RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis from Prior Knowledge

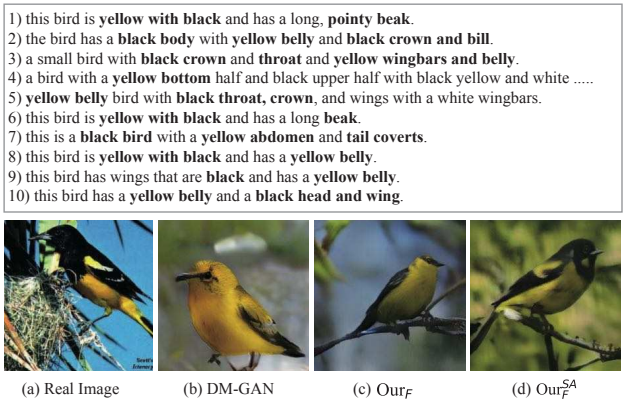
Jun Cheng[1,2], Fuxiang Wu[1,2*], Yanling Tian[3], Lei Wang[1,2], Dapeng Tao[4]
[1] CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, CAS, China
[2] The Chinese University of Hong Kong, Hong Kong, China
[3] Graduate School of Information, Production and Systems, Waseda University, Japan
[4] School of Information Science and Engineering, Yunnan University, China
{jun.cheng, fx.wu1, lei.wang1}@siat.ac.cn, tianyanling@fuji.waseda.jp, dptao@ynu.edu.cn

Abstract

Text-to-image synthesis is a challenging task that generates realistic images from a textual sequence, which usually contains limited information compared with the corresponding image and so is ambiguous and abstractive. The limited textual information only describes a scene partly, which will complicate the generation with complementing the other details implicitly and lead to low-quality images. To address this problem, we propose a novel rich feature generation text-to-image synthesis, called RiFeGAN, to enrich the given description. In order to provide additional visual details and avoid conflicting, RiFeGAN exploits an attention-based caption matching model to select and refine the compatible candidate captions from prior knowledge. Given enriched captions, RiFeGAN uses self-attentional embedding mixtures to extract features across them effectually and handle the diverging features further. Then it exploits multi-captions attentional generative adversarial networks to synthesize images from those features. The experiments conducted on widely-used datasets show that the models can generate images from enriched captions effectually and improve the results significantly.

1. Introduction

Generating realistic images from text descriptions is one of the most active research areas in recent years [10, 12, 16, 21, 24, 26, 33, 37]. Since natural language is one of the easiest ways to interact with people, text-to-image synthesis plays an important role in many areas, like the dual learning mechanism in captioning [31], and has wide potential applications, for instance, art generation, computer-aided design, and early-childhood education. Recently, many methods have focused on improving generators of Generative Adversarial Networks (GANs) [8], like BigGANs [5], and training methods, such as Wasserstein GANs [2, 9], to synthesize high-quality images. However, due to the ambiguity, abstraction and limited information of natural language, one caption lacks of much detailed information of objects. Thus the conditional generator is required to complement those details, which will make the generator complex and its training difficult.



(a) Real Image (b) DM-GAN (c) Our$_F$ (d) Our$_F^{SA}$

Figure 1. Captions and their corresponding images: *(a)* is a real image; *(b)* is generated by DM-GAN [37] with the first caption; *(c)* is synthesized by our model without SAEMs with all captions; *(d)* is synthesized by our model with all captions. The bold words in a caption indicate the prominent features of a bird, and a caption only describes part of the features.

As shown in Fig.1, each caption only describes part of the features of a bird *(a)*. Training with several corresponding captions from the same image, such as captions of *1)-10)*, can be simultaneously exploited to provide more detailed information. Therefore, the generated image *(d)* is closer to the real image, compared with image *(b)* synthesized with only one caption with limited information.

*F. Wu is the corresponding author.

Moreover, a large set of captions is hard to be handled directly to synthesize images. Thus image *(c)* generated by our model without self-attentional embedding mixtures (SAEMs) is worse than image *(d)*. To alleviate the problem of limited information and generate desirable visual details efficiently, additional captions, explicit complement, should be retrieved to enrich the description. To retrieve the compatible captions, we introduce an attention-based caption matching model to select candidate captions from prior knowledge built by the training dataset. The complement captions are selected from the candidates by comparing their embedding and the given one's to improve semantic consistency. For example, given the first caption 1) in Fig. 1, the others can be retrieved as complements to provide additional information.

More captions can provide more visual details, but understanding their semantic meaning will be much harder since even understanding a long caption is not a trivial task. To resolve this dilemma, we transform the complex task of understanding the full captions into a relatively simpler task of understanding one caption and fusing the representations of the captions. Thus, we extract features from each caption via using an attentional model, followed by self-attentional embedding mixtures (SAEMs) to fuse those embeddings.

In summary, this work has the following contributions:

• We propose a novel framework called RiFeGAN for enriching the given caption from prior knowledge, formed by the training dataset, to tackle the problem of limited information and improve the quality of synthesized images.

• We introduce a caption matching method, by using an attentional text-matching model, to retrieve compatible captions from prior knowledge automatically. Then, multi-captions attentional GANs with SAEMs to extract rich features are exploited to synthesize high-quality images. Consequently, we improve the performance on widely-used datasets significantly.

2. Related Work

In this section, we review the recent works of text matching and text-to-image synthesis.

Text Matching: Pang *et al.* [18] model text matching as image recognition. They construct a matching matrix represented the similarities between words and utilize a convolutional neural network to capture matching patterns. Wan *et al.* [29] propose a deep architecture exploiting positional sentence representations, generated by a bidirectional long short term memory (Bi-LSTM), k-Max pooling, and a multi-layer perceptron, to match two sentences. Lee *et al.* [14] propose a Stacked Cross Attention model to align image regions and words and compute the image-text similarity. Yang *et al.* [32] present a fast and strong RE2 with multiple alignment processes to match two sentences.

Most of GANs-based generative methods have achieved great progress on image generation and will be introduced from four aspects as follows:

Generating with One Caption: Reed *et al.* [21] exploit deep symmetric structured joint embedding strategy to create visually-discriminative embeddings of text descriptions and propose an effective conditional GAN to synthesize plausible images according to the embeddings. Zhang *et al.* [35] transform the complex generating problem into several sub-problems and exploit multiple generators arranged in a tree-like structure to synthesize images progressively. Besides, they introduce a conditioning augmentation to stabilize the training process. Zhang *et al.* [36] introduce a single-stream generator architecture, which applies the hierarchical-nested adversarial objectives to regularize mid-level representations, to adapt jointed discriminators better and generate high-resolution images.

Generating with One Caption and Understanding: Qiao *et al.* [19] propose a LeicaGAN to learn and imagine the prior of the diverse objects about semantics, textures, colors, shapes, and layouts. Qiao *et al.* [20] introduce a MirrorGAN to connect the dual tasks, text-to-image synthesis and text captioning, and constrain the regenerated caption aligning with the given caption.

Generating with One Caption Attentionally: Xu *et al.* [30] propose multi-stage AttnGANs that synthesize different parts of a scene by focusing on different words in a caption. At the first stage, the generator uses the sentence embedding to synthesize an image. In the next stage, the word-context features are calculated by an attentional model and fed into the next generator. Based on AttnGANs, Zhu *et al.*. [37] propose the Dynamic Memory Generative Adversarial Network (DM-GAN) to address the problems of heavily relying on the initial images and unchanged text representation in different stages.

Generating with Multi-captions: Sharma *et al.* [24] extend the caption by adding a dialogue describing the scene further. Then the Skip-thought [13] or recurrent neural networks (RNNs) [1, 23] are exploited to get the embedding of the dialogue, and the StackGAN [35] is employed to generate images. Joseph *et al.* [11] propose Cross-Caption Cycle-Consistent (C4Synth) models to synthesize an image from multiple captions. They are inspired by CycleGAN and construct image generators, text captioners, and the discriminators. The models take the noise and the first caption to generate an image, followed by a discriminator to classify it from the real one and a captioner to generate the caption that should be similar to the next real caption. Next, the models iteratively process the remaining captions as described previously by replacing the noise with the output features of the previous generator. Thus, to generate an image, the generator needs to run as many times as the number of captions.

Difference to Existing Works: Chatpainter [24] ex-

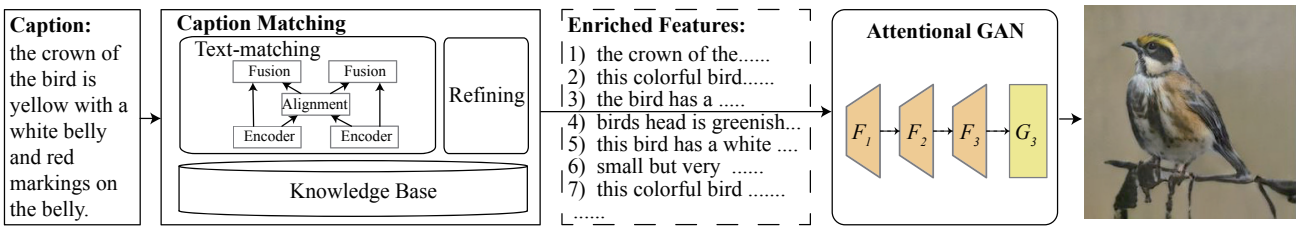


Figure 2. Model structure: given a caption, several captions, enriched features, will be retrieved from the knowledge base by exploiting text matching and refining models.

ploits Skip-thought or RNNs to encode the sentences of dialogue to compute the embedding, then directly feeds the embedding into the StackGAN to synthesize images. Our work utilizes caption matching to enrich the descriptions and exploit SAEMs to extract features from multi-captions in an attentional framework. The results demonstrate the effectivity of our models. C4Synth [11] needs to run many times to synthesize an image, and the models need captioning models to generate captions, which makes the models more complex in training. Different from them, our work directly exploits a caption with enriched or multiple captions so that the full generator only executes only one time per image, and does not need the captioning model to help the training. Moreover, our work, firstly, exploits caption enriching to generate rich features and SAEMs to utilize multi-captions more effectively and achieves significant improvement.

3. Text-to-Image Synthesis with Rich Feature

In Fig 2, given a caption, we firstly enrich it instead of synthesizing images directly. In the caption matching, since there are several captions of an image in common datasets, similar to human memory, we treat each image and its captions of the training part as an item in the knowledge base (memories). Thus, the enriching process will retrieve the compatible items from the knowledge base and refine the captions of the items to return the best complement as the middle part of Fig. 2. Then, given enriched captions, an attentional GAN with SAEMs is introduced to synthesize images with the captions efficiently.

3.1. Caption Matching with Prior Knowledge

Given a caption, caption matching needs to return its compatible captions to enhance it, which are hard tasks in NLP. In order to simplify this problem, we treat the problem as an information retrieval problem and recall the relevant captions from the training dataset. Thus, given a dataset, we treat it as prior knowledge, a knowledge base $\Omega = \{\omega_i\}$, where each item ω_i consists of an image I_i and its captions $\{t_{i,k}\}_{k=0}^{N^T}$. Given a caption t and an item ω_i, to evaluate their compatible score, we exploit RE2 [32], a fast and strong neural architecture for general purpose text matching, to measure their score as,

$$S_{compat}(t, \omega_i) = \frac{1}{N^T} \sum_{k=0}^{N^T} S_{RE2}(t, t_{i,k}) \tag{1}$$

where scorer $S_{RE2}(t_1, t_2)$ returns the matching score of given captions t_1 and t_2. The scorer consists of several encoders, alignment layers, and fusion layers as shown in the second block of Fig. 2. The encoders stack several convolutional networks with the same padding to extract context embedding of words instead of recurrent networks. The alignment layer computes aligned representations of two sequences $\{c_{1,i}\}$ and $\{c_{2,i}\}$ as,

$$\begin{cases} c'_{1,i} = \sum_j \alpha'_{i,j} \cdot c_{2,j} \\ c'_{2,i} = \sum_j \alpha'_{j,i} \cdot c_{1,j} \end{cases} \tag{2}$$

where $\alpha'_{i,j}$ is an attentional weight proportional to the dot product of $c_{1,i}$ and $c_{1,j}$. Fusion layers consist of feed-forward networks to fuse $c'_{*,i}$ and $c_{*,i}$. Then, the multi-layer feed-forward neural network is exploited to return their matching score.

Since the captions $\{t_{i,k}\}_{k=0}^{N^T}$ of item ω_i depict I_i simultaneously, they are compatible with each other. Thus, we construct a positive sample $(t_i, \omega_{i,c})$ by selecting a caption from an item ω_i as t_i randomly, and select the rest of the item as the context $\omega_{i,c}$. Because the captions of different classes are likely conflicted with each other, we construct the negative samples $(t_{r(i)}, \omega_{i,c})$ by selecting a caption $t_{r(i)}$ of $\omega_{r(i)}$, where $r(i) \neq i$ returns a random index to Ω, and the classes of indexed items are different. Therefore, similar to the Pairwise Ranking Loss focusing on relative preferences between items, the compatible score is formed as a logistic regression, and the training loss is as follows,

$$\begin{aligned} L_{compat}(\Omega) = \; & -\tfrac{1}{N^T} \textstyle\sum_{i=0}^{N^T} \sigma(S_{compat}(t_i, \omega_{i,c})) \\ & +\sigma(S_{compat}(t_{r(i)}, \omega_{i,c})) \end{aligned} \tag{3}$$

where σ is a sigmoid function. Given a caption t, the K-best candidate captions, denoted by $\Omega_K(t)$, can be retrieved from Ω by using Eq. 1. To improve the semantic consistency and exclude the conflicted captions further, we refine the captions by selecting N^{test} captions whose embeddings are more closer to that of t than others in terms of cosine similarity.

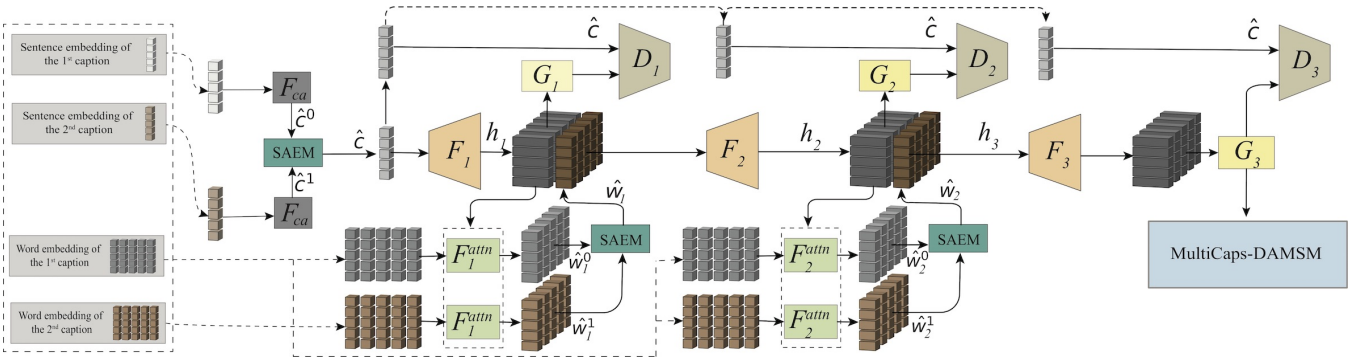


Figure 3. The processing flow of multi-captions attentional GANs with SAEMs: F_i is an upsampling module; G_i and D_i are the generator and discriminator respectively; F_i^{attn} is an attention module taking h_i and word embeddings as input; $SAEM$ and MultiCaps-DAMSM state for self-attentional embedding mixer and multi-captions deep attentional multimodal similarity model respectively.

3.2. Multi-captions Attentional GANs

In text-to-image synthesis, given a caption, its embedding $\boldsymbol{e}$ is created by using a text encoder and fed into conditional GANs to generate images. AttnGAN [30] can efficiently draw different sub-regions with different words. Thus, as described in Fig. 3, we construct the attentional GANs with SAEMs and MultiCaps-DAMSM, which will be elaborated in the following part, to support multi-captions. F_1 is an upsampling module which consists of a fully connected layer, followed by several upsampling modules constructed by an upsampling layer, a 3×3 convolutional layer, a batch normalization layer, and a gated linear layer. F_2 and F_3 are upsampling modules which consist of several residual networks and an upsampling module. The module G_i converts the inner feature $h_i \in R^{N_i \times N_w \times N_h}$ into an image by using a 3×3 convolutional layer and a tanh activation function. D_i is a discriminator constructed by several convolutional layers, batch normalization layers, and leaky rectified linear units. F_i^{attn} is an attention module which takes the word features w and the inner feature h_i as input and computed as follows,

$$F_i^{attn}(h_i, w) = [\textstyle\sum_{k=1}^{T} \alpha_{1,k} w_k, \cdots, \sum_{k=1}^{T} \alpha_{N_3,k} w_k] \quad (4)$$

where $N_3 = N_w \cdot N_h$; T is the length of w; the attentional weights are computed as,

$$\alpha_{j,k} = \frac{\exp(s_{j,k})}{\sum_k (\exp(s_{j,k}))} \quad (5)$$

where $s_{j,k}$ is the dot product of $h_{i,j}$ and w_k; F_{ca} is the Conditioning Augmentation [35] projecting the text embedding into a lower conditional space to enforce the smoothness and encourage robustness.

In Fig. 3, given a set of captions $\mathbf{T} = \{t_j\}_{j=0}^{N_T}$, we exploit text encoders f_{word}^{txt} and f_{cap}^{txt}, which are bi-direction Long Short-Term Memories (LSTMs) [23], to extract the word features and the sentence feature of t_j. In the first stage, F_1 takes $\hat{c}$, the total feature computed by a SAEM, as input to calculate the inner features h_1 and synthesize an image. In the next stage, the attentional model F_1^{attn} takes the word features and h_1 as input to get the attentional features for each caption, followed by a SAEM to compute the total attentional features $\hat{w}_1^j$. Then $\hat{w}_1^j$ and h_1 are combined to synthesize the larger image by F_2 and G_2. The third stage is similar to the second stage except MultiCaps-DAMSM will introduce an additional constraint in training.

3.2.1. Self-Attentional Embedding Mixture

Attention-based models have been applied successfully in many areas, like the dual task, captioning [7, 15]. Moreover, Zhang *et al.* [34] introduce a self-attention mechanism into convolutional GANs and achieve significant improvement in terms of Inception score [22]. Thus, we introduce SAEMs to fuse the embeddings of captions. Given the hidden state h_i, generated by F_i, for each t_j, its corresponding embeddings are calculated as follows,

$$\begin{cases} \hat{c}^j = F_{ca}(f_{cap}^{txt}(t_j)) \\ \hat{w}_i^j = F_i^{attn}(h_i, f_{word}^{txt}(t_j)) \end{cases} \quad (6)$$

where $\hat{c}^j \in R^{N_c}$ is the whole embedding of t_j; $\hat{w}_i^j \in R^{N_i \times N_i \times N_w}$ is the conditional embedding where each element focuses on different words. Therefore, in order to extract the whole embedding of captions **T**, we exploit self-attentional module [27] to fuse those embeddings as follows,

$$\begin{cases} \hat{c} = f_{max}(f_{posw}(L_{MHA}([\hat{c}^0, \hat{c}^1, \cdots, \hat{c}^{N_T}]))) \\ \hat{w}_i = f_{max}(f_{posw}(L_{MHA}([\hat{w}_i^0, \hat{w}_i^1, \cdots, \hat{w}_i^{N_T}]))) \end{cases} \quad (7)$$

where $f_{max}(x)$ returns a tensor, the element of which is maximal across columns of x; f_{posw} is the Position-wise Feed-Forward Networks, $L_{MHA}(v)$ is a multi-head atten-

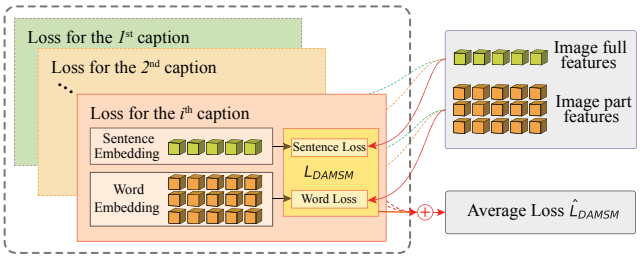


Figure 4. Multi-captions deep attentional multimodal similarity model.

tion layer as follows,

$$L_{MHA}(v) = L_{nl}\left(L_{drop}\left(\begin{bmatrix} H_0(v) \\ \cdots \\ H_{N_H}(v) \end{bmatrix}^T \cdot M_3\right) + v\right) \tag{8}$$

where $M_3 \in R^{N_H \cdot N_i \times N_i}$ is a projecting matrix; L_{nl} is a layer normalization. The attentional function $H_i(E)$ is defined as,

$$H_k(v) = L_{att}(M^k_{4,1} \cdot v, M^k_{4,2} \cdot v, M^k_{4,3} \cdot v) \tag{9}$$

where $M^j_{4,1}, M^j_{4,2}, M^j_{4,3} \in R^{N_i \times N_i}$ are the matrixes projecting the input embedding into question, key, value space, respectively. The function $L_{att}(Q, K, V) = softmax(\beta \cdot Q \cdot K^T) \cdot V$ is the Scaled Dot-Product Attention [27], where β is a scale value to counteract the problem of small gradients.

3.2.2. Multi-Caps DAMSM

As shown in Fig. 4, we consider the captions $\mathbf{T}$ simultaneously and employ the DAMSM [30] to guide the generator. Therefore, the Multi-Caps DAMSM loss is defined as,

$$\hat{L}_{DAMSM}(I, \mathbf{T}) = \sum_{k=0}^{N_T} \mu_k \cdot L_{DAMSM}(I, t_k) \tag{10}$$

where μ_k is the weight of the sentence t_k to indicate its importance. Eq. 10 forces the generated image I to fit the whole descriptions $\mathbf{T}$. $L_{DAMSM}(Q, D)$ is the loss of DAMSM as,

$$\begin{aligned} L_{DAMSM}(I, t_k) = & L^w_1(f^{img}_{part}(I), f^{txt}_{word}(t_k)) \\ & + L^w_2(f^{img}_{part}(I), f^{txt}_{word}(t_k)) \\ & + L^s_1(f^{img}_{full}(I), f^{txt}_{cap}(t_k)) \\ & + L^s_2(f^{img}_{full}(I), f^{txt}_{cap}(t_k)) \end{aligned} \tag{11}$$

where L^w_1, L^w_2 and L^s_1, L^s_2 are the word and sentence loss functions [30] describing the matching probability between the images and their corresponding captions. Given a batch of image-sentence pairs, L^w_1 computes the cross entropy loss of the similarities between images and captions; The similarity between an image I and a caption t is computed by using the cosine similarity between embeddings of words and their corresponding attentional representations extracted from I. L^w_2, L^s_1 and L^s_2 are handled similarly. f^{img}_{part} and f^{img}_{full} extract the sub-region features and the global feature by using an image encoder built upon the Inception-v3 model [25], followed by a 1×1 convolutional layer and a multi-layer perceptron, respectively.

3.2.3. Jointed Training Value Function

An image synthesized by multi-captions should be conformed to those captions. Thus training with that constraint is beneficial to the generation. With the constraint of multi-captions, the total value function can be written as,

$$\begin{aligned} V(D_1, \cdots, D_K, G_1, \cdots, G_K | \mathbf{T}) = & \\ \sum_{i=1}^{K} \{ & \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\mathbf{T})}[\log D_i(\boldsymbol{x}|\mathbf{T})] + \\ & \mathbb{E}_{\hat{\boldsymbol{x}} \sim p_{G_i}(\mathbf{T})}[\log(1 - D_i(\hat{x}|\mathbf{T}))] \} + \\ & \lambda \cdot \mathbb{E}_{\hat{\boldsymbol{x}} \sim p_{G_K}(\mathbf{T})}[\hat{L}_{DAMSM}(\hat{x}, \mathbf{T})]. \end{aligned} \tag{12}$$

where $\hat{\boldsymbol{x}} \sim p_{G_i}(\mathbf{T})$ is the image synthesized by the generator G_i, given the condition $\mathbf{T}$; λ is a hyper-parameter for adjusting the constraint; K is the number of the stages of the generators.

4. Experiments

4.1. Datasets

We conduct experiments on widely-used datasets, Caltech-UCSD Birds-200-2011 (CUB200) [28] and Oxford-Flower-102 (Oxford102) [17] datasets. Each image in the datasets has 10 captions to describe the fine-grained visual details. Following the works [30, 35], we use the same settings and employ the class zero-shot setting. Despite notable flaws [3] of Inception score, we adopt the fine-tuned Inception models[1] to measure the results since it prefers meaningful and diversifying images. In addition to the Inception score, following Xu *et al.* [30], we exploit the R-precision to measure the caption-image alignment. Specifically, given one ground truth caption and 99 mismatching captions selected randomly, if the cosine similarity between the ground truth and the image is higher than that of others, the retrieving is relevant. The R-precision is the ratio of the relevant in retrieving captions. Since the text-to-image synthesize utilizes several captions explicitly or implicitly, we report the average R-precision between an image x and its caption set T. Besides, we analyze the R-precision between an image and its captions as supplementary.

4.2. Quantitative Results

The Inception scores for CUB200 and Oxford102 are shown in Tab. 1, where the baseline systems are taken from AttnGAN [30] and DM-GAN [37]. We evaluate the models with the whole dataset as the knowledge base to retrieve

[1] https://github.com/hanzhanggit/StackGAN-v2

Table 1. Inception scores on CUB200 and Oxford102 datasets

Dataset	AttnGAN	DM-GAN	C4Synth	Our$_F$	Our$_{KB}$	Our$_F^{SA}$	Our$_{KB}^{SA}$
CUB200	4.36 ± 0.03	4.75 ± 0.07	4.07 ± 0.13	4.90 ± 0.07	4.79 ± 0.04	**5.23** ± 0.09	4.85 ± 0.08
Oxford102	3.91 ± 0.05	4.03 ± 0.05	3.52 ± 0.15	4.23 ± 0.03	4.09 ± 0.03	**4.53** ± 0.05	4.23 ± 0.05

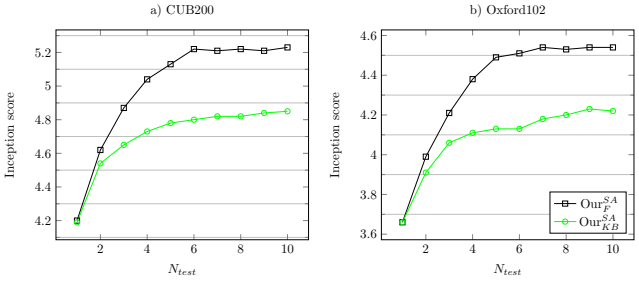


Figure 5. Diagram of Inception score and N_{test}.

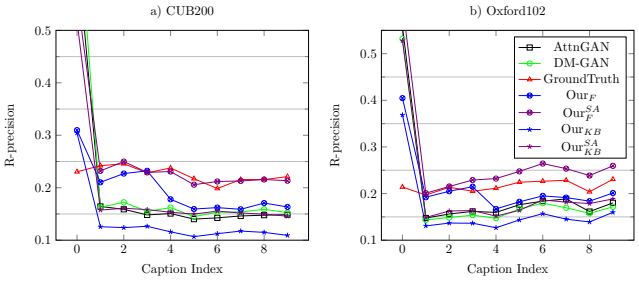


Figure 6. Diagram of R-precision and the caption index.

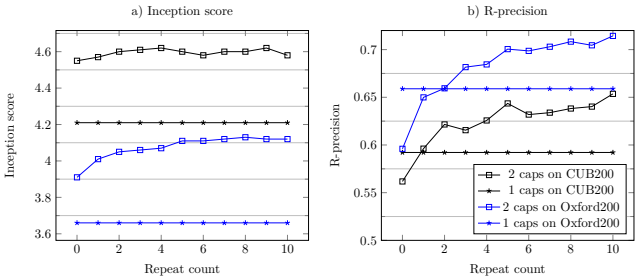


Figure 7. Diagram of Inception score and R-precision influenced by repeat times.

the best matching captions, denoted with the subscript "F". For the class zero-shot setting, we evaluate them with the knowledge base constructed by the training part of the corresponding dataset, denoted with the subscript "KB". Besides, in order to verify the generating ability, we exploit arbitrary N_{test} captions to synthesize images, denoted with the subscript "RND".

For comparison, we construct the multi-captions attentional GAN without SAEMs, Our$_F$ and Our$_{KB}$, by concatenating the multi-captions into a caption and training with the multi-caps DAMSM and jointed training value function. In the setting "F", the Inception scores of our models at least increase by **0.48** on CUB200 and **0.50** on Oxford102 over the baselines. In the setting "KB", the score of Our$_{KB}$ at least increases by **0.10** on CUB200 and **0.20** on Oxford102 over the baselines. Without SAEMs, the scores

Table 2. Average R-precision on CUB200 and Oxford102 datasets

	CUB200	Oxford102
AttnGAN	0.198 ± 0.014	0.203± 0.015
DM-GAN	0.215 ± 0.013	0.199± 0.014
Our$_F$	0.182 ± 0.012	0.213 ± 0.015
Our$_{KB}$	0.130 ± 0.012	0.163± 0.014
Our$_F^{SA}$	**0.238**± **0.015**	**0.267**± **0.016**
Our$_{KB}^{SA}$	0.183± 0.013	0.210 ± 0.014
Ground-truth	*0.225± 0.015*	*0.215± 0.014*

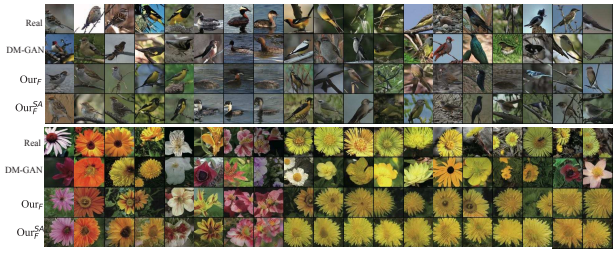


Figure 8. The images synthesized with the captions: Our$_F$ and Our$_F^{SA}$ synthesize more semantically consistent images with the real images than DM-GAN by using the original captions, while Our$_F^{SA}$ is better than Our$_F$.

will drop 0.33 on CUB200 and 0.30 on Oxford102. In Fig. 5, Inception score is in an upward trend and higher with more captions. Besides, we evaluate Our$_F^{SA}$ on MS-COCO, and it largely improves the score from 25.89 to 31.70 with the original captions, comparing with AttnGAN.

In Fig. 6, since the images are generated by using the first caption with the caption index 0, the R-precisions between the images and the 0th captions are higher than the others except for the ground truth. Besides, R-precisions between the images and the other captions are higher than that of the irrelevant selecting, because their models are trained with the pairs of an image and a caption selected from its captions randomly and would memory the visual details of other associated captions. Moreover, the R-precisions between the real image and its captions are about 0.22, which are smaller than those of synthesized images at index 0 and indicates that photo-realistic images should contain more visual details than that included in a caption. Our methods explicitly model that associative process by using multi-captions, and their R-precisions are almost higher than those of the real images for Our$_F^{SA}$, which shows that the synthesized images contain more relevant visual details than the real images depicted by the captions. For Our$_{KB}^{*}$, since they exploit the 0th original caption and the recalling captions, which may be not similar to other original captions, to synthesize images. The synthesized image will be close to those captions which may be different from the

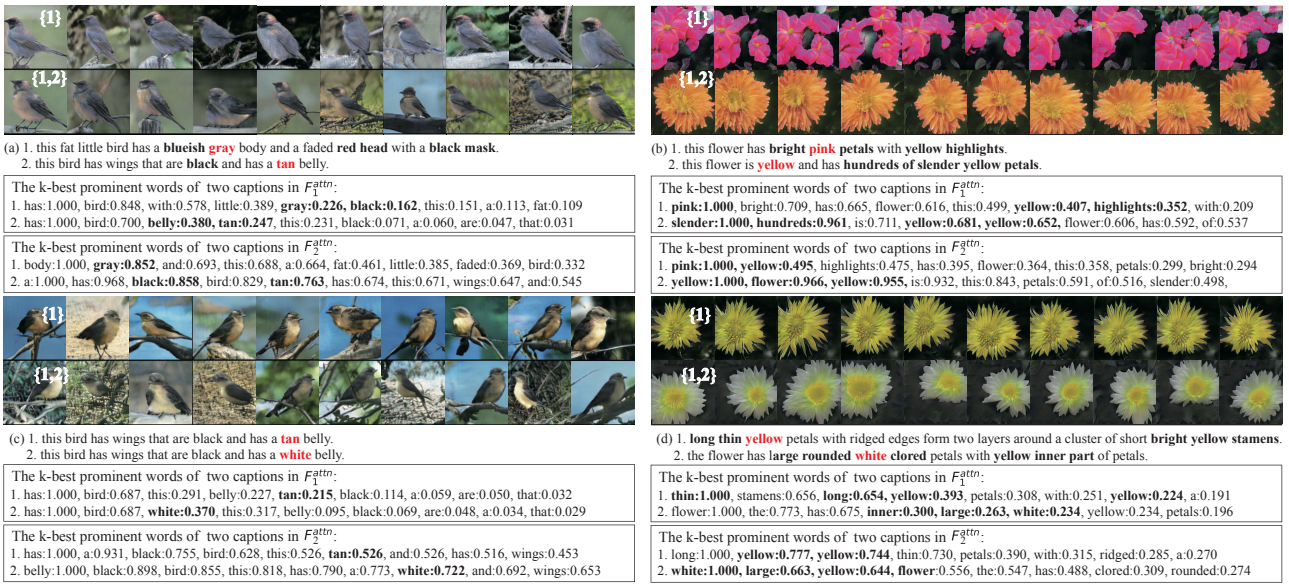


Figure 9. Examples synthesized by two captions: the black bold words indicate prominent visual details, while the red words indicate conflicting visual details in captions. Words in black boxes are prominent features in generating steps, specifically, F_1^{attn} and F_2^{attn}.

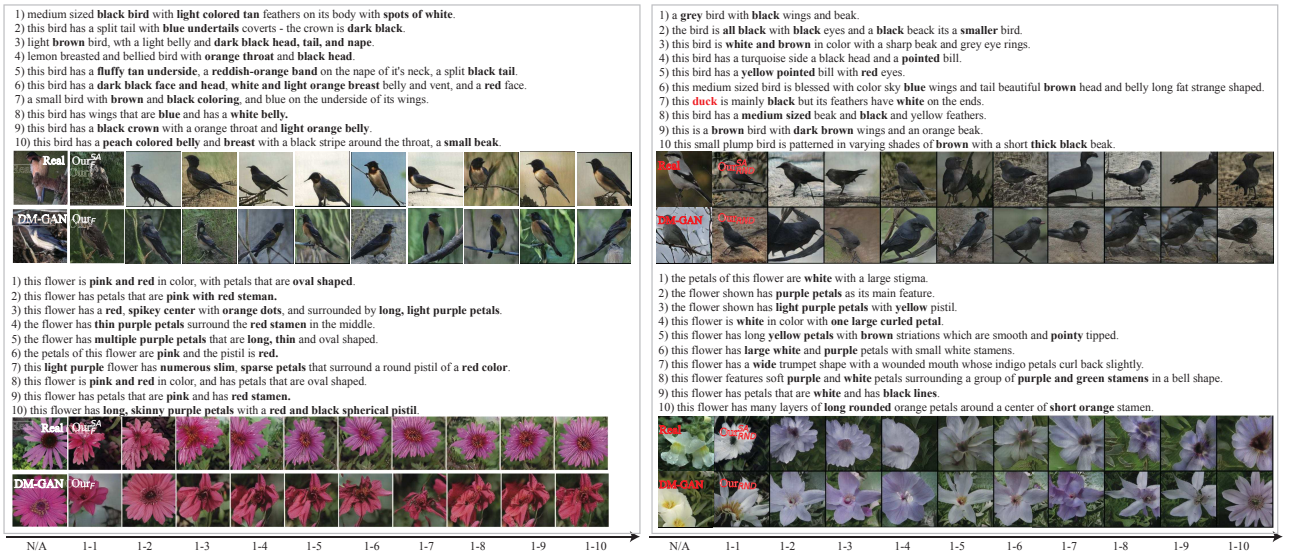


Figure 10. The synthesized images with increasing captions: the bold words in a caption indicate the prominent features, and *1-b* indicates the image is generated with the captions from the first one to the b^{th} one.

original ones because of including additional compatible details. Thus, the R-precisions of the images and the other captions will be lower than those of real images.

In Fig. 7, the R-precision (Inception score) of images synthesized with two captions will be lower (higher) than that with one caption. However, R-precision can be improved by repeating the corresponding caption simply to emphasize it in SAEMs as shown in Fig. 7b). The scores are listed in Tab. 2. The average scores of Our_F^{SA} are higher than others, specifically, increasing by **0.023** on CUB200 and **0.064** on Oxford102 over the baselines. For Our_{KB}^{SA}, the score is higher than those of baselines on Oxford102. On CUB200, since classes of birds are more than that of flowers on Oxford102, the recalling captions may be less relevant to the other original captions than those of flowers, which will result in lower scores compared with that of baselines.

4.3. Qualitative Results

In Fig. 8, the results show that images generated by Our_F and Our_F^{SA}, with the original captions, are more semantically consistent than that synthesized by DM-GAN, and Our_F^{SA} is more stable than Our_F because of the difficulty of extracting the correct visual details from a very long textual sequence. In Fig. 9(a), our models can synthesize the features "black wing" and "tan belly" in the sec-

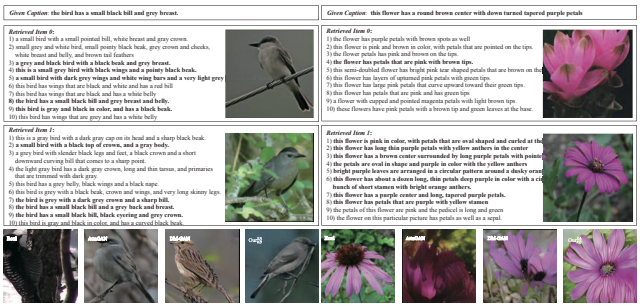


Figure 11. Synthesized examples by exploiting the recalling captions: given a caption, caption matching will retrieve the compacted items and select their captions, masked as bold, to synthesize images.

ond caption while retaining the features "gray body" and "black mask" in the first caption. The Fig. 9(b)-(d) show similar results, which indicates our model can extract features from multi-captions effectively and synthesize images in an incremental manner. Besides, for the captions contained some conflicting features, our models will combine the conflicting features into an intermediate feature. For examples, in Fig. 9(c), the conflicting features "tan belly" and "white belly" will be combined into "pale tan belly", in Fig. 9(b), the inner of petals is pink-yellow; in Fig. 9(d), the weight of "yellow" is smaller than that of "white", namely white:1.000$>$ yellow:0.744, thus the feature "yellow petals" is unapparent. Therefore, our models can handle the conflicting features, and combine them into some reasonable intermediate representation.

In Fig. 10, we demonstrate the generated images by adding more captions gradually, which shows that Our$_F$ and Our$_F^{SA}$ generate more realistic images than the baseline by utilizing more captions. Besides, it is shown that Our$_F^{SA}$ can generate more relevant images than Our$_F$. For example, in the top left example, Our$_F$ synthesizes an image with a blue crown when the 8^{th} caption is touched. However, "black crown/head" is mentioned four times before the 8^{th} caption, which indicates the bird should have a black crown as the images synthesized by our model Our$_F^{SA}$. Therefore, the long textual sequence constructed by concatenating captions will confuse Our$_F$ that it is hard to select and retain the prominent words. Our model Our$_F^{SA}$ can alleviate those issues by considering each caption individually to divide the complex problem into the easier sub-problems. In the right part of Fig. 10, we further exploit randomly selected captions to synthesize images to explore the generating ability. There are many conflicting features in the randomly selected captions, and the results show that our models can synthesize images with reasonable semantics meaning. For example, in the top right example, when the 7th caption is exploited, "duck" is the prominent feature that causes the synthesized bird similar to a duck. When using 1)-8) captions or more captions, the synthesized images retain the "mainly black" feature. However, there may be many conflicting features or unseen combination of some visual features in the randomly selecting captions. Therefore, the synthesized images will degrade and may be worse than the one synthesized by one caption since the large set of those features will not provide useful information to the generator but hinder it from extracting the correct features.

In Fig. 11, we present enriching examples with the given captions. Given a caption "the bird has a small black bill and grey breast.", the matching will retrieve the compact items, "retrieved item 0" and "retrieved item 1" in Fig. 11, in the knowledge base. The features of items contain both "a small black beak" and "gray breast". The images in the last row show that our model can synthesize high-quality images with more compact visual details than that of AttnGAN. For flowers, the "retrieved item 1" include the features of the given caption, while the "retrieved item 0" only contains "pink petals", thus the matching selects more captions of item 1 than that of item 0. The results demonstrate our model can synthesize high-quality images as well.

4.4. Limitation and Discussion

Our models synthesize images based on enriched multi-captions and provide more information to the generator, which alleviates the problem of limited information. The experiments show that it can improve the quality of generated images. Moreover, our model supports incrementally generating, increasing N_{test}, by some interactive operations as the Neural Painter [4], and the SeqAttnGAN [6]. However, images synthesized from multi-captions is not a trivial task. It needs more sophisticated methods in natural language understanding and image synthesis to improve the performance further, such as semantic sentence embeddings [38] and BigGANs [5], which will be our future works.

5. Conclusion

In this paper, to address the problem of limited information on the text descriptions and extract high-quality features, we propose a novel text-to-image synthesis model, RiFeGAN, to enrich the given caption and exploit enriched multi-captions to synthesize images. The experiments conducted on widely-used datasets show that our models can synthesize more realistic images and improve Inception scores significantly. Moreover, the results demonstrate that the models can effectively extract visual details across multi-captions even in the conflicting captions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (U1713213, 61772508, 61772455, U1913202, U1813205), in part by CAS Key Technology Talent Program.

References

[1] Abrar H. Abdulnabi, Bing Shuai, Zhen Zuo, Lap-Pui Chau, and Gang Wang. Multimodal recurrent neural networks with information transfer layers for indoor scene labeling. *IEEE Transactions on Multimedia*, 20(7):1656–1671, 2018.

[2] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In *Proceedings of International Conference on Machine Learning, ICML*, volume 1, pages 298–321, 2017.

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

[4] Ryan Y Benmalek, Claire Cardie, Serge Belongie, Xiadong He, and Jianfeng Gao. The neural painter: Multi-turn image generation. *arXiv preprint arXiv:1806.06183*, 2018.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *Proceedings of International Conference on Learning Representations, ICLR*, 2019.

[6] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. Sequential attention gan for interactive image editing via dialogue. *arXiv preprint arXiv:1812.08352*, 2018.

[7] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems, NeurIPS*, pages 2672–2680, 2014.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of Advances in Neural Information Processing Systems, NeurIPS*, pages 5768–5778, 2017.

[10] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1219–1228, 2018.

[11] KJ Joseph, Arghya Pal, Sailaja Rajanala, and Vineeth N Balasubramanian. C4synth: Cross-caption cycle-consistent text-to-image synthesis. *arXiv preprint arXiv:1809.10238*, 2018.

[12] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

[13] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Proceedings of Advances in Neural Information Processing System, NeurIPS*, pages 3294–3302, 2015.

[14] Kuanghuei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of European Conference on Computer Vision, ECCV*, pages 212–228, 2018.

[15] Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. Gla: Global-local attention for image description. *IEEE Transactions on Multimedia*, 20(3):726–737, 2018.

[16] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. *arXiv preprint arXiv:1812.02784*, 2018.

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of Computer Vision, Graphics & Image Processing, ICVGIP*, pages 722–729.

[18] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *AAAI Conference on Artificial Intelligence*, pages 2793–2799, 2016.

[19] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. In *Advances in Neural Information Processing Systems*, pages 885–895. 2019.

[20] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

[21] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of International Conference on Machine Learning, ICML*, pages 1681–1690, 2016.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of Advances in Neural Information Processing Systems, NeurIPS*, pages 2234–2242, 2016.

[23] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[24] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. Chatpainter: Improving text to image generation using dialogue. In *Proceedings of International Conference on Learning Representations Workshop*, 2018.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2818–2826, 2016.

[26] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating abstract scenes from textual descriptions. *arXiv preprint arXiv:1809.01110*, 2018.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems, NeurIPS*, pages 5998–6008, 2017.

[28] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. (CNS-TR-2011-001), 2011.

[29] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A deep architecture for semantic matching with multiple positional sentence representations.

In *AAAI Conference on Artificial Intelligence*, volume 16, pages 2835–2841, 2016.

[30] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1316–1324, 2018.

[31] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, 21(4):1047–1061, 2019.

[32] Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. Simple and effective text matching with richer alignment features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4699–4709, 2019.

[33] Mingkuan Yuan and Yuxin Peng. Text-to-image synthesis via symmetrical distillation networks. In *Proceedings of ACM Multimedia Conference, ACM MM*, pages 1047–1415, 2018.

[34] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[35] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, doi:10.1109/TPAMI.2018.2856256.

[36] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6199–6208, 2018.

[37] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5802–5810, 2019.

[38] Xunjie Zhu, Tingfeng Li, and Gerard de Melo. Exploring semantic properties of sentence embeddings. In *Proceedings of Annual Meeting of the Association for Computational Linguistics, ACL*, pages 632–637, 2018.