Dual Contrastive Loss and Attention for GANs

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

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].
Among these works, contrastive learning is used as an auxiliary task in various unsupervised learning works [26, 61, 76, 8, 9] and generation works [63, 40, 106]. Within the dataset which are referred to as negative examples.

After validating our optimal configuration, we compare it to the state of the art in Section 4.

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:

In Case I:

$$L_{\text{real}}^{\text{contr}}(G, D) = \mathbb{E}_{x \sim p(x)} \left[ \log \frac{e^{D(x)}}{e^{D(x)} + \sum_{z \sim N(0, I_d)} e^{D(G(z))}} \right]$$

$$= - \mathbb{E}_{x \sim p(x)} \left[ \log \left( 1 + \sum_{z \sim N(0, I_d)} e^{D(G(z))} - D(x) \right) \right]$$

(1)

In Case II:

$$L_{\text{fake}}^{\text{contr}}(G, D) = \mathbb{E}_{z \sim N(0, I_d)} \left[ \log \frac{e^{-D(G(z))}}{e^{-D(G(z))} + \sum_{x \sim p(x)} e^{-D(x)}} \right]$$

$$= - \mathbb{E}_{z \sim N(0, I_d)} \left[ \log \left( 1 + \sum_{x \sim p(x)} e^{D(G(z))} - D(x) \right) \right]$$

(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 \(\log(1 + e^{D(x)})\) with a batch of inner terms and using discriminator logit contrasts between real and fake samples. Finally, our adversarial objective is:

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

(3)

Investigation on loss designs. We extensively validate the effectiveness of dual contrastive loss compared to other
we reason the success of the dual loss to its formulation that Wasserstein loss is better than ours on CLEVR dataset, but is worse than the original discriminator representation. We measure much our contrastive representation is 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 \( T \in \mathbb{R}^{h \times w \times c} \) be the input tensor to a convolutional layer in the original architecture. Following the mainstream protocol of self-attention calcula-
We loop over all the \((i, j)\) to constitute an output tensor \(O_{self} \in \mathbb{R}^{h \times w \times c}\) and define it as the self-attention output. Finally, we replace the original convolution output with \(O_{self} \in \mathbb{R}^{h \times w \times c}\), a residual version of this self-attention output.

\[
O_{self} = o(i, j), \forall i = 1, \ldots, h, j = 1, \ldots, w
\]
\[
O_{self} = \text{attn}(K(T), Q(T), V(T))
\]
\[
O_{self} = O_{self} + T
\]

It is worth noting that \(w\) 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. \(w\) 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, \(w\) 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
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.

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 \( T_{\text{ref}} \in \mathbb{R}^{h \times w \times c} \) from the reference branch to calculate the key and query tensors, and use the tensor \( T_{\text{pri}} \in \mathbb{R}^{h \times w \times c} \) 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:

\[
O^{\text{ref}} = \text{attn}(K(T_{\text{ref}}), Q(T_{\text{ref}}), V(T_{\text{pri}})) + T_{\text{pri}} \quad (8)
\]

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
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 StyleGAN [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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Loss</th>
<th>FFHQ</th>
<th>Bedroom</th>
<th>Church</th>
<th>Horse</th>
<th>CLEVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigGAN [5]</td>
<td>Adv</td>
<td>11.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>U-Net GAN [69]</td>
<td>Adv</td>
<td>7.48</td>
<td>17.6</td>
<td>11.7</td>
<td>20.2</td>
<td>33.3</td>
</tr>
<tr>
<td>StyleGAN2 [43]</td>
<td>Adv</td>
<td>4.86</td>
<td>4.01</td>
<td>4.54</td>
<td>3.91</td>
<td>9.62</td>
</tr>
<tr>
<td>StyleGAN2 w/ attn</td>
<td>Adv</td>
<td>5.13</td>
<td>3.48</td>
<td>4.38</td>
<td>3.59</td>
<td>8.96</td>
</tr>
<tr>
<td>StyleGAN2</td>
<td>Contr</td>
<td>3.98</td>
<td>3.86</td>
<td>3.72</td>
<td>3.70</td>
<td>6.06</td>
</tr>
<tr>
<td>StyleGAN2 w/ attn</td>
<td>Contr</td>
<td>4.63</td>
<td>3.31</td>
<td>3.38</td>
<td>2.97</td>
<td>5.05</td>
</tr>
</tbody>
</table>

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 *underline*. “w/ attn” indicates using the self-attention in the generator. “Contr” indicates using our dual contrastive loss instead of conventional GAN loss.

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

[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
[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
[28] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. IJCAI, 2021. 1
[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
[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
[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
[61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv, 2018. 2, 3
[68] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Lopez, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In ICML, 2016. 1


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


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


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


[98] Han Zhang, Tao Xu, Hongsheng Li, Shaoteng 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

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


