Cross-Modal Contrastive Learning for Text-to-Image Generation

Han Zhang*
Google Research
zhanghan@google.com

Jing Yu Koh*†
Google Research
jykoh@google.com

Jason Baldridge
Google Research
jridge@google.com

Honglak Lee‡
University of Michigan
honglak@umich.edu

Yinfei Yang
Google Research
yinfeiy@google.com

Abstract

The output of text-to-image synthesis systems should be coherent, clear, photo-realistic scenes with high semantic fidelity to their conditioned text descriptions. Our Cross-Modal Contrastive Generative Adversarial Network (XMC-GAN) addresses this challenge by maximizing the mutual information between image and text. It does this via multiple contrastive losses which capture inter-modality and intra-modality correspondences. XMC-GAN uses an attentional self-modulation generator, which enforces strong text-image correspondence, and a contrastive discriminator, which acts as a critic as well as a feature encoder for contrastive learning. The quality of XMC-GAN's output is a major step up from previous models, as we show on three challenging datasets. On MS-COCO, not only does XMC-GAN improve state-of-the-art FID from 24.70 to 9.33, but–more importantly–people prefer XMC-GAN by 77.3% for image quality and 74.1% for image-text alignment, compared to three other recent models. XMC-GAN also generalizes to the challenging Localized Narratives dataset (which has longer, more detailed descriptions), improving state-of-the-art FID from 48.70 to 14.12. Lastly, we train and evaluate XMC-GAN on the challenging Open Images data, establishing a strong benchmark FID score of 26.91.

1. Introduction

Compared to other kinds of inputs (*e.g.*, sketches and object masks), descriptive sentences are an intuitive and flexible way to express visual concepts for generating images. The main challenge for text-to-image synthesis lies in learning from unstructured description and handling the different statistical properties between vision and language inputs.

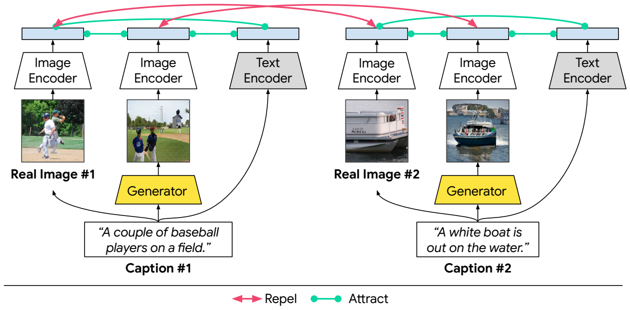


Figure 1: Inter-modal and intra-modal contrastive losses in our proposed XMC-GAN text-to-image synthesis model.

Generative Adversarial Networks (GANs) [12] have shown promising results on text-to-image generation [44, 61, 62], using a conditional GAN formulation [11]. AttnGAN [58] proposes a multi-stage refinement framework to generate fine-grained details by attending to relevant words in the description. These models generate high fidelity images on single domain datasets (*e.g.*, birds [56] and flowers [35]), but struggle on complex scenes with many objects—such as those in MS-COCO [30]. Recent methods [18, 27, 16, 22] propose object-driven, hierarchical approaches that explicitly model object instances within an image. Given the text description, they first infer a semantic layout (*e.g.*, object bounding boxes, segmentation masks, or a combination), and then generate an image from the layout. These hierarchical methods are cumbersome to apply to real-world scenarios; generation becomes a multi-step process (box-to-mask-to-image), and the model requires much more fine-grained object labels to train.

We study contrastive learning in the context of text-to-image synthesis and demonstrate that a simple one-stage GAN *without* object-level annotation can outperform prior object-driven and multi-stage approaches. Besides generating realistic images, we also hope (1) the image should holistically match the description; (2) generated images should match real images when they are conditioned on the

[*] Equal contribution.
[†] Work done as a member of the Google AI Residency program.
[‡] Work performed at Google Research.

same description; (3) individual image regions should be recognizable and consistent with words in the sentence. To fulfill these desiderata and achieve strong language alignment, we propose to maximize the mutual information between the corresponding pairs through contrastive learning. Our method, the Cross(X)-Modal Contrastive Generative Adversarial Network (XMC-GAN), uses image to sentence, image region to word, and image to image contrastive losses to enforce alignment between generated images and their captions (Fig. 1). Our primary contributions include:

- We propose XMC-GAN, a simple one-stage GAN that employs several contrastive losses. XMC-GAN produces dramatic improvements over previous models, e.g. reducing FID [15] from 24.70 to 9.33 on MS-COCO and from 48.70 to 14.12 on LN-COCO (the MS-COCO portion of Localized Narratives [40]).
- We conduct thorough human evaluations comparing XMC-GAN to three recent models. These show that people prefer XMC-GAN 77.3% of the time for image realism, and 74.1% for image-text alignment.
- We establish a strong benchmark on the challenging LN-OpenImages (Open Images subset of Localized Narratives). To the best of our knowledge, this is the first text-to-image results training and testing on the diverse images and descriptions for Open Images.
- We conduct a thorough analysis of contrastive losses used in XMC-GAN to provide general modeling insights for contrastive learning in conditional GANs.

XMC-GAN consistently produces images that are more coherent and detailed than previous models. In addition to greater realism (with clearer, more delineated objects), they better capture the full image description, including the presence of named objects and background compositions.

2. Related Work

Text-to-image synthesis Generating images from text descriptions has been quickly improved with deep generative models, including pixelCNN [55, 45], approximate Langevin sampling [34], variational autoencoders (VAEs) [21, 13] and Generative Adversarial Networks (GANs) [12, 44]. GAN-based models in particular have shown better sample quality [61, 64, 58, 66, 59, 26, 52, 42, 24]. GAN-INT-CLS [44] was the first to use conditional GANs for text to image generation. StackGAN [61, 62] improves this with a coarse-to-fine framework that progressively generates images at different resolutions for high-resolution synthesis. AttnGAN [58] introduces cross-modal attention to better capture details. DM-GAN [66] adaptively refines generated images with a memory module that writes and reads text and image features. MirrorGAN [43] enforces text-image consistency via caption generation on the generated images. SD-GAN [59] proposes word-level conditional batch normalization and dual encoder structure with triplet loss to improve text-image alignment. Compared with the triplet loss, our contrastive loss does not require mining for informative negatives and thus lowers training complexity. CP-GAN [28] proposes an object-aware image encoder and fine-grained discriminator. Its generated images obtain high Inception Score [46]; however, we show it performs poorly when evaluated with the stronger FID [15] metric and in human evaluations (see Sec. 6.1). To create a final high resolution image, these approaches rely on multiple generators and discriminators to generate images at different resolutions. Others have proposed hierarchical models that explicitly generate different objects after inferring semantic layouts [18, 16, 22]. A drawback of these is that they need fine-grained object labels (*e.g.*, object bounding boxes or segmentation maps), so generation is a multi-step process. Compared to these multi-stage and multi-step frameworks, our proposed XMC-GAN only has a single generator and discriminator trained end-to-end, and it generates much higher quality images.

Contrastive learning and its use in GANs Contrastive learning is a powerful scheme for self-supervised representation learning [36, 14, 5, 57]. It enforces consistency of image representations under different augmentations by contrasting positive pairs with negative ones. It has been explored under several adversarial training scenarios [25, 65, 9, 41]. Cntr-GAN [65] uses a contrastive loss as regularization on image augmentations for unconditional image generation. ContraGAN [20] explores contrastive learning for class-conditional image generation. DiscoFaceGAN [9] adds contrastive learning to enforce disentanglement for face generation. CUT [39] proposes patch-based contrastive learning for image-to-image translation by using positive pairs from the same image location in input and output images. Unlike prior work, we use intra-modality (image-image) and inter-modality (image-sentence and region-word) contrastive learning in text-to-image synthesis (Fig. 1).

3. Preliminaries

3.1. Contrastive Representation Learning

Given two random variables v_1 and v_2, often known as *views* of the data, *contrastive learning* aims to find useful representations of v_1 and v_2 by learning a function that measures the dependence of two views [53], *i.e.*, whether samples are from the joint distribution $p(v_1)p(v_2|v_1)$ or the product of the marginals $p(v_1)p(v_2)$. The resulting function is an estimator of the mutual information $I(v_1; v_2)$. As directly maximizing the mutual information is challenging [37, 3, 50], the InfoNCE loss [36] was proposed to maximize a lower bound of the mutual information $I(v_1; v_2)$. Specifically, given a query sample $v_{1,i}$, minimizing the InfoNCE loss is to score the matching positive sample $v_{2,i} \sim$

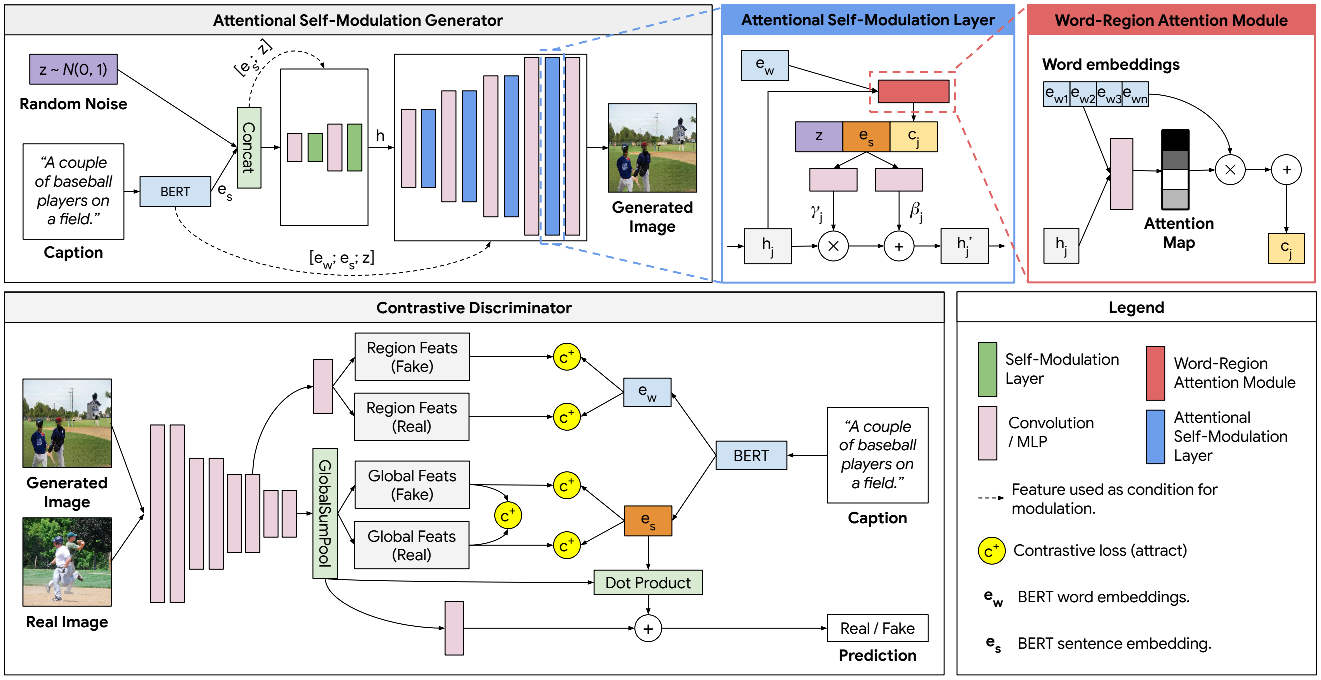

Attentional Self-Modulation Generator
z ~ N(0, 1)
Random Noise
"A couple of baseball players on a field."
Caption
BERT
e_s
[e_s; z]
Concat
h
[e_w; e_s; z]
Generated Image
Attentional Self-Modulation Layer
e_w
z
e_s
c_j
γ_j
β_j
h_j
h_j'
Word-Region Attention Module
Word embeddings
e_w1 e_w2 e_w3 e_wn
h_j
Attention Map
c_j
Contrastive Discriminator
Generated Image
Real Image
Region Feats (Fake)
Region Feats (Real)
Global Feats (Fake)
Global Feats (Real)
GlobalSumPool
c+
e_w
e_s
BERT
"A couple of baseball players on a field."
Caption
Dot Product
Real / Fake
Prediction
Legend
Self-Modulation Layer
Word-Region Attention Module
Convolution / MLP
Attentional Self-Modulation Layer
Feature used as condition for modulation.
Contrastive loss (attract)
e_w BERT word embeddings.
e_s BERT sentence embedding.

Figure 2: Overview of the proposed XMC-GAN.

$p(v_2|v_{1,i})$ higher than $M-1$ negative samples $v_{2,j} \sim p(v_2)$. The overall objective can be summarized as follows:

$$I(v_1; v_2) \geq \log(M) - \mathcal{L}_{NCE},$$
$$\text{where } \mathcal{L}_{NCE} = -\mathbb{E}\left[\log \frac{\exp(\mathcal{S}(v_{1,i}, v_{2,i}))}{\sum_{j=1}^{M} \exp(\mathcal{S}(v_{1,i}, v_{2,j}))}\right].$$

Here, $\mathcal{S}(\cdot, \cdot)$ is the score function, which typically has two parameterized feature encoders for v_1 and v_2. The encoders can share parameters if v_1 and v_2 are from the same domain. There are many ways to construct v_1 and v_2: different augmentations of the same image [14, 5]; spatially adjacent image patches [36]; a video as v_1 and its aligned audio as v_2 for video representation learning [33, 8].

3.2. Generative Adversarial Networks (GANs)

GANs [12] are generative models that employ both a generator and a discriminator. The generator G maps a latent variable $z \sim p(z)$ (usually sampled from a Gaussian distribution) to a real data distribution p_{data}. The discriminator D is trained to distinguish whether inputs are synthesized or sampled from real data. The generator G is trained to synthesize images that the discriminator will classify as real.

A large amount of work has focused on designing the adversarial objective to improve training [12, 1, 31, 47, 29, 54]. A notable example is the hinge loss:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{x \sim p_{\text{data}}}\left[\min(0, -1 + D(x))\right] \\ &-\mathbb{E}_{z \sim p(z)}\left[\min(0, -1 - D(G(z)))\right], \\ \mathcal{L}_G = &-\mathbb{E}_{z \sim p(z)}\left[D(G(z))\right].\end{aligned}$$

The hinge loss has been used in state-of-the-art GANs for image generation [32, 60, 4, 63]. For conditional GANs, the generator and the discriminator are provided with an additional condition c, yielding $G(z, c)$ and $D(x, c)$. For conditional generation, the generated sample should be both realistic and also match the condition c.

4. Method

We describe the losses and components of XMC-GAN below. See Fig. 2 for an overview.

4.1. Contrastive Losses for Text-to-Image Synthesis

Text-to-image synthesis is a conditional generation task. Generated images should both be realistic and well-aligned with a given description. To achieve this, we propose to maximize the mutual information between the corresponding pairs: (1) image and sentence, (2) generated image and real image with the same description, and (3) image regions and words. Directly maximizing mutual information is difficult (see Sec. 3.1), so we maximize the lower bound of the mutual information by optimizing contrastive losses.

Image-text contrastive loss. Given an image x and its corresponding description s, we define the score function following previous work in contrastive learning [14, 5, 36]:

$$\mathcal{S}_{\text{sent}}(x, s) = \cos(f_{\text{img}}(x), f_{\text{sent}}(s))/\tau,$$

where $\cos(u, v) = u^T v/\|u\|\|v\|$ denotes cosine similarity, and τ denotes a temperature hyper-parameter. f_{img} is an image encoder to extract the overall image feature vector and f_{sent} is a sentence encoder to extract the global sentence feature vector. This maps the image and sentence representations into a joint embedding space $\mathbb{R}^D$. The contrastive loss between image x_i and its paired sentence s_i is computed as:

$$\mathcal{L}_{\text{sent}}(x_i, s_i) = -\log \frac{\exp(\cos(f_{\text{img}}(x_i), f_{\text{sent}}(s_i))/\tau)}{\sum_{j=1}^{M} \exp(\cos(f_{\text{img}}(x_i), f_{\text{sent}}(s_j))/\tau)}.$$

This form of contrastive loss is also known as the normalized temperature-scaled cross entropy loss (*NT-Xent*) [5].

Contrastive loss between fake and real images with shared description. This contrastive loss is also defined with *NT-Xent*. The main difference is that a shared image encoder f'_{img} extracts features for both real and fake images. The score function between two images is $\mathcal{S}_{\text{img}}(x, \tilde{x}) = \cos(f'_{\text{img}}(x), f'_{\text{img}}(\tilde{x}))/\tau$. The image-image contrastive loss between real image x_i and generated image $G(z_i, s_i)$ is:

$$\mathcal{L}_{\text{img}}(x_i, G(z_i, s_i)) = -\log \frac{\exp(\mathcal{S}_{\text{img}}(x_i, G(z_i, s_i)))}{\sum_{j=1}^{M} \exp(\mathcal{S}_{\text{img}}(x_i, G(z_j, s_j)))}.$$

Contrastive loss between image regions and words. Individual image regions should be consistent with corresponding words in an input description. We use attention [58] to learn connections between regions in image x and words in sentence s, without requiring fine-grained annotations that align words and regions. We first compute the pairwise cosine similarity matrix between all words in the sentence and all regions in the image; then, we compute the soft attention $\alpha_{i,j}$ for word w_i to region r_j as:

$$\alpha_{i,j} = \frac{\exp(\rho_1 \cos(f_{\text{word}}(w_i), f_{\text{region}}(r_j)))}{\sum_{h=1}^{R} \exp(\rho_1 \cos(f_{\text{word}}(w_i), f_{\text{region}}(r_h)))},$$

where f_{word} and f_{region} represent word and region feature encoders respectively, R is the total number of regions in the image and ρ_1 is a sharpening hyper-parameter to reduce the entropy of the soft attention. The aligned region feature for the i^{th} word is defined as $c_i = \sum_{j=1}^{R} \alpha_{i,j} f_{\text{region}}(r_j)$. The score function between all the regions in image x and all words in sentence s can then be defined as:

$$\mathcal{S}_{\text{word}}(x, s) = \log \Big(\sum_{h=1}^{T} \exp(\rho_2 \cos(f_{\text{word}}(w_h), c_h)) \Big)^{\frac{1}{\rho_2}} /\tau,$$

where T is the total number of words in the sentence. ρ_2 is a hyper-parameter that determines the weight of the most aligned word-region pair, *e.g.*, as $\rho_2 \to \infty$, the score function approximates to $\max_{h=1}^{T} \cos(f_{\text{word}}(w_h), c_h)$. Finally the contrastive loss between the words and regions in image x_i and its aligned sentence s_i can be defined as:

$$\mathcal{L}_{\text{word}}(x_i, s_i) = -\log \frac{\exp(\mathcal{S}_{\text{word}}(x_i, s_i))}{\sum_{j=1}^{M} \exp(\mathcal{S}_{\text{word}}(x_i, s_j))}.$$

Algorithm 1 XMC-GAN Training Algorithm.

```
Input: generator and discriminator parameters $\theta_G, \theta_D$,
    contrastive loss coefficients $\lambda_1$, $\lambda_2$, $\lambda_3$, Adam hyperparameters $\beta_1, \beta_2$,
    generator and discriminator learning rate $lr_G$, $lr_D$, batch size $M$,
    number of discriminator iterations per generator iteration $N_D$
for number of training iterations do
    for $t = 1, ..., N_D$ do
        Sample $\{z_i\}_{i=1}^{M} \sim p(z)$
        Sample $\{(x_i, s_i)\}_{i=1}^{M} \sim p_{\text{data}}(x, s)$
        $\mathcal{L}_{\text{sent}}^{\text{r}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_{\text{sent}}(x_i, s_i)$
        $\mathcal{L}_{\text{word}}^{\text{r}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_{\text{word}}(x_i, s_i)$
        $\mathcal{L}_{\text{GAN}}^{D} \leftarrow -\frac{1}{M}\sum_{i=1}^{M} \min(0, -1 + D(x_i, s_i)) -$
            $\frac{1}{M}\sum_{i=1}^{M} \min(0, -1 - D(G(z_i, s_i), s_i))$
        $\mathcal{L}_D \leftarrow \mathcal{L}_{\text{GAN}}^{D} + \lambda_1 \mathcal{L}_{\text{sent}}^{\text{r}} + \lambda_2 \mathcal{L}_{\text{word}}^{\text{r}}$
        $\theta_D \leftarrow \text{Adam}(\mathcal{L}_D, lr_D, \beta_1, \beta_2)$
    end for
    Sample $\{z_i\}_{i=1}^{M} \sim p(z)$, $\{(x_i, s_i)\}_{i=1}^{M} \sim p_{\text{data}}(x, s)$
    $\mathcal{L}_{\text{sent}}^{\text{f}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_{\text{sent}}(G(z_i, s_i), s_i)$
    $\mathcal{L}_{\text{word}}^{\text{f}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_{\text{word}}(G(z_i, s_i), s_i)$
    $\mathcal{L}_{\text{img}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_{\text{img}}(G(z_i, s_i), x_i)$
    $\mathcal{L}_{\text{GAN}}^{G} \leftarrow \frac{1}{M}\sum_{i=1}^{M} -(D(G(z_i, s_i), s_i))$
    $\mathcal{L}_G \leftarrow \mathcal{L}_{\text{GAN}}^{G} + \lambda_1 \mathcal{L}_{\text{sent}}^{\text{f}} + \lambda_2 \mathcal{L}_{\text{word}}^{\text{f}} + \lambda_3 \mathcal{L}_{\text{img}}$
    $\theta_G \leftarrow \text{Adam}(\mathcal{L}_G, lr_G, \beta_1, \beta_2)$
end for
```

4.2. Attentional Self-Modulation Generator

We propose a one-stage generator to directly generate the image at the desired resolution. This is much simpler than previous multi-stage generators that create images at multiple, different resolutions. We first sample noise z from a standard Gaussian distribution. We obtain the global sentence embedding e_s and the word embeddings e_w from a pretrained BERT [10] module. e_s and z are concatenated to form the global condition, which is passed through several up-sampling blocks (see appendix for details) to generate a 16×16 feature map. The global condition is also used as the condition to calculate scale parameter γ and shift parameter β in conditional batch normalization layers. This formulation is also known as self-modulation [6].

The self-modulation layer improves consistency of the hidden feature with the conditional inputs, but it lacks finer details for each sub-region. To generate fine-grained, recognizable regions, we propose the *attentional self-modulation layer*. Specifically, besides random noise z and global sentence embedding e_s, we modify the attention mechanism [58] to calculate the word-context vector as the additional modulation parameter for each sub-region. For the j^{th} region with feature h_j, the word-context vector c_j is:

$$c_j = \sum_{i=1}^{T} \tilde{\alpha}_{j,i} e_{w_i}, \text{ where } \tilde{\alpha}_{j,i} = \frac{\exp(\rho_0 \cos(e_{w_i}, h_j))}{\sum_{k=1}^{T} \exp(\rho_0 \cos(e_{w_k}, h_j))},$$

where T is the total number of words in the sentence and ρ_0 is a sharpening hyper-parameter. Then, the modulated feature h'_j for the j^{th} region can be defined as:

$$h'_j = \gamma_j(\text{concat}(z, e_s, c_j)) \odot \frac{h_j - \mu}{\sigma} + \beta_j(\text{concat}(z, e_s, c_j)),$$

where μ and σ are the estimated mean and standard deviation from aggregating both batch and spatial dimensions. $\gamma_j(\cdot)$ and $\beta_j(\cdot)$ represent any function approximators; in our work we simply use linear projection layers. Further details of the generator can be found in the appendix.

4.3. Contrastive Discriminator

Our proposed discriminator has two roles: (1) to act as a critic to determine whether an input image is real or fake, and (2) to act as an encoder to compute global image and region features for the contrastive loss. The image is passed through several down-sampling blocks until its spatial dimensions are reduced to 16×16 (see Fig. 2, bottom left). Then, a 1×1 convolution is applied to obtain region features, where the feature dimensions are consistent with the dimensions of the word embedding. The original image feature is fed through two more down-sampling blocks and a global pooling layer. Finally, a projection head computes the logit for the adversarial loss, and a separate projection head computes image features for the image-sentence and image-image contrastive loss. Note that it is important to only use the *real images* and their descriptions to train these discriminator projection heads. The reason is that the generated images are sometimes not recognizable, especially at the start of training. Using such generated image and sentence pairs hurts the training of the image feature encoder projection heads. Therefore, the contrastive losses from *fake images* are only applied to the generator. In addition to the discriminator projection layers, we use a pretrained VGG network [49] as an image encoder for an additional supervisory image-image contrastive loss (see Sec. 6.2). Algorithm 1 summarizes the XMC-GAN training procedure. For simplicity, we set all contrastive loss coefficients ($\lambda_1, \lambda_2, \lambda_3$ in Algorithm 1) to 1.0 in our experiments.

5. Evaluation

5.1. Data

We perform a comprehensive evaluation of XMC-GAN on three challenging datasets (summarized in Table 1).

Dataset	COCO-14		LN-COCO		LN-OpenImages	
	train	val	train	val	train	val
#samples	82k	40k	134k	8k	507k	41k
caption/image	5		1		1	
avg. caption length	10.5		42.1		35.6	

Table 1: Statistics of datasets.

MS-COCO [30] is commonly used for text-to-image synthesis. Each image is paired with 5 short captions. We follow most prior work to use the 2014 split (COCO-14) for evaluation.

Localized Narratives [40] contains long form image descriptions for several image collections. We benchmark results on **LN-COCO**, which contains *narratives* for images in the 2017 split of MS-COCO (COCO-17). Narratives are four times longer than MS-COCO captions on average and they are much more descriptive (see Figure 4). Narratives also contain disfluencies since they are spoken and then transcribed. These factors make text-to-image synthesis for LN-COCO much more challenging than MS-COCO.

We also train and evaluate using **LN-OpenImages**, the Open Images [23] split of Localized Narratives. Its images are both diverse and complex (8.4 objects on average). LN-OpenImages is also much larger than MS-COCO and LN-COCO (see Table 1). To the best of our knowledge, we are the first to train and evaluate a text-to-image generation model for Open Images. XMC-GAN is able to generate high quality results, and sets a strong benchmark for this very challenging task.

5.2. Evaluation Metrics

Following previous work, we report validation results by generating images for 30,000 random captions[1]. We evaluate comprehensively using several measures.

Image quality. We use standard automated metrics for assessing image quality. *Inception Score (IS)* [46] calculates *KL*-divergence between the conditional class distribution and the marginal class distribution given a pre-trained image classifier. *Fréchet Inception Distance (FID)* [15] is the Fréchet distance between two multivariate Gaussians fit to Inception [51] features of generated and real images. While IS and FID have both been shown to correlate with human judgements of generated image quality, IS is likely less informative as it overfits easily and can be manipulated to achieve much higher scores using simple tricks [2, 17]. This is further emphasized by our results (Sec. 6.1) showing that FID correlates better with human judgments of realism.

Text-Image Alignment. Following previous work [58, 27], we use *R-precision* to assess whether a generated image can be used to retrieve its conditioning description. However, we notice that previous work computes R-precision

[1] We oversample the images and captions if there are less than 30,000 samples in the validation set.

Model	IS ↑	FID ↓	R-prec (CC) ↑	SOA-C ↑	SOA-I ↑
Real Images	34.88	6.09	69.36	74.97	80.84
AttnGAN [58]	23.61	33.10	-	25.88	39.01
Obj-GAN [27]	24.09	36.52	-	27.14	41.24
DM-GAN [66]	32.32	27.34	-	33.44	48.03
OP-GAN [17]	27.88	24.70	49.80	35.85	50.47
SD-GAN [59]	35.69	29.35†	51.68	-	-
CP-GAN [28]	**52.73**	55.82‡	59.05	**77.02**	**84.55**
XMC-GAN (ours)	30.45	**9.33**	**71.00**	50.94	71.33

Table 2: Comparison of XMC-GAN with previous models on COCO-14. *R-prec (CC)* are R-precision scores computed from a model trained on Conceptual Captions (see Sec. 5.2). † indicates scores computed from images shared by the original paper authors, and ‡ indicates scores computed from images generated from the open-sourced models.

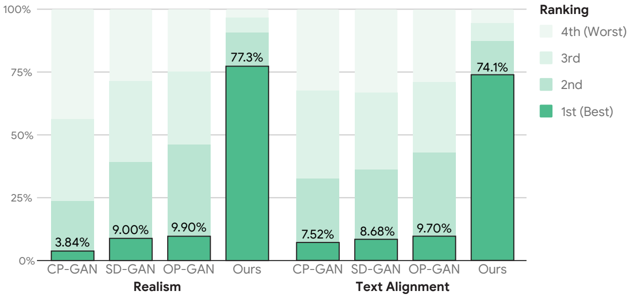


Figure 3: Human evaluation on COCO-14 for image quality and text alignment. Annotators rank (anonymized and order-randomized) generated images from best to worst.

Model	IS ↑	FID ↓	R-prec ↑	SOA-C ↑	SOA-I ↑
Real Images	34.40	8.01	61.52	66.08	67.39
AttnGAN [58]	20.80	51.80	43.88	-	-
TReCS [22]	21.30	48.70	37.88	-	-
XMC-GAN (ours)	**28.37**	**14.12**	**66.92**	36.76	48.14

Table 3: Comparison of XMC-GAN on LN-COCO. SOA metrics together with others are computed from 30,000 random examples.

using image-text encoders from AttnGAN [58], and many others use these encoders as part of their optimization function during training. This skews results: many generated models report R-precision scores significantly higher than real images. To alleviate this, we use an image-text dual-encoder[2] [38] pretrained on *real images* in the Conceptual Captions dataset [48], which is disjoint from MS-COCO. We find that computing R-precision with independent encoders better correlates with human judgments.

Caption retrieval metrics assess whether the entire image matches the caption. In contrast, *Semantic Object Accuracy (SOA)* [17] evaluates the quality of individual regions and objects within an image. Like previous work, we report SOA-C (*i.e.*, the percentage of images per class in which a desired object is detected) and SOA-I (*i.e.*, the percentage of images in which a desired object is detected). Further details of SOA can be found in [17]. SOA was originally designed for COCO-14, and can take very long to compute as it requires generating multiple samples for each MS-COCO class label. We use the official code to compute the metrics reported in Table 2, but approximate results for LN-COCO and other ablation experiments where we compute results over 30,000 random samples.

Human evaluation. Automated metrics are useful while iterating on models during experimentation, but they are no substitute for human eyes. We conduct thorough human evaluations on generated images from 1000 randomly selected captions. For each caption, we request 5 independent human annotators to rank the generated images from best to worst based on (1) realism, and (2) language alignment.

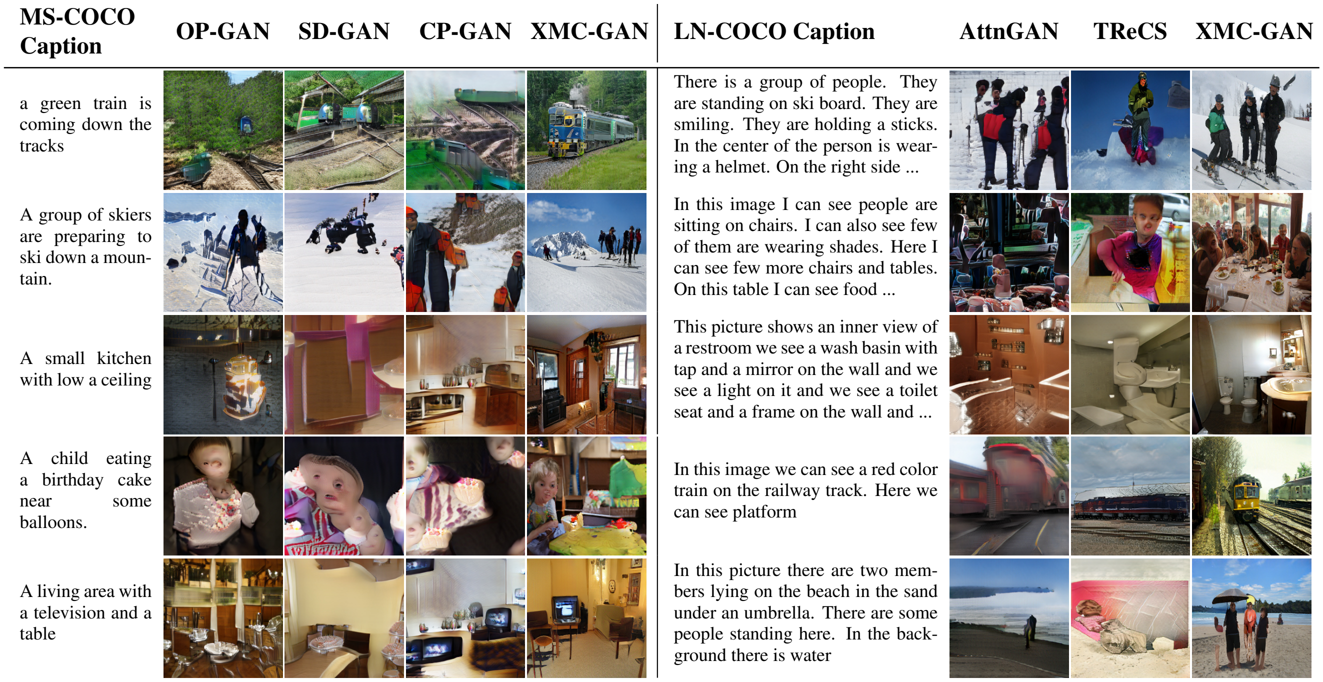
6. Experiments

6.1. Results

COCO-14. Figure 3 shows *human evaluations* comparing XMC-GAN to three recent strong models: CP-GAN [28], SD-GAN [59], and OP-GAN [17]. Given images (anonymized and randomly ordered) generated from the same caption by the four models, annotators are asked to rank them from best to worst. Realism and text alignment judgments are collected independently. XMC-GAN is the clear winner on both: its output is ranked best in 77.3% of realism comparisons, and 74.1% of text alignment ones. OP-GAN is a distant second, at 9.90% and 9.70%, respectively. XMC-GAN achieves this while being a simpler, one-stage model, whereas OP-GAN is multi-stage and needs object bounding boxes. Visual inspection of selected images (Fig. 4) convincingly shows the large quality improvement. XMC-GAN's images are much higher fidelity compared to others, and depict clearer objects and more coherent scenes. This also holds for more random samples (see appendix).

Table 2 provides comprehensive COCO-14 results for *automated* metrics. XMC-GAN dramatically improves FID from 24.70 to 9.33, a 62.2% relative improvement over the next best model, OP-GAN [17]. XMC-GAN also outperforms others (71% vs. 59%) for R-precision computed with our *independently trained* encoders, indicating a large im-

[2] This model will be publicly released to facilitate future evaluations.



a green train is coming down the tracks

A group of skiers are preparing to ski down a mountain.

A small kitchen with low a ceiling

A child eating a birthday cake near some balloons.

A living area with a television and a table

There is a group of people. They are standing on ski board. They are smiling. They are holding a sticks. In the center of the person is wearing a helmet. On the right side ...

In this image I can see people are sitting on chairs. I can also see few of them are wearing shades. Here I can see few more chairs and tables. On this table I can see food ...

This picture shows an inner view of a restroom we see a wash basin with tap and a mirror on the wall and we see a light on it and we see a toilet seat and a frame on the wall and ...

In this image we can see a red color train on the railway track. Here we can see platform

In this picture there are two members lying on the beach in the sand under an umbrella. There are some people standing here. In the background there is water

Figure 4: Generated images for selected examples from COCO-14 and LN-COCO. XMC-GAN generated images are generally of much higher quality and depict clearer scenes. More random samples are available in the appendix.

provement in fidelity of generated images to the captions they are conditioned on—and consistent with human judgments. Although CP-GAN achieves higher IS and SOA scores, both our human evaluations and visual inspection of randomly selected images indicates XMC-GAN's image quality is much higher than CP-GAN's. This may be due to the issue that IS and SOA do not penalize intra-class mode dropping (low diversity within a class)—a model that generates one "perfect" sample for each class can achieve good scores on IS and SOA. Our findings are consistent with other works [27, 2], which suggest that FID may be a more reliable metric for measuring text-to-image synthesis quality.

LN-COCO. Localized Narratives [40] contains much longer descriptions, which increases the difficulty of text-to-image synthesis (see Sec. 5.1). Table 3 shows that XMC-GAN provides massive improvements over prior work. Compared to TReCS [22], XMC-GAN improves IS and FID, by 7.07 and 34.58 (absolute), respectively. It also improves R-precision by 23.04% absolute over AttnGAN [58], indicating much better text alignment. This is supported by qualitative comparison of randomly selected outputs: XMC-GAN's images are decisively clearer and more coherent (see Fig. 4). We stress that TReCS exploits LN-COCO's mouse trace annotations—incorporating this training signal in XMC-GAN in future should further boost performance.

S	W	I	IS ↑	FID ↓	R-prec ↑	SOA-C ↑	SOA-I ↑
Real Images [17]			34.88	6.09	69.36	76.17	80.12
			15.89	39.28	21.41	8.99	25.72
✓			23.50	19.25	53.57	24.57	45.41
	✓		20.72	24.38	44.42	20.50	39.12
		D	18.90	29.71	31.16	12.73	30.89
		VGG	21.54	39.58	35.89	17.41	35.08
		D + VGG	23.61	21.14	47.04	23.87	44.41
✓	✓		26.02	14.25	64.94	30.49	51.60
✓	✓	D	28.06	12.96	65.36	34.21	54.23
✓	✓	VGG	30.55	**11.12**	**70.98**	39.36	59.10
✓	✓	D + VGG	**30.66**	11.93	69.86	**39.85**	**59.78**

Table 4: Ablation results with different contrastive losses on COCO-14. **S** indicates the sentence-image loss. **W** indicates the region-word loss. **I** indicates the image-image loss, where D represents using the discriminator to extract image features, and VGG represents using a pre-trained VGG network to extract image features.

LN-OpenImages. We train XMC-GAN on Open Images dataset, which is much more challenging than MS-COCO due to greater diversity in images and descriptions. XMC-GAN achieves an IS of 24.90, FID of 26.91, and R-precision of 57.55, and manages to generate high quality images (see appendix). To the best of our knowledge, XMC-GAN is the first text-to-image model trained and evaluated on Open Images. Its strong automated scores establish strong benchmark results on this challenging dataset.

6.2. Ablations

We thoroughly evaluate the different components of XMC-GAN and analyze their impact. Table 4 summarizes

Modulation	IS ↑	FID ↓	R-prec ↑	SOA-C ↑	SOA-I ↑
Self-modulation	28.98	13.59	64.65	35.18	55.54
Attentional self-modulation	**30.66**	**11.93**	**69.86**	**39.85**	**59.78**

Table 5: Comparison of different modulation layers.

VGG Loss	IS ↑	FID ↓	R-prec ↑	SOA-C ↑	SOA-I ↑
l_2 loss	12.46	52.86	22.62	8.27	25.48
Contrastive (InfoNCE) loss	**21.54**	**39.58**	**35.89**	**17.41**	**35.08**

Table 6: Comparison of different VGG losses.

our ablations[3] on the COCO-14 validation set. To study the effects of each contrastive loss component used in XMC-GAN, we experiment with four losses: (1) image-sentence, (2) region-word, (3) image-image using discriminator features, and (4) image-image using VGG features. For (3), we use the discriminator encoder projection (indicated by D in Table 4) to extract image features. For (4), we extract image features from a VGG-19 network [49] pretrained on ImageNet.

Individual contrastive losses. Table 4 shows that using any of the contrastive losses improves all metrics compared to the baseline. During experimentation, we also found that including any contrastive loss greatly improves training stability. The largest improvements come from the *inter-modal* image-sentence and region-word contrastive losses, which improve FID from 39.28 to 19.25 and 24.38, respectively. This is much larger compared to the image-image *intra-modal* contrastive losses, *e.g*., including the loss from the discriminator feature encoder (D) only improves FID to 29.71. These ablations highlight the effectiveness of inter-modal contrastive losses: sentence and word contrastive losses each greatly improve the text-alignment metrics, as well as improving image quality.

Combined contrastive losses. Combining contrastive losses provides further gains. For example, using both image-sentence and region-word losses achieves better performance (FID 14.25) than alone (FID 19.25 and 24.38, respectively). This demonstrates that local and global conditions are complementary. Moreover, using both inter-modal losses (sentence and words) outperforms the intra-modal losses (D + VGG): FID scores are 14.25 and 21.14, respectively. These results further emphasize the effectiveness of cross-modal contrastive learning. Nevertheless, the *inter-modal* and *intra-modal* contrastive losses also complement each other: the best FID score comes from combining image-sentence, region-word, and image-image (VGG) losses. Performance on IS and text alignment further improves when using the image-image (D + VGG) loss. To obtain our final results (Table 2), we train a model (with base channels dimension 96) using all 4 contrastive losses.

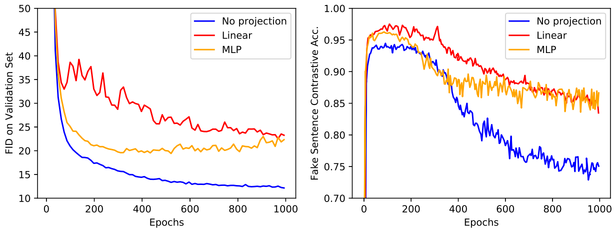


Figure 5: Comparison between different contrastive heads.

Deeper contrastive heads. In unsupervised representation learning [5, 7], adding non-linear layers generally improves performance. To study this, we increase the depth of the projection head in the discriminator. Training curves for FID and contrastive accuracy [5] on fake images are in Fig. 5, across 1000 epochs. We find that using no additional projection layers gives the best FID (12.61, compared to 19.42 of the 2-layer MLP). Moreover, we also find that the contrastive accuracy increases on fake images (from 76.56% to 88.55%) when more layers are added to the projection head. We posit that the discriminator overfits to the contrastive learning task in this configuration, resulting in poorer performance on the adversarial task as a critic and hence worse as a supervisory signal for the generator.

Attentional Self-Modulation. We compare two generator setups: (1) self-modulation layers [6] in all residual blocks, and (2) attentional self-modulation layers (see Sec. 4.2) for blocks with input resolution larger than 16×16. Table 5 shows that the proposed attentional self-modulation layer outperforms self-modulation on all metrics.

Loss types. A frequently used loss function in generative models is the l_2 loss over VGG [49] outputs between fake images and corresponding real images. This is also commonly known as the perceptual loss [19]. Table 6 shows that contrastive losses outperform such perceptual losses. This demonstrates that repelling mismatched samples is more effective than simply pulling together aligned samples. Given this superior performance, replacing perceptual losses with contrastive losses may help other generative tasks.

7. Conclusion

In this work, we present a cross-modal contrastive learning framework to train GAN models for text-to-image synthesis. We investigate several cross-modal contrastive losses that enforce correspondence between image and text. With both human and automated evaluations on multiple datasets, XMC-GAN establishes a marked improvement over previous models: it generates higher quality images that better match their input descriptions, including for long, detailed narratives. It does so while being a simpler, end-to-end model. We believe that these advances are strong leaps towards creative applications for image generation from natural language descriptions.

[3] All ablation results (Fig. 5, Tables 4, 5, and 6) are reported using metrics re-implemented in TensorFlow. SOA is approximated using 30,000 random samples. Ablation models use a reduced base channels dimension of 64. Implementation details are provided in the appendix.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[2] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

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

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

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[8] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP*, 2019.

[9] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, 2020.

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

[11] Jon Gauthier. Conditional generative adversarial networks for convolutional face generation. *Technical report*, 2015.

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

[13] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, 2015.

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

[15] 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.

[16] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. In *ICLR*, 2019.

[17] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *TPAMI*, 2020.

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

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[20] Minguk Kang and Jaesik Park. ContraGAN: Contrastive Learning for Conditional Image Generation. In *NeurIPS*, 2020.

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

[22] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. *WACV*, 2021.

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[24] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, and Thomas Fevens. Dual adversarial inference for text-to-image synthesis. In *ICCV*, 2019.

[25] 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.

[26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. In *NeurIPS*, 2019.

[27] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, 2019.

[28] Jiadong Liang, Wenjie Pei, and Feng Lu. CPGAN: Full-spectrum content-parsing generative adversarial networks for text-to-image synthesis. *ECCV*, 2020.

[29] Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv:1705.02894*, 2017.

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

[31] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *ICCV*, 2017.

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

[33] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *arXiv preprint arXiv:2004.12943*, 2020.

[34] Anh Nguyen, Jason Yosinski, Yoshua Bengio, Alexey Dosovitskiy, and Jeff Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017.

[35] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGIP*, 2008.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[37] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 2003.

[38] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. *arXiv:2004.15020*, 2020.

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

[40] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. *ECCV*, 2020.

[41] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018.

[42] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. In *NeurIPS*, 2019.

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

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

[45] Scott E. Reed, Aäron van den Oord, Nal Kalchbrenner, Sergio Gomez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando de Freitas. Parallel multiscale autoregressive density estimation. In *ICML*, 2017.

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

[47] Tim Salimans, Han Zhang, Alec Radford, and Dimitris N. Metaxas. Improving GANs using optimal transport. In *ICLR*, 2018.

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[50] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *ICLR*, 2020.

[51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[52] Hongchen Tan, X. Liu, Xin Li, Y. Zhang, and B. Yin. Semantics-enhanced adversarial nets for text-to-image synthesis. In *ICCV*, 2019.

[53] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020.

[54] Dustin Tran, Rajesh Ranganath, and David M. Blei. Hierarchical implicit models and likelihood-free variational inference. In *NeurIPS*, 2017.

[55] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016.

[56] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

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

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

[59] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *CVPR*, 2019.

[60] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019.

[61] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[62] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stack-GAN++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 2018.

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

[64] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018.

[65] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for GAN training. *arXiv preprint arXiv:2006.02595*, 2020.

[66] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *CVPR*, 2019.