# CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis

Jiadong Liang[1,†], Wenjie Pei[2,†], and Feng Lu[1,3,⋆]

[1] State Key Lab. of VR Technology and Systems, School of CSE, Beihang University
[2] Harbin Institute of Technology, Shenzhen [3] Peng Cheng Laboratory, Shenzhen

**Abstract.** Typical methods for text-to-image synthesis seek to design effective generative architecture to model the text-to-image mapping directly. It is fairly arduous due to the cross-modality translation. In this paper we circumvent this problem by focusing on parsing the content of both the input text and the synthesized image thoroughly to model the text-to-image consistency in the semantic level. Particularly, we design a memory structure to parse the textual content by exploring semantic correspondence between each word in the vocabulary to its various visual contexts across relevant images during text encoding. Meanwhile, the synthesized image is parsed to learn its semantics in an object-aware manner. Moreover, we customize a conditional discriminator to model the fine-grained correlations between words and image sub-regions to push for the text-image semantic alignment. Extensive experiments on COCO dataset manifest that our model advances the state-of-the-art performance significantly (from **35.69** to **52.73** in Inception Score).

**Keywords:** Text-to-Image Synthesis · Content-Parsing · Generative Adversarial Networks · Memory Structure · Cross-modality

## 1 Introduction

Text-to-image synthesis aims to generate an image according to a textual description. The synthesized image is expected to be not only photo-realistic but also consistent with the description in the semantic level. It has various potential applications such as artistic creation and interactive entertainment. Text-to-image synthesis is more challenging than other tasks of conditional image synthesis like label-conditioned synthesis [29] or image-to-image translation [13]. On one hand, the given text contains much more descriptive information than a label, which implies more conditional constraints for image synthesis. On the other hand, the task involves cross-modality translation which is more complicated than image-to-image translation. Most existing methods [4, 9–11, 17, 32, 34, 39, 43, 44, 49, 50,

[†] Both authors contributed equally.
[⋆] Corresponding Author: Feng Lu (lufeng@buaa.edu.cn)

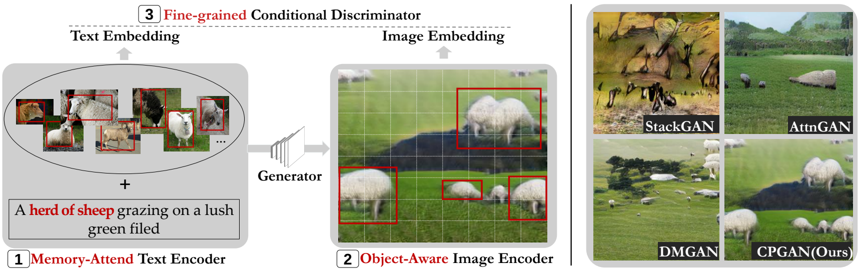

3 Fine-grained Conditional Discriminator
Text Embedding
Image Embedding
+
A herd of sheep grazing on a lush green filed
Generator
1 Memory-Attend Text Encoder
2 Object-Aware Image Encoder
StackGAN
AttnGAN
DMGAN
CPGAN(Ours)

Fig. 1: Our model parses the input text by a customized memory-attended mechanism and parses the synthesized image in an object-aware manner. Besides, the proposed Fine-grained Conditional Discriminator is designed to push for the text-image alignment in the semantic level.Consequently, our CPGAN is able to generate more realistic and more consistent image than other methods.

52], for text-to-image synthesis are built upon the GANs [8], which has been validated its effectiveness in various tasks on image synthesis [2, 25, 48]. A pivotal example is StackGAN [49] which is proposed to synthesize images iteratively in a coarse-to-fine framework by employing stacked GANs. Subsequently, many follow-up works focus on refining this generative architecture either by introducing the attention mechanism [43, 52] or modeling an intermediate representation to smoothly bridge the input text and generated image [10, 11, 17]. Whilst substantial progress has been made by these methods, one potential limitation is that these methods seek to model the text-to-image mapping directly during generative process which is fairly arduous for such cross-modality translation. Consider the example in Figure 1, both StackGAN and AttnGAN can hardly correspond the word 'sheep' to an intact visual picture for a sheep correctly. It is feasible to model the text-to-image consistency more explicitly in the semantic level, which however requires thorough understanding for both text and image modalities. Nevertheless, little attention is paid by these methods to parsing content semantically for either the input text or the generated image. Recently this limitation is investigated by SD-GAN [44], which leverages the Siamese structure in the discriminator to learn semantic consistency between two textual descriptions. However, direct content-oriented parsing in the semantic level for both input text and the generated image is not performed in depth.

In this paper we focus on parsing the content of both the input text and the synthesized image thoroughly and thereby modeling the semantic correspondence between them. On the side of text modality, we design a memory mechanism to parse the textual content by capturing the various visual context information across relevant images in the training data for each word in the vocabulary. On the side of image modality, we propose to encode the generated image in an object-aware manner to extract the visual semantics. The obtained text embeddings and the image embeddings are then utilized to measure the text-image consistency in the semantic space. Besides, we also design a conditional discriminator to push for the semantic text-image alignment by modeling the fine-grained correlations locally between words and image sub-regions. Thus, a full-spectrum content parsing is performed by the resulting model, which we refer to as Content-Parsing Generative Adversarial Networks (CPGAN), to better align the input text and the generated image semantically and thereby improve the performance of text-to-image synthesis. Going back to the example in Figure 1, our CPGAN successfully translates the textual description 'a herd of sheep grazing on a greed field' to a correct visual scene, which is more realistic than the generated results of other methods. We evaluate the performance of our CPGAN on COCO dataset both quantitatively and qualitatively, demonstrating that CPGAN pushes forward the state-of-the-art performance by a significant step. Moreover, the human evaluation performed on a randomly selected subset from COCO test set consistently shows that our model outperforms other two methods(StackGAN and AttnGAN). To conclude, the idea of our CPGAN to **parse the content on both the text side (by MATE, in Sec. 3.2) and the image side (by OAIE, in Sec. 3.3)** is novel, which tackles the cross-modality semantic alignment problem effectively and clearly distinguishes our CPGAN from existing methods. Along with a customized fine-grained conditional discriminator (FGCD, in Sec. 3.4), the CPGAN pushes forward the state-of-the-art performance significantly, from 35.69 to 52.73 in Inception Score.

## 2 Related Work

**Text-to-Image Synthesis.** Text-to-image synthesis was initially investigated based on pixelCNN [35, 37], which suffers from highly computational cost during the inference phase. Meanwhile, the variational autoencoder (VAE) [23] was applied to text-to-image synthesis. A potential drawback of VAE-based synthesis methods is that the generated images by VAE tend to be blurry presumably. This limitation is largely mitigated by the GANs [8], which was promptly extended to various generative tasks in computer vision [2, 13, 25, 48, 51, 28, 3, 20, 19]. After Reed [34] made the first attempt to apply GAN to text-to-image synthesis, many follow-up works [10, 11, 17, 32, 43, 49, 50, 52] focus on improving the generative architecture of GAN to refine the quality of generated images. A well-known example is StackGAN [49, 50], which proposes to synthesize images in a coarse-to-fine framework. Following StackGAN, AttnGAN [43] introduces the attention mechanism which was widely used in computer vision tasks [45, 21] to this framework. DMGAN [52] further refines the attention mechanism by utilizing a memory scheme. MirrorGAN [32] develops a text-to-image-to-text cycle framework to encourage text-image consistency. Another interesting line of research introduces an intermediate representation as a smooth bridge between the input text and the synthesized image [10, 11, 17, 46]. To improve the semantic consistency between the generated image and the input text, ControlGAN [16] applies the matching scheme of DAMSM in AttnGAN [43] in all 3-level discriminators. In contrast, our Fine-Grained Conditional Discriminator (FGCD) proposes a novel discriminator structure to capture the local semantic correlations between each caption word and image regions. Whist these methods have brought about sub-

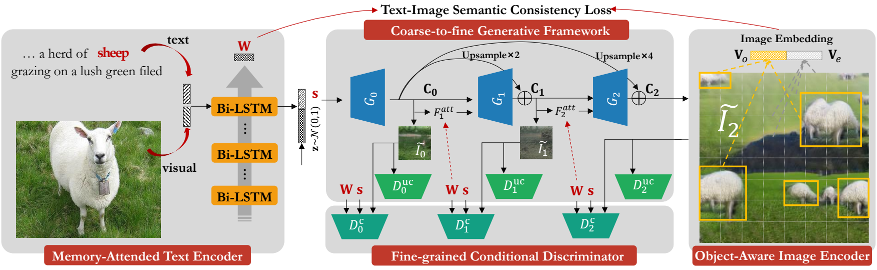


Fig. 2: Architecture of the proposed CPGAN. It follows the coarse-to-fine generative framework. We customize three components to perform content parsing: Memory-Attended Text Encoder for text, Object-Aware Image Encoder for image, and Finegraind Conditional Discriminator for the text-image alignment.

stantial progress, they seek to model the text-to-image mapping directly during generative process. Unlike these methods, we focus on content-oriented parsing of both text and image to obtain a thorough understanding of involved multimodal information. Recently Siamese network is leveraged to explore the semantic consistence either between two textual descriptions by SD-GAN [44] or two images by SEGAN [39]. LeicaGAN [31] adopts text-visual co-embeddings to replace input text with corresponding visual features. Lao et al. [15] parses the input text by learning two variables that are disentangled in the latent space. Text-SeGAN [5] focuses on devising a specific discriminator to regress the semantic relevance between text and image. CKD [47] parses the image content by a hierarchical semantic representation to enhance the semantic consistency and visual quality of synthesized images. However, deep content parsing in the semantic level for both text and image modalities is not performed.

**Memory Mechanism.** Memory networks were first proposed to tackle the limited memory of recurrent networks [14, 38]. It was then extensively applied in tasks of natural language processing (NLP) [6, 7, 24, 41] and computer vision (CV)[22, 26, 30]. Different from the initial motivation of memory networks that is to enlarge the modeling memory, we design a specific memory mechanism to build the semantic correspondence between a word to all its relevant visual features across training data during text parsing.

## 3 Content-Parsing Generative Adversarial Networks

The proposed Content-Parsing Generative Model for text-to-image synthesis focuses on parsing the involved multimodal information by three customized components. To be specific, the Memory-Attended Text Encoder employs the memory structure to explore the semantic correspondence between a word and its various visual contexts; the Object-Aware Image Encoder is designed to parse the generated image in the semantic level; the Fine-grained Conditional Discriminator is proposed to measure the consistency between the input text and the generated image for guiding optimization of the whole model. We will first present the overall architecture of the proposed CPGAN illustrated in Figure 2, which follows the coarse-to-fine generative framework, then we will elaborate on the three aforementioned components specifically designed for content parsing.

### 3.1 Coarse-to-fine Generative Framework

Our proposed model synthesizes the output image from the given textual description in the classical coarse-to-fine framework, which has been extensively shown to be effective in generative tasks [17, 32, 43, 44, 49, 50]. As illustrated in Figure 2, the input text is parsed by our Memory-Attended Text Encoder and the resulting text embedding is further fed into three cascaded generators to obtain coarse-to-fine synthesized images. Two different types of loss functions are employed to optimize the whole model jointly: 1) Generative Adversarial Losses to push the generated image to be realistic and meanwhile match the descriptive text by training adversarial discriminators and 2) Text-Image Semantic Consistency Loss to encourage the text-image alignment in the semantic level.

Formally, given a textual description $X$ containing $T$ words, the parsed text embeddings by the Memory-Attended Text Encoder (Sec. 3.2) are denoted as: $\mathbf{W}, \mathbf{s} = \text{TextEnc}(X)$. Herein $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots \mathbf{w}_T\}$ consists of embeddings of $T$ words in which $\mathbf{w}_t \in \mathbb{R}^d$ denotes the embedding for the $t$-th word. $\mathbf{s} \in \mathbb{R}^d$ is the global embedding for the whole sentence. Three cascaded generators $\{G_0, G_1, G_2\}$ are then employed to sequentially synthesize coarse-to-fine images $\{\widetilde{I}_0, \widetilde{I}_1, \widetilde{I}_2\}$. We apply similar structure as Generative Network in AttnGAN [43]:

$$\widetilde{I}_0, \mathbf{C_0} = G_0(\mathbf{z}, \mathbf{s}), \qquad \widetilde{I}_i, \mathbf{C_i} = G_i(\mathbf{C_{i-1}}, F_i^{att}(\mathbf{W}, \mathbf{C_{i-1}})), i = 1, 2, \tag{1}$$

where $\mathbf{C}_i$ are the generated intermediate feature maps by $G_i$ and $F_i^{att}$ is an attention model designed to attend to the word embeddings $\mathbf{W}$ to each pixel of $\mathbf{C_{i-1}}$ in $i$-th generation stage. Note that the first-stage generator $G_0$ takes as input the noise vector $\mathbf{z}$ sampled from a standard Gaussian distribution to introduce the randomness. In practice, $F_i^{att}$ and $G_i$ are modeled as convolutional neural networks (CNNs), which are elaborated in the supplementary material. Different from AttnGAN, we introduce extra residual connection from $\mathbf{C_0}$ to $\mathbf{C_1}$ and $\mathbf{C_2}$ (via up-sampling) to ease the information propagation between generators.

To optimize the whole model, the generative adversarial losses are utilized by training generators and the corresponding discriminators alternately. In particular, we train two discriminators for each generative stage: 1) an unconditional discriminator $D^{\text{uc}}$ to push the synthesized image to be realistic and 2) a conditional discriminator $D^{\text{c}}$ to align the synthesized image and the input text. The generators are trained by minimizing following adversarial losses:

$$\mathcal{L}_G = \sum_{i=0}^{2} \mathcal{L}_{G_i}, \quad \mathcal{L}_{G_i} = -\frac{1}{2}\mathbb{E}_{\widetilde{I}_i \sim p_{G_i}} D_i^{\text{uc}}(\widetilde{I}_i) - \frac{1}{2}\mathbb{E}_{\widetilde{I}_i \sim p_{G_i}} D_i^{\text{c}}(\widetilde{I}_i, X). \tag{2}$$

Accordingly, the adversarial loss for the corresponding discriminators in the $i$-th generative stage is defined as:

$$\begin{aligned}\mathcal{L}_{D_i} = &\frac{1}{2}\mathbb{E}_{I_i \sim p_{\mathrm{data}_i}}[\max(0, 1 - D_i^{\mathrm{uc}}(I_i))] + \frac{1}{3}\mathbb{E}_{\widetilde{I}_i \sim p_{G_i}}[\max(0, 1 + D_i^{\mathrm{uc}}(\widetilde{I}_i))] \\ &+ \frac{1}{2}\mathbb{E}_{I_i \sim p_{\mathrm{data}_i}}[\max(0, 1 - D_i^{\mathrm{c}}(I_i, X))] + \frac{1}{3}\mathbb{E}_{\widetilde{I}_i \sim p_{G_i}}[\max(0, 1 + D_i^{\mathrm{c}}(\widetilde{I}_i, X))] \\ &+ \frac{1}{3}\mathbb{E}_{I_i \sim p_{\mathrm{data}_i}}[\max(0, 1 + D_i^{\mathrm{c}}(I_i, \overline{X}))],\end{aligned} \tag{3}$$

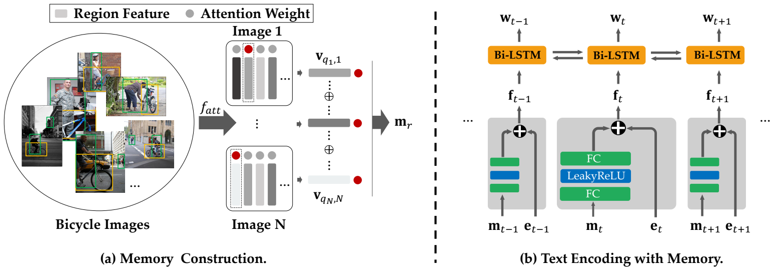
where $X$ is the input descriptive text and $I_i$ is the corresponding groudtruth image for the $i$-th generative stage. The negative pairs $(I_i, \overline{X})$ are also involved to improve the training robustness. Note that we formulate the adversarial losses in the form of Hinge loss rather than the negative log-likelihood due to the empirical superior performance of Hinge loss [25, 48]. The modeling of unconditional discriminator $D_i^{\mathrm{uc}}$ is straightforward by CNNs (check supplementary material for details), it is however non-trivial to design an effective conditional discriminator $D_i^{\mathrm{c}}$. For this reason, we propose the Fine-grained Conditional Discriminator in Section 3.4. While the adversarial losses in Equation 2, 3 push for the text-image consistency in an adversarial manner by the conditional discriminator, Text-Image Semantic Consistency Loss (TISCL) is proposed to optimize the semantic consistency directly. Specifically, the synthesized image and the input text are encoded respectively, then the obtained image embedding and the text embedding are projected to the same latent space to measure their consistency. We adopt DAMSM [43] to compute the non-matching loss between a textual description $X$ and the corresponding image $\widetilde{I}$:

$$\mathcal{L}_{\mathrm{TISCL}}(\widetilde{I}, X) = \mathcal{L}_{\mathrm{DAMSM}}(\mathrm{ImageEnc}(\widetilde{I}), \mathrm{TextEnc}(X)). \tag{4}$$

The key difference between our TISCL and DAMSM lies in encoding mechanisms for both input text (TextEnc) and the synthesized image (ImageEnc). Our proposed Memory-Attended Text Encoder and Object-Aware Image Encoder focus on 1) distilling the underlying semantic information contained in text and image, and 2) capturing the semantic correspondence between them. We will discuss these two encoders in subsequent sections concretely.

### 3.2 Memory-Attended Text Encoder

The Memory-Attended Text Encoder is designed to parse the input text and learn meaningful text embeddings for downstream generators to synthesize realistic images. A potential challenge during text encoding is that a word may have multiple (similar but not identical) visual context information and correspond to more than one relevant images in training data. Typical text encoding methods which encode the text online during training can only focus on the text-image correspondence of the current training pair. Our Memory-Attended Text Encoder aims to capture full semantic correspondence between a word to



(a) Memory Construction.

(b) Text Encoding with Memory.

Fig. 3: (a) The memory $\mathbf{m}_r$ is constructed by considering salient regions from all relevant images across training data. (b) The learned memory and the word embedding are fused via LSTM structure to incorporate temporal information.

various visual contexts from all its relevant images across training data. Thus, our model can achieve more comprehensive understanding for each word in the vocabulary and synthesize images of higher quality with more diversity.

**Memory Construction** The memory is constructed as a mapping structure, wherein each item maps a word to its visual context representation. To learn the meaningful visual features from each relevant image for a given word, we detect salient regions in each image to the word and extract features from them. There are many ways to achieve this goal. We resort to existing models for image captioning, which is the sibling task of text-to-image synthesis, since we can readily leverage the capability of image-text modeling. In particular, we opt for the Bottom-Up and Top-Down(BUTD) Attention model [1] which extracts the salient visual features for each word in a caption at the level of objects. Specifically, given an image-text pair $\langle I, X \rangle$, object detection is first performed on image $I$ by pretrained Yolo-V3 [33] to select top-36 sub-regions (indicated by bounding boxes) w.r.t. the confidence score and the extracted features are denoted as $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{36}\}$. Note that we replace the Faster R-CNN with Yolo-V3 for object detection for computational efficiency. Then the pretrained BUTD Attention model is employed to measure the salience of each of 36 sub-regions for each word in the caption (text) $X$ based on attention mechanism. In practice we only retain the visual feature of the most salient sub-region from each relevant image. Since a word may correspond to multiple relevant images, we extract salient visual features for each of the images the word is involved in. As shown in Figure 3 (a), the visual context features in the memory $\mathbf{m}_r$ for the $r$-th word in the vocabulary is modeled as the weighted average feature:

$$q_n = \operatorname{argmax}_{i=1}^{36} a_{i,n}, \quad \mathbf{m}_r = \frac{\sum_{n=1}^{N} a_{q_n,n} \mathbf{v}_{q_n,n}}{\sum_{n=1}^{N} a_{q_n,n}}, \quad n = 1, \ldots, N, \tag{5}$$

where $N$ is the number of relevant images in the training data to the $r$-th word; $a_{i,n}$ is the attention weight on $i$-th sub-regions for the $n$-th relevant image and $q_n$ is the index of the most salient sub-region of the $n$-th relevant image. To avoid potential feature pollution, we extract features from top-$K$ most relevant images instead of all $N$ images where $K$ is a hyper-parameter tuned on a validation set. The benefits of parsing visual features by such memory mechanism are twofold: 1) extract precise semantic features from the most salient region of relevant images for each word; 2) capture full semantic correspondence between a word to its various visual contexts. It is worth mentioning that both Yolo-V3 and BUTD Attention model are pretrained on MSCOCO dataset [18] which is also used for text-to-image synthesis, hence we do not utilize extra data in our method.

**Text Encoding with Memory** Apart from the learned memory which parses the text from visual context information, we also encode the text by learning latent embedding directly for each word in the vocabulary to characterize the semantic distance among all words. To be specific, we aim to learn an embedding matrix $\mathbf{E} \in \mathbb{R}^{d \times K}$ consisting of $d$-dim embeddings for in total $K$ words in the vocabulary. The learned word embedding $\mathbf{e}_i = \mathbf{E}[:, i]$ for the $i$-th word in the vocabulary is then fused with the learned memory $\mathbf{m}_i$ by concatenation: $\mathbf{f}_i = [\mathbf{e}_i; p(\mathbf{m}_i)]$, where $p(\mathbf{m}_i)$ is a nonlinear projection function to balance the feature dimensions between $\mathbf{m}_i$ and $\mathbf{e}_i$. In practice, we perform $p$ by two fully-connected layers with a LeakReLU layer [42] in between, as illustrated in Figure 3 (b).
Given a textual description $X$ containing $T$ words, we employ a Bi-LSTM [12] structure to obtain final word embedding for each time step, which incorporates the temporal dependencies between words: $\mathbf{W}, \mathbf{s} = \text{Bi-LSTM}(\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_T)$. Herein, $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_T\}$ consists of the embeddings of $T$ words. We use $\mathbf{w}_T$ as the sentence embedding $\mathbf{s}$.

### 3.3 Object-Aware Image Encoder

The Object-Aware Image Encoder is proposed to parse the synthesized image by our generator in the semantic level. The obtained image-encoded features are prepared for the proposed TISCL (Equation 4) to guide the optimization of the whole model by minimizing the semantic discrepancy between the input text and the synthesized image. Thus, the quality of the parsed image features are crucial to the performance of image synthesis by our model.
Besides learning global features of the whole image, typical way of attending to local image features is to extract features from equally-partitioned image sub-regions [43]. We propose to parse the image in object level to extract more physically-meaningful features. In particular, we employ Yolo-V3 (pretrained on MSCOCO) to detect salient bounding boxes with top confidence of object detection and learn features from them, which is exactly same as the corresponding operations by Yolo-V3 in the section of memory construction 3.2. Formally, we extract visual features (1024-dim) of top 36 bounding boxes by Yolo-V3 for a given image $I$, denoted as $\mathbf{V}_o \in \mathbb{R}^{1024 \times 36}$. Another benefit of parsing images in object level is that it is consistent with our Memory-Attended Text Encoder, which parses text based on visual context information in object level.
The synthesized image in the early stage of training process cannot be sufficiently meaningful for performing object (salience) detection by Yolo-V3, which

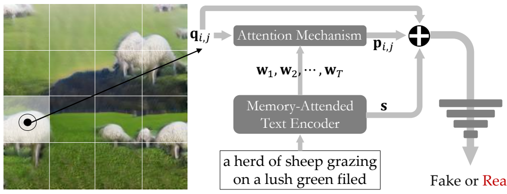

$\mathbf{q}_{i,j}$
Attention Mechanism
$\mathbf{p}_{i,j}$
$\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T$
Memory-Attended Text Encoder
$\mathbf{s}$
a herd of sheep grazing on a lush green filed
Fake or Real

Fig. 4: The structure of Fine-grained Conditional Discriminator.

would adversely affects the image encoding quality. Hence, we also incorporate local features extracted from equally-partitioned sub-regions ($8 \times 8$ in our implementation) like AttnGAN [43], which is denoted as $\mathbf{V}_e \in \mathbb{R}^{768\times 64}$. This kind of two-pronged image encoding scheme is illustrated in Figure 2.

Two kinds of extracted features $\mathbf{V}_o$ and $\mathbf{V}_e$ are then projected into latent spaces with the same dimension by linear transformation and concatenated together to derive the final image encoding features $\mathbf{V}_c$:

$$\mathbf{V}'_o = \mathbf{M}_o\mathbf{V}_o + \mathbf{b}_o, \quad \mathbf{V}'_e = \mathbf{M}_e\mathbf{V}_e + \mathbf{b}_e, \quad \mathbf{V}_c = [\mathbf{V}'_o; \mathbf{V}'_e], \tag{6}$$

where $\mathbf{M}_o \in \mathbb{R}^{256\times 1024}$ and $\mathbf{M}_e \in \mathbb{R}^{256\times 768}$ are transformation matrices. The obtained image encoding feature $\mathbf{V}_c \in \mathbb{R}^{256\times 100}$ is further fed into the DAMSM in Equation 4 to compute the TISCL by measuring the maximal semantic consistency between each word of the input text and different sub-region of the image by attention mechanism[†].

### 3.4 Fine-grained Conditional Discriminator

Conditional discriminator is utilized to distinguish whether a textual caption matches the image in a pair, thus to push the semantic alignment between the synthesized image and the input text by the corresponding adversarial loss. Typical way of designing conditional discriminator is to extract a feature embedding from the text and the image respectively, and then train a discriminator directly on the aggregated features. A potential limitation of such method is that only the global compatibility between the text and the image is considered whereas the local correlations between a word in the text and a sub-region of the image are not explored. Nevertheless, most salient correlations between an image and a caption are always reflected locally. To this end, we propose the Fine-grained Conditional Discriminator, which focuses on modeling local correlations between an image and a caption to measure their compatibility more accurately.

Inspired by PatchGAN [13], we partition the image into $N \times N$ patches and extract visual features for each patch. Then learn the contextual features from the text for each patch by attending to each of the word in the text. As illustrated in Figure 4, suppose the extracted visual features for the $(i, j)$-th patch in the image are denoted as $\mathbf{q}_{i,j}, i, j \in 1, 2, \ldots, N$ and the word features in the

[†] Details are provided in the supplementary file.

text extracted by our text encoder are denoted as $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T\}$. We compute the contextual features for the $(i, j)$-th patch by attention mechanism:

$$a_n = \frac{\exp(\mathbf{q}_{i,j}^\top \mathbf{w}_n)}{\sum_{k=1}^{T} \exp(\mathbf{q}_{i,j}^\top \mathbf{w}_k)}, \quad \mathbf{p}_{i,j} = \sum_{k=1}^{T} a_k \mathbf{w}_k, \quad n = 1, 2, \ldots, T \tag{7}$$

where $a_n$ is the attention weight for $n$-th word in the text. The obtained contextual feature $\mathbf{p}_{i,j}$ is concatenated together with the visual feature $\mathbf{q}_{i,j}$ as well as the sentence embedding $\mathbf{s}$ for the discrimination to be real for fake. Note that the patch size (or the value of $N$) should be tuned to balance between capturing fine-grained local correlations and global text-image correlations.

# 4 Experiments

To evaluate the performance of CPGAN, we conduct experiments on COCO dataset [18] which is a widely used benchmark of text-to-image synthesis.

## 4.1 Experimental Setup

**Dataset.** Following the official 2014 data splits, COCO dataset contains 82,783 images for training and 40,504 images for validation. Each image has 5 corresponding textual descriptions by human annotation. Note that CUB [40] and Oxford-102 [27] are not selected since they are too easy (only one object is contained per image) to fully explore the potential of our model.

**Evaluation Metrics.** We adopt three metrics for quantitative evaluation: Inception score [36], R-precision [43] and SOA [10]. Inception score is extensively used to evaluate the quality of synthesized images taking into account both the authenticity and diversity of images. R-precision is used to measure the semantic consistency between the textual description and the synthesized image. SOA adopts a pre-trained object detection network to measure whether the objects specifically mentioned in the caption are recognizable in the generated images. Specifically, it includes two sub-metrics: SOA-C (average recall w.r.t. object category) and SOA-I (average recall w.r.t. image sample), which are defined as:

$$\text{SOA-C} = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|I_c|} \sum_{i_c \in I_c} \text{Det}(i_c), \quad \text{SOA-I} = \frac{1}{\sum_{c \in C} |I_c|} \sum_{c \in C} \sum_{i_c \in I_c} \text{Det}(i_c), \tag{8}$$

where $C$ and $I_C$ refer to the set of categories and set of images in the category $c$ respectively. $\text{Det}(i_c) \in \{0, 1\}$ indicates whether the pre-trained detector successfully recognizes an object corresponding to class $c$ in the image $i_c$.

**Implementation Details.** Our model is designed based on AttnGAN [43], hence AttnGAN is an important baseline to evaluate our model. We make several minor technical improvements over AttnGAN, which yield much performance gain. Specifically, we replace the binary cross-entropy function for adversarial

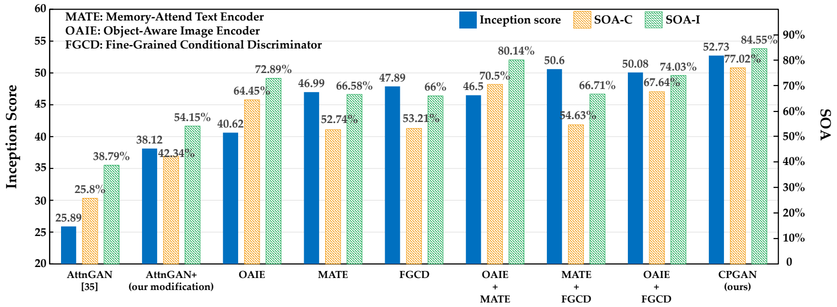


Fig. 5: Performance of ablation study both in Inception score and SOA.

loss with hinge-loss form. Besides, we adopt truncated Gaussian noise [2] as input noise for synthesis ($\mathbf{z}$ in Equation 1). We observe that larger batch size in the training process can also lead to better performance. In our implementation, we use batch size of 72 samples instead of 14 samples in AttnGAN. Finally, the hyper-parameters in AttnGAN are carefully tuned. We call the resulting version based on these improvements as AttnGAN$^+$.

### 4.2 Ablation Study

We first conduct experiments to investigate the effectiveness of our proposed three modules respectively, i.e., Memory-Attended Text Encoder (MATE), Object-Aware Image Encoder (OAIE) and Fine-Grained Conditional Discriminator (FGCD). To this end, we perform ablation experiments which begins with AttnGAN$^+$, and then incrementally augments the text-to-image synthesis system with three modules. Figure 5 presents the performance measured by Inception score and SOA of all ablation experiments.

**AttnGAN$^+$ versus AttnGAN.** It is shown in Figure 5 that AttnGAN$^+$ performs much better than original AttnGAN, which benefits from the aforementioned technical improvements. We observe that increasing the batch size (from 14 to 72) during training brings about the largest performance gain (around 7 points in Inception score). Additionally, fine-tuning the hyper-parameters contributes another 4 points of improvement in Inception score to the performance. Besides, the substantial performance gains in SOA show that AttnGAN$^+$ could synthesis images containing more recognizable objects than AttnGAN.

**Effect of single module.** Equipping the system with each of three proposed modules individually boosts the performance substantially. Compared to AttnGAN$^+$, the performance is improved by 8.9 points, 2.5 points, and 9.8 points by MATE, OAIE and FGCD respectively in Inception score. SOA evaluation results also show large improvements by each of three modules. It is worth noting that OAIE performs best among three modules on SOA metrics emphasizing more on object-level semantics in synthesized images, which in turn validates that OAIE could effectively parse the image in object level. These improvements

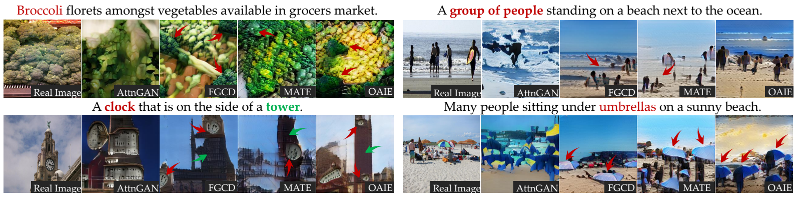

Broccoli florets amongst vegetables available in grocers market.
A group of people standing on a beach next to the ocean.
Real Image
AttnGAN
FGCD
MATE
OAIE
A clock that is on the side of a tower.
Many people sitting under umbrellas on a sunny beach.

Fig. 6: Qualitative comparison between different modules of our model for ablation study, the results of AttnGAN are also provided for reference.

demonstrate the effectiveness of all three modules. Whilst sharing same generators with AttnGAN$^+$, all our three modules focus on parsing the content of the input text or the synthesized image. Therefore, it is implied that deeper semantic content parsing for the text by the memory-based mechanism helps the downstream generators to understand the input text more precisely. On the other hand, our OAIE encourages generators to generate more consistent images with the input text in object level under the guidance of our TISCL. Besides, FGCD steers the optimization of generators to achieve better alignment between the text and the image by the corresponding adversarial losses.

**Effect of combined modules.** We then combine every two of three modules together to further augment the text-to-image synthesis system. The experimental results in Figure 5 indicate that the performances in Inception score are further enhanced compared to the results of single-module cases with the exception of MATE + OAIE. We surmise that this is because MATE performs similar operations as OAIE when learning the visual context information from images in the object level for each word in the vocabulary. Nevertheless, OAIE still advances the performances after being mounted over the single FGCD or MATE + FGCD. It can be observed that combined modules also perform much better than the corresponding single module on SOA metrics. Employing all three modules leads to our full CPGAN model and achieves the best performance in all metrics, which is better than all other single-module or double-module cases.

**Qualitative evaluation.** To gain more insight into effectiveness of our three modules, we visualize the synthesized images for several examples by systems equipped with different modules and the baseline AttnGAN. Figure 6 presents the qualitative comparison. Compared to AttnGAN, the synthesized images by each of our three modules are more realistic and more consistent with the input text, which again reveals advantages of our proposed modules over AttnGAN. Benefiting from the content-oriented parsing mechanisms, our modules tend to generate more intact and realistic objects corresponding to the meaningful words in the input text, which are indicated with red or green arrows.

### 4.3 Comparison with State-of-the-arts

In this set of experiments, we compare our model with the state-of-the-art methods for text-to-image synthesis on COCO dataset.

Table 1: Performance of different text-to-image synthesis models on COCO dataset in terms of Inception score, R-precision SOA-C, SOA-I and model size.

| Model | Inception score | R-precision | SOA-C | SOA-I | #Parameters |
|---|---|---|---|---|---|
| Reed [34] | 7.88 ± 0.07 | – | – | – | – |
| StackGAN [49] | 8.45 ± 0.03 | – | – | – | 996M |
| StackGAN++ [50] | 8.30 ± 0.03 | – | – | – | 466M |
| Lao [15] | 8.94 ± 0.20 | – | – | – | – |
| Infer [11] | 11.46 ± 0.09 | – | – | – | – |
| MirrorGAN [32] | 26.47 ± 0.41 | – | – | – | – |
| SEGAN [39] | 27.86 ± 0.31 | – | – | – | – |
| ControlGAN [16] | 24.06 ± 0.60 | – | – | – | – |
| SD-GAN [44] | 35.69 ± 0.50 | – | – | – | – |
| DMGAN [52] | 30.49 ± 0.57 | 88.56% | 33.44% | 48.03% | 223M |
| AttnGAN [43] | 25.89 ± 0.47 | 82.98% | 25.8% | 38.79% | 956M |
| objGAN[17] | 30.29 ± 0.33 | 91.05% | 27.14% | 41.24% | – |
| OP-GAN[10] | 28.57 ± 0.17 | 87.90% | 33.11% | 47.95% | 1019M |
| AttnGAN$^{+}$<br>(our modification) | 38.12 ± 0.68 | 92.58% | 42.34% | 54.15% | 956M |
| CPGAN (ours) | **52.73** ± 0.61 | **93.59%** | **77.02%** | **84.55%** | 318M |

**Quantitative Evaluation.** Table 1 reports the quantitative experimental results. Our model achieves the best performance in all four metrics and outperforms other methods significantly in terms of Inception score and SOA, which is owing to joint contributions from all three modules we proposed. Particularly, our CPGAN boosts the state-of-the-art performance by 47.74% in inception score, 130.32% in SOA-C and 78.12% in SOA-I. It proves that the synthesized images by our model not only have higher authenticity and diversity, but also are semantically consistent with the corresponding captions in object level. It is worth mentioning that our CPGAN contains much less parameters than StackGAN and AttnGAN, which also follow the coarse-to-fine generative framework. The reduction of model size mainly benefits from two aspects: 1) a negligible amount of parameters are introduced by our proposed MATE and OAIE, 2) The parameter number of three-level discriminators are substantially reduced due to the adoption of Patch-based discriminating behavior in our proposed FGCD.

**Human Evaluation.** As a complement to the standard evaluation metrics, we also perform a human evaluation to compare our model with two classical models: StackGAN and AttnGAN. We randomly select 50 test samples and ask 100 human subjects to compare the quality of synthesized images by these three models and vote for the best for each sample. Note that three models' synthesized results are presented to human subjects randomly for each test sample. We calculate the rank-1 ratio for each model as the comparison metric, presented in Table 2. Averagely, our model achieves 63.73% of votes while AttnGAN wins on 28.33% votes and StackGAN performs worst. This human evaluation result is consistent with the quantitative results in terms of Inception score in Table 1.

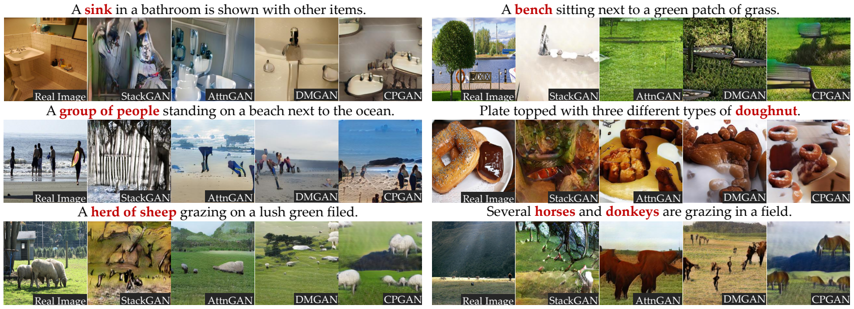

A sink in a bathroom is shown with other items.
A bench sitting next to a green patch of grass.
A group of people standing on a beach next to the ocean.
Plate topped with three different types of doughnut.
A herd of sheep grazing on a lush green filed.
Several horses and donkeys are grazing in a field.
Real Image
StackGAN
AttnGAN
DMGAN
CPGAN

Fig. 7: Qualitative comparison between our CPGAN with other classical models.

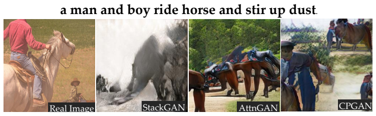


Fig. 8: Challenging examples.

| Model | Rank-1 ratio |
|---|---|
| StackGAN [49] | 7.94% |
| AttnGAN [43] | 28.33% |
| CPGAN(ours) | **63.73%** |

Table 2: Human evaluation results.

**Qualitative Evaluation.** To obtain a qualitative comparison, we visualize the synthesized images on randomly selected text samples by our models and other three classical models: StackGAN, AttnGAN and DMGAN, which is shown in Figure 7. It can be observed that our model is able to generate more realistic images than other two models, like 'sheep' , 'doughnuts' or 'sink'. Besides, the scenes in the generated image by our model are also more consistent with the given text than the other models, such as 'bench next to a patch of grass'.

Image synthesis from text is indeed a fairly challenging task that is far from solved. Take Figure 8 as a challenging example, all models can hardly precisely interpret the the interaction ('ride') between 'man' and 'horse'. Nevertheless, our model still synthesizes more reasonable images than other two methods.

## 5 Conclusions

In this work, we have presented the Content-Parsing Generative Adversarial Networks (CPGAN) for text-to-image synthesis. The proposed CPGAN focuses on content-oriented parsing on both the input text and the synthesized image to learn the text-image consistency in the semantic level. Further, we also design a fine-grained conditional discriminator to model the local correlations between words and image sub-regions to push for the text-image alignment. Our model significantly improves the state-of-the-art performance on COCO dataset.

**Acknowledgements.** This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61972012 and 61732016.

## References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 6077–6086 (2018)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis (2017)
3. Cao, C., Lu, F., Li, C., Lin, S., Shen, X.: Makeup removal via bidirectional tunable de-makeup network. IEEE Transactions on Multimedia(TMM) **21**(11), 2750–2761 (2019)
4. Cha, M., Gwon, Y., Kung, H.: Adversarial nets with perceptual losses for text-to-image synthesis. In: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2017)
5. Cha, M., Gwon, Y.L., Kung, H.: Adversarial learning of semantic relevance in text to image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence(AAAI). vol. 33, pp. 3272–3279 (2019)
6. Das, R., Zaheer, M., Reddy, S., Mccallum, A.: Question answering on knowledge bases and text using universal schema and memory networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(ACL) p. 358–365 (2017)
7. Feng, Y., Zhang, S., Zhang, A., Wang, D., Abel, A.: Memory-augmented neural machine translation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing(EMNLP) p. 1390–1399 (2017)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems(NIPS). pp. 2672–2680 (2014)
9. Hao, D., Yu, S., Chao, W., Guo, Y.: Semantic image synthesis via adversarial learning. In: Proceedings of the IEEE international conference on computer vision(ICCV). pp. 5706–5714 (2017)
10. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. arXiv:1910.13321 (2019)
11. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 7986–7994 (2018)
12. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). pp. 1125–1134 (2017)
14. J. Weston, S.Chopra, A., A.Bordes: Memory networks. International Conference on Learning Representations(ICLR) (2015)
15. Lao, Q., Havaei, M., Pesaranghader, A., Dutil, F., Jorio, L.D., Fevens, T.: Dual adversarial inference for text-to-image synthesis. In: Proceedings of the IEEE International Conference on Computer Vision(ICCV). pp. 7567–7576 (2019)
16. Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. In: Advances in Neural Information Processing Systems(NeurIPS). pp. 2063–2073 (2019)
17. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 12174–12182 (2019)

18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision(ECCV). pp. 740–755 (2014)
19. Liu, Y., Li, Y., You, S., Lu, F.: Unsupervised learning for intrinsic image decomposition from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 3248–3257 (2020)
20. Liu, Y., Lu, F.: Separate in latent space: Unsupervised single image layer separation. In: Proceedings of the AAAI conference on artificial intelligence(AAAI). pp. 11661–11668 (2020)
21. Lv, F., Lu, F.: Attention-guided low-light image enhancement. arXiv preprint arXiv:1908.00682 (2019)
22. Ma, C., Shen, C., Dick, A., Den Hengel, A.V.: Visual question answering with memory-augmented networks. Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR) p. 6975–6984 (2018)
23. Mansimov, E., Parisotto, E., Ba, J., Salakhutdinov, R.: Generating images from captions with attention. International Conference on Learning Representations(ICLR) (2016)
24. Maruf, S., Haffari, G.: Document context neural machine translation with memory networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL) p. 1275–1284 (2018)
25. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. International Conference on Learning Representations(ICLR) (2018)
26. Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., Moschitti, A.: Automatic stance detection using end-to-end memory networks. arXiv preprint arXiv:1804.07581 (2018)
27. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
28. Niu, Y., Gu, L., Lu, F., Lv, F., Wang, Z., Sato, I., Zhang, Z., Xiao, Y., Dai, X., Cheng, T.: Pathological evidence exploration in deep retinal image diagnosis. In: Proceedings of the AAAI conference on artificial intelligence(AAAI). vol. 33, pp. 1093–1101 (2019)
29. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. Proceedings of the 34 rd International Conference on Machine Learning(ICML) p. 2642–2651 (2017)
30. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., Tai, Y.W.: Memory-attended recurrent network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 8347–8356 (2019)
31. Qiao, T., Zhang, J., Xu, D., Tao, D.: Learn, imagine and create: Text-to-image generation from prior knowledge. In: Advances in Neural Information Processing Systems((NeurIPS)). pp. 885–895 (2019)
32. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 4321–4330 (2019)
33. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
34. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. Proceedings of the 33 rd International Conference on Machine Learning(ICML) (2016)

35. Reed, S., Den Oord, A.V., Kalchbrenner, N., Colmenarejo, S.G., Wang, Z., Chen, Y., Belov, D., De Freitas, N.: Parallel multiscale autoregressive density estimation. Proceedings of the 34 rd International Conference on Machine Learning(ICML) pp. 2912–2921 (2017)
36. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems(NIPS). pp. 2234–2242 (2016)
37. Scott Reed, Aaron Van Den Oord, N.K.V.B.M.B.N.D.F.: Generating interpretable images with controllable structure. International Conference on Learning Representations(ICLR) (2017)
38. Sukhbaatar, S., Weston, J., Fergus, R., et al.: End-to-end memory networks. In: Advances in neural information processing systems(NIPS). pp. 2440–2448 (2015)
39. Tan, L., Li, Y., Zhang: Semantics-enhanced adversarial nets for text-to-image synthesis. Proceedings of the IEEE international conference on computer vision(ICCV) p. 10501–10510 (2019)
40. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
41. Wang, S., Mazumder, S., Liu, B., Zhou, M., Chang, Y.: Target-sensitive memory networks for aspect sentiment classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL) pp. 957–967 (2018)
42. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
43. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 1316–1324 (2018)
44. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 2327–2336 (2019)
45. Yu, H., Cai, M., Liu, Y., Lu, F.: What i see is what you see: Joint attention learning for first and third person video co-analysis. In: Proceedings of the 27th ACM International Conference on Multimedia(ACMMM). pp. 1358–1366 (2019)
46. Yuan, M., Peng, Y.: Bridge-gan: Interpretable representation learning for text-to-image synthesis. IEEE Transactions on Circuits and Systems for Video Technology(TCSVT) (2019)
47. Yuan, M., Peng, Y.: Ckd: Cross-task knowledge distillation for text-to-image synthesis. IEEE Transactions on Multimedia(TMM) (2019)
48. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. Proceedings of the 36 rd International Conference on Machine Learning(ICML) (2019)
49. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision(ICCV). pp. 5907–5915 (2017)
50. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence(TPAMI) **41**(8), 1947–1962 (2018)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision(ICCV). pp. 2223–2232 (2017)

52. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis pp. 5802–5810 (2019)