates the feature is more distinguishable between real and fake. We find our dual contrastive features to be consistently more distinguishable than the original discriminator features as shown in Table 2 and Fig. 3, which back-propagates more effective gradients to incentivize our generator.

### 3.2. Self-attention in the generator

The majority of the GAN-based image generators rely solely on convolutional layers to extract features [67, 3, 25, 60, 41, 42, 43], even though the local and stationary convolution primitive in the generator can not model the long-range dependencies in an image. Among recent GAN-based models, SAGAN [98] uses the self-attention block [83] and demonstrates improved results. BigGAN [5] also follows this choice and uses a similar attention module for better performance. After that, however, StyleGAN [42] and StyleGAN2 [43] redefine the state of the art with various modifications in the generator architecture which do not include any attention mechanism. StyleGAN2 also shows that generation results can be improved by larger networks with an increased number of convolution filters. Therefore, it is now not clear what the role of attention is in the state-of-the-art image generation models. Does attention still improve the network performance? Which attention mechanism benefits the most and in the trade of how many additional parameters? To answer these questions, we experiment with previously proposed self-attention modules: Dynamic Filter Networks (DFN) [37], Visual Transformers (VT) [85], Self-Attention GANs (SAGAN) [98], as well as the state-of-the-art patch-based spatially-adaptive self-attention module, SAN [103].

All the above self-attention modules are benefited from their adaptive data-dependent parameter space while they have their own hand-crafted architecture designs and interpretability. DFN [37] keeps the convolution primitive but makes the convolutional filter condition to its input tensor. VT [85] compresses input tensor to a set of 1D feature vectors, interprets them as semantic tokens, and leverages language transformer [78] for tensor propagation. SAN [103] generalizes the self-attention block [83] (as used in SAGAN [98]) by replacing the point-wise softmax attention with a patch-wise fully-connected transformation.

We show the diagram of self-attention in Figure 4, with a specific instantiation from SAN [103] due to its generalized and state-of-the-art design. Note that the attention module is agnostic to network backbone and can be switched to other options for fair comparisons. For conceptual and technical completeness, we formulate our SAN-based self-attention below.

In details, let $\mathbf{T} \in \mathbb{R}^{h \times w \times c}$ be the input tensor to a convolutional layer in the original architecture. Following the mainstream protocol of self-attention calcula-

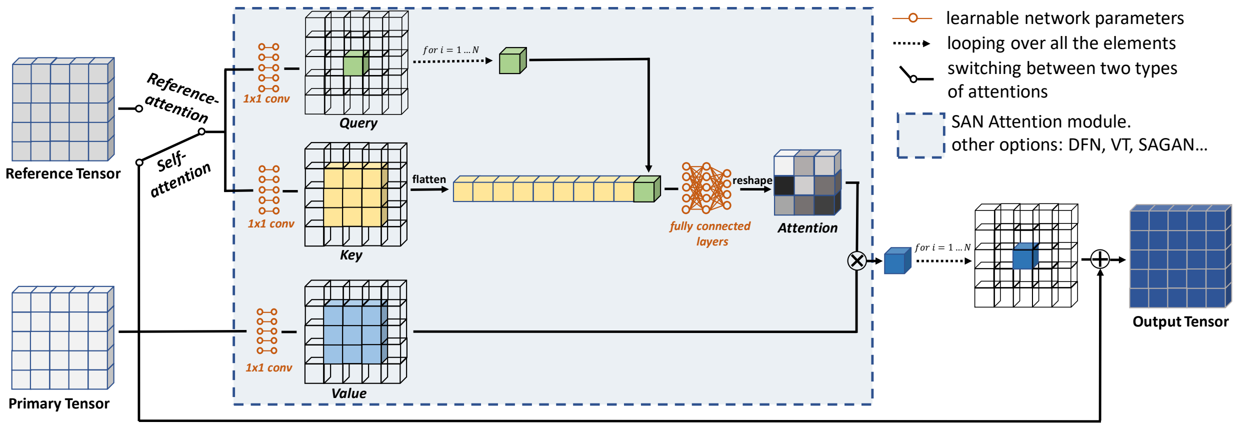


Figure 4. The diagram of self-attention and reference-attention schemes. The attention module is instantiated by SAN [103] but is agnostic to network backbone. It can flexibly switch to other options and be plug-and-play. We switch between the sources that are used to calculate the Key and Query tensors, so as to implement self-attention and reference-attention respectively.

tion [83, 98, 65], we obtain the corresponding key, query, and value tensors $\mathbf{K}(\mathbf{T}), \mathbf{Q}(\mathbf{T}), \mathbf{V}(\mathbf{T}) \in \mathbb{R}^{h\times w\times c}$ separately using $1\times 1$ convolutional kernel followed by bias and leaky ReLU. For each location $(i, j)$ within the tensor spatial dimensions, we extract a large patch with size $s$ from $\mathbf{K}$ centered at $(i, j)$, denoted as $\mathbf{k} \in \mathbb{R}^{s\times s\times c}$. We then flatten the patch and concatenate it along the channel dimension with $\mathbf{q} \in \mathbb{R}^{1\times 1\times c}$, the query vector at $(i, j)$, to obtain $\mathbf{p} \in \mathbb{R}^{1\times 1\times (s^2c+c)}$:

$$\begin{aligned} \mathbf{k} &= \mathbf{K}_{(i-\frac{s}{2}:i+\frac{s}{2}+1,\; j-\frac{s}{2}:j+\frac{s}{2}+1)} \\ \mathbf{q} &= \mathbf{Q}_{(i,j)} \\ \mathbf{p} &= \mathsf{concat}\left(\mathsf{flatten}(\mathbf{k}), \mathbf{q}\right) \end{aligned} \tag{4}$$

In order to cooperate between the key and query, we feed $\mathbf{p}$ through two fully-connected layers followed by bias and leaky ReLU and obtain a vector with size $\tilde{\mathbf{w}} \in \mathbb{R}^{1\times 1\times s^2c}$:

$$\begin{aligned} \hat{\mathbf{w}} &= \mathsf{leakyReLU}(\mathbf{p}\mathbf{M}_{w1} + \mathbf{b}_{w1}) \\ \tilde{\mathbf{w}} &= \hat{\mathbf{w}}\mathbf{M}_{w2} + \mathbf{b}_{w2} \end{aligned} \tag{5}$$

$\mathbf{M}_{w1} \in \mathbb{R}^{(s^2c+c)\times s^2c}$, $\mathbf{M}_{w2} \in \mathbb{R}^{s^2c\times s^2c}$, and $\mathbf{b}_{w1}, \mathbf{b}_{w2} \in \mathbb{R}^{1\times 1\times s^2c}$ are the learnable parameters in the fully connected layers and biases.

On one hand we reshape $\tilde{\mathbf{w}}$ back to the patch size $\mathbf{w} \in \mathbb{R}^{s\times s\times c}$; on the other hand we extract a patch with the same size from $\mathbf{V}$ centered at $(i, j)$, denoted as $\mathbf{v} \in \mathbb{R}^{s\times s\times c}$. Next, we aggregate $\mathbf{v}$ over spatial dimensions with the correponding weights from $\mathbf{w}$ to derive an output vector $\mathbf{o} \in \mathbb{R}^{1\times 1\times c}$:

$$\begin{aligned} \mathbf{w} &= \mathsf{reshape}(\tilde{\mathbf{w}}) \\ \mathbf{v} &= \mathbf{V}_{(i-\frac{s}{2}:i+\frac{s}{2}+1,\; j-\frac{s}{2}:j+\frac{s}{2}+1)} \\ \mathbf{o}(i,j) &= \sum_{m,n=1}^{s} \mathbf{w}_{(m,n)}\mathbf{v}_{(m,n)} \end{aligned} \tag{6}$$

We loop over all the $(i, j)$ to constitute an output tensor $\bar{\mathbf{O}}^{\mathit{self}} \in \mathbb{R}^{h\times w\times c}$ and define it as the self-attention output. Finally, we replace the original convolution output with $\mathbf{O}^{\mathit{self}} \in \mathbb{R}^{h\times w\times c}$, a residual version of this self-attention output.

$$\begin{aligned} \bar{\mathbf{O}}^{\mathit{self}}_{(i,j)} &= \mathbf{o}(i,j), \quad \forall i = 1,\ldots,h,\; j = 1,\ldots,w \\ \bar{\mathbf{O}}^{\mathit{self}} &\doteq \mathsf{attn}\left(\mathbf{K}(\mathbf{T}), \mathbf{Q}(\mathbf{T}), \mathbf{V}(\mathbf{T})\right) \\ \mathbf{O}^{\mathit{self}} &= \bar{\mathbf{O}}^{\mathit{self}} + \mathbf{T} \end{aligned} \tag{7}$$

It is worth noting that $\mathbf{w}$ plays a conceptually equivalent role as the softmax attention map of the traditional key-query aggregation [83, 98, 65], except it is not identical across channels anymore but rather generalized to optimize for each channel. $\mathbf{w}$ also aligns in spirit with the concept of DFN [37], except the spatial size $s\times s$ is empirically set much larger than $3\times 3$, and more importantly, $\mathbf{w}$ is not "sliding" anymore but rather generalized to optimize at each location.

**Investigations on self-attention modules.** In Table 3 we extensively compare among a variety of self-attention modules by replacing the default convolution in the $32\times 32$-resolution layer in StyleGAN2 [43] config E backbone with one of them. We justify that SAN [103] significantly improves over the StyleGAN2 baseline and outperforms the other attention variants on several datasets. DFN [37] is better than ours on CelebA dataset, but is the worst on most other datasets. We provide additional ablation studies on network architectures in the supplementary material.

We visualize the attention map examples of the best performing generator (StyleGAN2 + SAN) in Fig. 5. We find attention maps to strongly correspond to the semantic layout and structures of the generated images.

**Complexity of self-attention modules.** We also compare in Table 4 the time and space complexity of these self-attention modules. We observe that DFN [37] and VT [85] moderately improve the generation quality yet in the trade

| | CelebA | Animal Face | Bedroom | Church |
|---|---|---|---|---|
| StyleGAN2 [43] | 9.84 | 36.55 | 19.33 | 11.02 |
| + DFN [37] | **8.41** | 35.10 | 26.86 | 11.31 |
| + VT [85] | 9.18 | 34.70 | 16.85 | 10.64 |
| + SAGAN [98] | 9.35 | 34.83 | 17.94 | 10.65 |
| + SAN [103] | 8.60 | **32.72** | **16.36** | **9.62** |

Table 3. Comparisons in FID among different attention modules in the generator. StyleGAN2 config E which does not include an attention module is used as a backbone. For computationally efficient comparisons, we use the 30k subset of each dataset at 128×128 resolution.

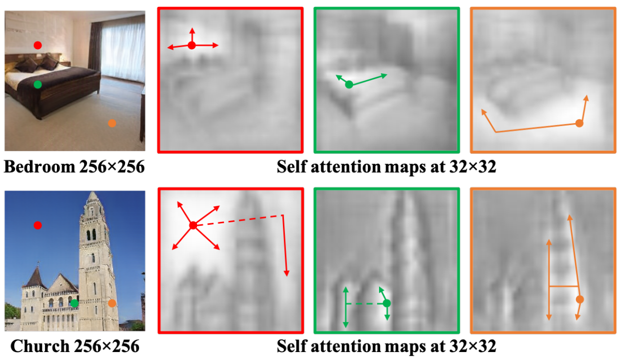


Figure 5. StyleGAN2 + SAN generated samples and their self-attention maps in the generator for the corresponding dot positions. Considering there is an attention weight kernel $\mathbf{w} \in \mathbb{R}^{s \times s \times c}$ for each position, we visualize the norm for each spatial position of $\mathbf{w}$. The attention maps strongly align to the semantic layout and structures of the generated images, which enable long-range dependencies across objects. See more samples in the supplementary material.

| Method | FLOPS (G) | #parameters (M) |
|---|---|---|
| StyleGAN2 [43] | 1.08 | 48.77 |
| + DFN [37] | 4.20 | 177.60 |
| + VT [85] | 7.39 | 240.09 |
| + SAGAN [98] | 0.99 | 44.99 |
| + SAN [103] | 1.08 | 48.43 |

Table 4. Time complexity in FLOPS and space complexity in the number of parameters for each method.

of undesirable $> 3.6\times$ complexity. On the contrary, the improvements from SAGAN [98] or SAN [103] are not at the cost of complexity, but rather benefited from the more representative attention designs. They use a fewer number of convolution channels and the multi-head trick [83] to control their complexity. These results show that the improved performance does not come from any additional parameters but rather the attention structure itself.

### 3.3. Reference-attention in the discriminator

First, we apply SAN [103], the best attention mechanism we validated in the generator, to the discriminator. However, we do not see a benefit of such design as shown in Table 5. Then, we explore an advanced attention scheme given that two classes of input (real vs. fake) are fed to the discriminator. We allow the discriminator to take two image inputs at the same time: the *reference* image and the *primary* image where we set the reference image to always be a real sample while the primary image to be either a real or generated sample. The reference image is encoded to represent one part of the attention components. These components are learned to guide the other part of the attention components, which are encoded from the *primary* image. There are three insights in this advancement. (1) An effective discriminator encodes real images and generated images differently, so that reference-attention is capable of learning positive feedback given both images from the real class and negative feedback given two images from different classes. Such a scheme amplifies the representation difference between real and fake, and in turn potentially strengthens the power of the discriminator. (2) Reference-attention enables distribution estimation in the discriminator feature level beyond the discriminator logit level in the original GAN framework, which guides generation more strictly towards the real distribution. (3) Reference-attention learns to cooperate real and generated images explicitly in one round of back-propagation, instead of individually classifying them and trivially averaging the gradients over one batch. Arbitrarily pairing up images mitigates discriminator from overfitting, similar to the spirit of random data augmentation, but we instead conduct random feature augmentation using attention.

In detail, we first encode the reference image and the primary image through the original discriminator layers prior to the convolution at a certain resolution. To align feature embeddings, we apply the Siamese architecture [6, 15] to share layer parameters as shown in Fig. 1. We then apply the same attention scheme as used in the generator, except we use the tensor $\mathbf{T}_{ref} \in \mathbb{R}^{h \times w \times c}$ from the reference branch to calculate the key and query tensors, and use the tensor $\mathbf{T}_{pri} \in \mathbb{R}^{h \times w \times c}$ from the primary branch to calculate the value tensor and the residual shortcut. Finally, we replace the original convolution output with our reference-attention output:

$$\mathbf{O}^{ref} \doteq \mathsf{attn}\left(\mathbf{K}(\mathbf{T}_{ref}), \mathbf{Q}(\mathbf{T}_{ref}), \mathbf{V}(\mathbf{T}_{pri})\right) + \mathbf{T}_{pri} \quad (8)$$

After the reference-attention layer, the two Siamese branches fuse into one and are followed by the remaining discriminator layers to obtain the classification logit. We show in Fig. 4 the diagram of reference-attention. Eq. 8 provides the flexibility how to cooperate between reference and primary images. We empirically explore the other compositions of sources to the key, query, and value components of reference-attention in the supplementary material as well as additional ablation studies on network architectures.

From Table 5 we validate reference-attention mechanism (ref attn) to improve the results whereas self-attention to be

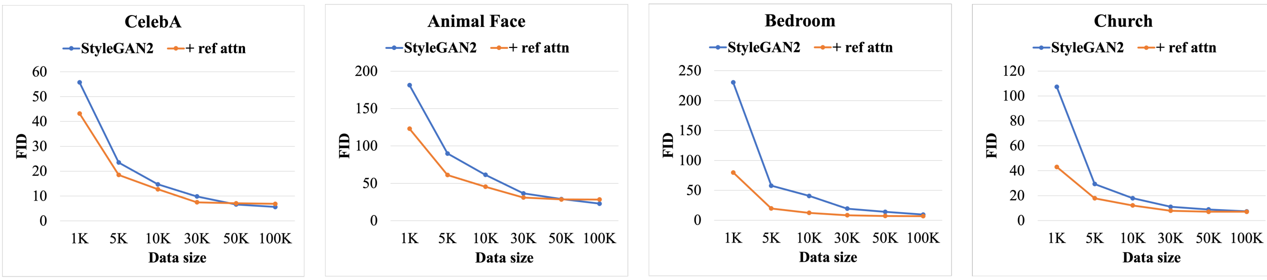


Figure 6. Comparisons in FID between StyleGAN2 config E baseline (blue) and that with our reference-attention in the discriminator (orange). Our method consistently improves the baseline when dataset size varies between 1k and 30k images. For computationally efficient comparisons, we use each dataset at 128×128 resolution. See the supplementary material for the values in these plots.

| | CelebA | Animal Face | Bedroom | Church |
|---|---|---|---|---|
| StyleGAN2 [43] | 9.84 | 36.55 | 19.33 | 11.02 |
| + self attn in D | 10.49 | 42.41 | 17.22 | 11.06 |
| + ref attn in D | **7.48** | **31.08** | **8.32** | **7.86** |

Table 5. Comparisons in FID among different attention configurations in the discriminator. StyleGAN2 config E which does not include any attention module is used as a backbone. For computationally efficient comparisons, we use the 30k subset of each dataset at 128×128 resolution.

barely benefiting for the discriminator. Encouraged with these findings, we run the proposed reference-attention on full-scale datasets but do not see any improvements. Therefore, we dive deep into reference-attentions behavior in the discriminator with respect to the dataset size as given in Fig. 6. We find that the reference-attention in the discriminator consistently improves the performance when dataset size varies between 1k and 30k images, and on contrary slightly deteriorates the performance when dataset sizes increase further. We reason that the arbitrary pair-up of the reference and primary image inputs to prevent overfitting when data size is small but causing underfitting with the increase of data size Even though in this paper our main scope is GANs on large-scale datasets, we believe these findings to be very interesting for researchers to design their networks for limited-scale datasets. We summarize our comparisons on limited-scale datasets in the supplementary material.

## 4. Comparisons to the state of the art

**Implementation details.** All our models are built upon the most recent state-of-the-art unconditional StyleGAN2 [43] config E for its high performance and reasonable speed. We leverage the plug-and-play advantages of all our improvement proposals to strictly follow StyleGAN2 official setup and training protocol, which facilitates reproducibility and fair comparisons. For dual contrastive loss, we first warm up training with the default non-saturating loss for about 20 epochs, and then switch to train with our loss.

**Datasets.** We use several benchmark datasets, 70K FFHQ face dataset [42], 3M LSUN Bedroom dataset [91], 120K LSUN Church dataset [91], 2M LSUN Horse dataset [91], CelebA face dataset [57] and Animal Face dataset [55], and 70K CLEVR [38] dataset which contains rendered images with random compositions of 3D shapes, uniform materials, uniform colors, point lighting, and a plain background. It poses different challenges from the other common datasets: compositional scenes with occlusions, shadows, reflections, and mirror surfaces. We use 256×256 resolution images for each of these datasets except the CelebA and Animal Face datasets which are used in 128×128 resolutions. We do not experiment with 1024×1024 resolution of FFHQ as it takes 9 days to train StyleGAN2 base model. Instead, we run extensive experiments on the mentioned various datasets. If not otherwise noted, we use the whole dataset.

**Evaluation.** FID [29] is regarded as the golden standard to quantitatively evaluate generation quality. We follow the protocol in StyleGAN2 [43] to report the FID between 50K generated images and 50K real testing images. The smaller the more desirable. In the supplementary material, we report various other metrics that are proposed in StyelGAN [42] or StyleGAN2 [43] but are less benchmarked in other literature, Perceptual Path Length, Precision, Recall, and Separability.

**Comparisons.** Besides StyleGAN2 [43], we also compare to a parallel state-of-the-art study, U-Net GAN [69], which was build upon and improved on BigGAN [5]. We train U-Net by adapting it to the better backbone of StyleGAN2 [43] for fair comparison, and obtain better results than their official release on non-FFHQ datasets. As shown in Table 6, our self-attention generator improves on four out of five large-scale datasets, up to 13.3% relative improvement on Bedroom dataset. This highlights the benefits of attention to details and to long-range dependencies on complex scenes. However, self-attention does not improve on the extensively studied FFHQ dataset. We reason that the image pre-processing of facial landmark alignment compensates for the lack of attention schemes, which makes previous works also overlook them on other datasets.

Our dual contrastive loss improves effectively on all the

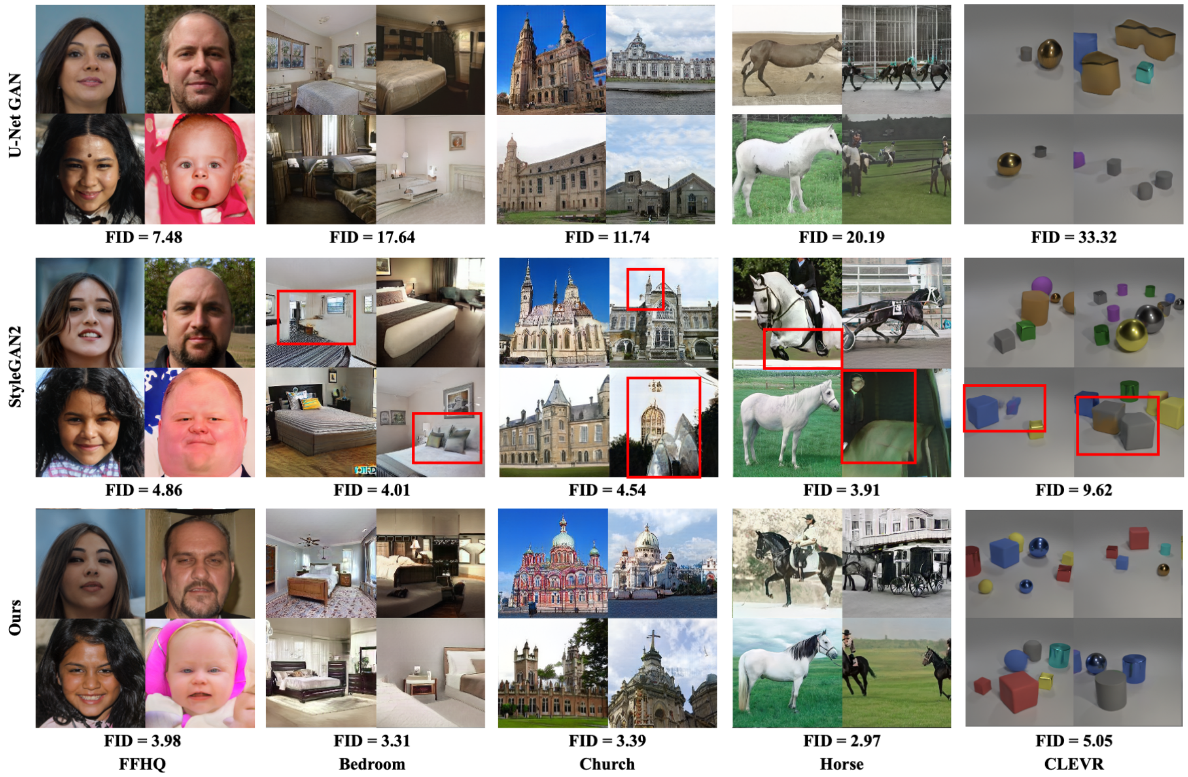


Figure 7. Uncurated generated samples. To align the comparisons, we use the same real query images for pre-trained generators to reconstruct. Artifacts from StyleGAN2 are highlighted with red boxes. Zoom in for details. In particular, our generation significantly outperforms the baselines on CLEVR images which strongly rely on long-range dependencies (occlusions, shadows, reflections, etc) and consistency (consistent shadow directions, consistent specularity, regular shapes, uniform colors, etc). See more samples in the supplementary material.

| Method | Loss | FFHQ | Bedroom | Church | Horse | CLEVR |
|---|---|---|---|---|---|---|
| BigGAN [5] | Adv | 11.4 | - | - | - | - |
| U-Net GAN [69] | Adv | 7.48 | 17.6 | 11.7 | 20.2 | 33.3 |
| StyleGAN2 [43] | Adv | 4.86 | 4.01 | 4.54 | 3.91 | 9.62 |
| StyleGAN2 w/ attn | Adv | 5.13 | <u>3.48</u> | 4.38 | <u>3.59</u> | 8.96 |
| StyleGAN2 | Contr | **3.98** | 3.86 | <u>3.73</u> | 3.70 | <u>6.06</u> |
| StyleGAN2 w/ attn | Contr | <u>4.63</u> | **3.31** | **3.39** | **2.97** | **5.05** |

Table 6. Comparisons in FID to the state-of-the-art GANs on the large-scale datasets. We highlight the best in **bold** and second best with <u>underline</u>. "w/ attn" indicates using the self-attention in the generator. "Contr" indicates using our dual contrastive loss instead of conventional GAN loss.

datasets, up to 37% improvement on CLEVR dataset. This highlights the benefits of contrastive learning on generalized representation, especially on aligned datasets, e.g. FFHQ and CLEVR, that can easily make a traditional discriminator overfit. The synergy effective between self-attention and contrastive learning is significant and consistent, resulting in at least 17.5% and up to 47.5% relative improvement on CLEVR. Especially for CLEVR, our generator handles more realistically for occlusions, shadows, reflections, and mirror surfaces. As shown in Fig. 7, our method suppresses artifacts that were previously visible in StyleGAN2 baseline outputs, with red boxes, e.g., the artifacts on the wall in Bedroom images, discontinuities in the structure in Church images, as well as color leakage between objects in CLEVR images.

# 5. Conclusion

The advancements in attention schemes and contrastive learning generate opportunities for new designs of GANs. Our attention schemes serve as a beneficial replacement for local and stationary convolutions, so as to equip generation and discriminator representation with long-range adaptive dependencies. In particular, our reference-attention discriminator cooperates between real reference images and primary images, mitigates discriminator overfitting, and leads to further boost on limited-scale datasets. Additionally, our novel contrastive loss generalizes discriminator representations, makes them more distinguishable between real and fake, and in turn incentivizes better generation quality.


# Acknowledgement

This work was partially supported by the DARPA SAIL-ON (W911NF2020009) program. Ning Yu is partially supported by Twitch Research Fellowship. We thank Tero Karras, Xun Huang, and Tobias Ritschel for constructive advice.

## References


[1] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv*, 2020. 1

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017. 3

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*. 1, 2, 4

[4] Anand Bhattad, Aysegul Dundar, Guilin Liu, Andrew Tao, and Bryan Catanzaro. View generalization for single image textured 3d models. In *CVPR*, 2021. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 1, 2, 4, 7, 8

[6] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *IJPRAI*, 1993. 2, 6

[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 3

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 2, 3

[10] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks. In *ICLR*, 2019. 1, 2

[11] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *CVPR*, 2019. 2

[12] Lu Chi, Zehuan Yuan, Yadong Mu, and Changhu Wang. Non-local neural networks with grouped bilinear attentional transforms. In *CVPR*, 2020. 2

[13] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 1

[14] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1

[15] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 2, 6

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *CVPR*, 2017. 2

[17] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019. 2

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[20] Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, and Bryan Catanzaro. Panoptic-based image synthesis. In *CVPR*, 2020. 1

[21] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 1

[22] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[23] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2, 3, 4

[25] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 1, 2, 4

[26] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2, 3

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[28] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. *IJCAI*, 2021. 1

[29] 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 *NeurIPS*, 2017. 7

[30] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, 2019. 2

[31] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *NeurIPS*, 2018. 2

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2

[33] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1

[34] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ToG*, 2017. 1

[35] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[36] Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *ICLR*, 2021. 1, 2

[37] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, 2016. 2, 4, 5, 6

[38] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1, 2, 7

[39] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *ICLR*, 2019. 1

[40] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In *NeurIPS*, 2020. 1, 2, 3

[41] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 2, 4

[42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 4, 7

[43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[44] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 1

[45] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 1

[46] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1

[47] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *WACV*, 2021. 2

[48] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. In *NeurIPS*, 2020. 3

[49] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv*, 2018. 2

[50] Jae Hyun Lim and Jong Chul Ye. Geometric gan. In *arXiv*, 2017. 4

[51] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: generation by parts via conditional coordinating. In *ICCV*, 2019. 2

[52] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 1

[53] Guilin Liu, Rohan Taori, Ting-Chun Wang, Zhiding Yu, Shiqiu Liu, Fitsum A Reda, Karan Sapra, Andrew Tao, and Bryan Catanzaro. Transposer: Universal texture synthesis using feature maps as transposed convolution filter. *arXiv*, 2020. 1

[54] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 1

[55] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 7

[56] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 2, 3

[57] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 7

[58] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 1, 2

[59] Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, and Bryan Catanzaro. Neural ffts for universal texture image synthesis. In *NeurIPS*, 2020. 1

[60] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 1, 2, 4

[61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018. 2, 3

[62] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020. 1

[63] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 1, 2, 3

[64] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1

[65] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 2, 5

[66] Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurelien Lucchi. Convolutional generation of textured 3d meshes. In *NeurIPS*, 2020. 3

[67] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 1, 2, 4

[68] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 1

[69] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *CVPR*, 2020. 1, 2, 7, 8

[70] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. In *CVPR*, 2020. 1

[71] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. In *ICLR*, 2017. 1

[72] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *NeurIPS*, 2017. 2

[73] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2

[74] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *CVPR*, 2019. 1

[75] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020. 3

[76] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv*, 2019. 2, 3

[77] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, 2018. 1

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 4

[79] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 2

[80] Hui-Po Wang, Ning Yu, and Mario Fritz. Hijack-gan: Unintended-use of pretrained, black-box gans. In *CVPR*, 2021. 1

[81] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020. 1

[82] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 1, 2

[83] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 4, 5, 6

[84] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 2

[85] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. In *arXiv*, 2020. 2, 4, 5, 6

[86] Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *ICLR*, 2019. 2

[87] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[88] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[89] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019. 2

[90] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2

[91] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv*, 2015. 1, 7

[92] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 1

[93] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 1

[94] Ning Yu, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Michal Lukác. Texture mixer: A network for controllable synthesis and interpolation of texture. In *CVPR*, 2019. 1

[95] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019. 1

[96] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *ECCV*, 2020. 1, 2

[97] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015. 2

[98] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 1, 2, 4, 5, 6

[99] 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 *ICCV*, 2017. 1

[100] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *PAMI*, 2018. 1

[101] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *ICLR*, 2020. 1, 2

[102] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1

[103] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020. 2, 3, 4, 5, 6

[104] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 2

[105] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. *arXiv*, 2020. 1, 2

[106] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. In *arXiv*, 2020. 1, 2, 3

[107] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1

[108] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 1

[109] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 2