Vision-Language Matching for Text-to-Image Synthesis via Generative Adversarial Networks

Qingrong Cheng, Keyu Wen, and Xiaodong Gu

arXiv:2208.09596v1 [cs.CV] 20 Aug 2022

***Abstract*—Text-to-image synthesis is an attractive but challenging task that aims to generate a photo-realistic and semantic consistent image from a specific text description. The images synthesized by off-the-shelf models usually contain limited components compared with the corresponding image and text description, which decreases the image quality and the textual-visual consistency. To address this issue, we propose a novel Vision-Language Matching strategy for text-to-image synthesis, named VLMGAN*, which introduces a dual vision-language matching mechanism to strengthen the image quality and semantic consistency. The dual vision-language matching mechanism considers textual-visual matching between the generated image and the corresponding text description, and visual-visual consistent constraints between the synthesized image and the real image. Given a specific text description, VLMGAN* firstly encodes it into textual features and then feeds them to a dual vision-language matching-based generative model to synthesize a photo-realistic and textual semantic consistent image. Besides, the popular evaluation metrics for text-to-image synthesis are borrowed from simple image generation, which mainly evaluate the reality and diversity of the synthesized images. Therefore, we introduce a metric named Vision-Language Matching Score (VLMS) to evaluate the performance of text-to-image synthesis which can consider both the image quality and the semantic consistency between synthesized image and the description. The proposed dual multi-level vision-language matching strategy can be applied to other text-to-image synthesis methods. We implement this strategy on two popular baselines, which are marked with VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{DFGAN}}$. The experimental results on two widely-used datasets show that the model achieves significant improvements over other state-of-the-art methods.**

***Index Terms*—Text-to-image synthesis, Generative Adversarial Networks, vision-language matching.**

I. Introduction

Photo-realistic image synthesis [1], [2] has drawn lots of attention in recent years, which has many potential applications, such as art design, computer graphics, and so on. Deep Neural Networks (DNNs) is powerful model for image related tasks, such as image generation [3], [2] and image encoding [4]. Remarkably, Generative Adversarial Networks (GANs) [5] is a milestone for image synthesis. GANs-based methods have achieved incredible results on various image synthesis tasks, especially in high-resolution image generation [6], image style transfer [7] , and text-to-image synthesis [8], [9], [10], [11], [12], [13], [14]. Among them, text-to-image synthesis is one key research sub-direction of Generative Adversarial networks.

Qingrong Cheng, Keyu Wen, and Xiaodong Gu are with Department of Electronic Engineering, Fudan University, Shanghai 200433, China (corresponding author: Xiaodong Gu;email: xdgu@fudan.edu.cn). This work was supported by National Natural Science Foundation of China under grants 61771145 and 61371148.

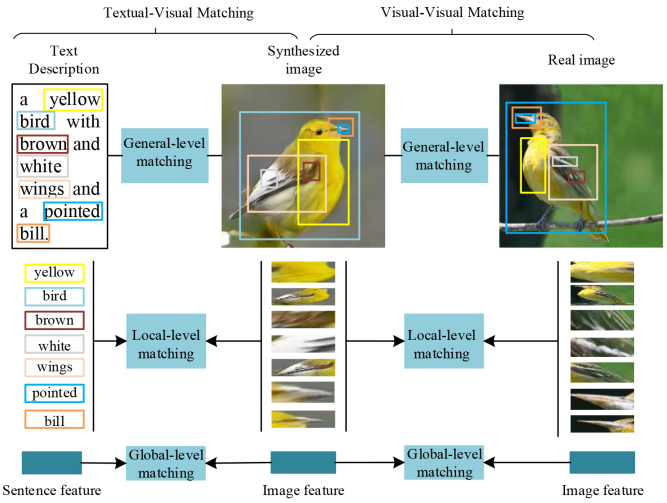


Fig. 1. Illustration of the dual multi-level Vision-Language Matching that presents the basic idea of learning text-to-image synthesis by strengthening the semantic and visual matching of generated image with the real image and the corresponding text description. The bounding boxes indicate different image local-level feature, which are extracted by pre-trained model.

To be specific, text-to-image synthesis aims to generate photo-realistic and text-consistent image based on a specific text description. Image synthesis conditioned on natural language description has become an attractive direction, which presents great potential application in practical. For this task, a large mount of approaches [11], [8], [15], [9], [16], [17], [10], [18], [19], [20], [21], [22] have been proposed to deal with this issue. The technique direction of text-to-image contains traditional DNNs and GANs. The former [23] adopts a DNNs to recover the image like other image reconstructions [24], [2]. Generally, GANs-based text-to-image synthesis has two branches of techniques, one-stage framework and multi-stage framework. One-stage framework follows the conventional GANs framework, which contains only one generator and one discriminator, such as GAN-INT-CLS [11], DFGAN [17]. For example, GAN-INT-CLS [11] concatenates the text embedding vector with a random noise vector and then feeds it into the generator to synthesis text-conditioned image. Multi-stage framework for text-to-image synthesis consists of multiple generators and discriminators, which are stacked in a pipeline, such as StackGAN [21], StackGAN++ [22], DMGAN [10], MA-GAN [25]. Compared with one-stage framework, multi-stage framework [21], [22], [10] is also a popular solution for text-to-image synthesis, which firstly generates relatively blur and low-resolution images and then refines them to photo-

realistic and high-resolution images. Since attention mechanism [26] shows excellent performance in various tasks such as language translation, combining attention mechanism with multi-stage GANs [8] shows excellent performance in text-to-image synthesis. AttnGAN becomes an popular baseline and many researches follow their work, such as [15] , [10], CFA-HAGAN [9]. Cheng et al. proposed CFA-HAGAN [9] for text-to-image synthesis, which contains cross-modal attention and self-attention in the generation framework. SAMGAN [27] also adopts self-attention to support text-to-image synthesis in GAN.

Text-to-image synthesis is significantly different from simple image synthesis, which contains two challenges, visual reality and textual-visual semantic consistency. The visual reality of image synthesis has been fully studied with the development of generative models, especially GANs [5]. Some approaches [1], [6] can generate highly realistic images, which are even difficult for our human to distinguish. The visual-textual semantic consistency is the key challenge for text-to-image synthesis on account of the variegated text description. Although many approaches have ability in synthesizing relatively fine-grained and realistic images, especially for simple datasets such as CUB dataset [28], they rarely concentrate on the multi-level semantic consistency between the generated images and the corresponding texts. Recent approaches can synthesize relatively realistic images while they may fail to generate images that are semantically consistent with the text. For example, for a description "small bird, with white breast, red head and black wings and back", the images synthesized by DMGAN [10] and AttnGAN [8] do not identify with the description especially "red head" as well as the ground truth image although they look realistic, as shown in the fourth column of Figure 4. Therefore, under the condition of photo-reality of image, text-to-image synthesis should focus on both the textual-visual matching and visual-visual matching simultaneously. Textual-visual matching can keep the image content consistent with the text description. The visual-visual matching consider the image quality and semantics of image content.

Besides, how to fairly evaluate the performance of the text-to-image synthesis is a significant issue that needs to be dealt with. As mentioned before, text-to-image synthesis aims at generating both realistic and semantic consistent image. Therefore, the evaluation metric should also include the two aspects, visual reality and textual-visual consistence. The popular evaluation metrics (IS [29] and FID [30]) mainly consider the visual reality, which are widely-used in image generation and image in-painting. To be specific, the IS calculates the KL divergence between the generated data and the original data, which are extracted by the pre-trained Inception-v3 model. The FID measures the Fréchet Distance between synthetic data and real data, which is extracted by the pre-trained Inception-v3 model. As is known, the accepted Inception-V3 is pre-trained by classification task on ImageNet [31] dataset, which contains up to several million images. Besides, most images only contain one object, which is usually located in the center. Therefore, this a gap between the distribution of ImageNet and the chosen datasets. Besides, the text-to-image synthesis is a pairing translation task, which should consider the synthesized image's quality and the semantic consistency with the text description. However, both FID and IS only consider the synthesized images while ignoring the text description. Thus, text-to-image synthesis needs a new evaluation metric which takes the consistency between the description and the synthesized image to make up for deficiencies of FID and IS. Meanwhi, the text-to-image task is suitable for a pairing evaluation metric that can take the two aspect into account.

Motivated by the mentioned observations, we aim to synthesize highly semantically consistent and photo-realistic images from the perspective of dual visual-textual matching and evaluate them under both visual reality and visual-textual semantic consistency. To this end, we first propose a novel vision-language matching model (VLM) that can effectively explore the similarity between image and text based on metric learning. Then, we view the proposed VLM as an additional constraint block and insert it into a multi-stage GANs-based text-to-image synthesis framework. Besides, multi-level matching between the synthesized image and the real image is also considered. Figure 1 shows the basic idea of the proposed method, which aims to strengthen both the textual-visual matching between the synthesized image and the text description and the visual-visual consistency between the synthesized image and the real photo-realistic image.

According to the basic idea, we propose a novel metric for evaluating text-to-image synthesis performance in another view, called Vision-Language Matching Score (VLMS). As mentioned before, text-to-image synthesis focuses on both the image quality and the semantic consistency between the image contents and text description. The proposed VLMS is obtained by a pre-trained Vision-Language Matching model, which is trained by ground-truth image-text pair data with metric learning. Experiments and analyses show that this metric can consider both the visual reality and textual-visual consistency.

The critical contributions of our VLMGAN approach are listed as follows.

1) We design a dual semantically consistent text-to-image synthesis framework that can strengthen the textual-visual consistency between the visual content and textual description and the visual-visual consistency between the synthesized image and real image. This mechanism is plug and play, which can be applied to any other text-to-image task.
2) We propose a novel multi-level Vision-Language Matching model to learn the similarity between image and text, which can consider the global-level matching, fine-grained local-level matching, and general-level matching. This model is optimized by metric learning, which can push the image and text into interpretable representation space.
3) A novel evaluation metric (Vision-Language Matching Score, VLMS) is introduced in text-to-image synthesis to evaluate the performance. The VLMS considers both the visual reality of generated image and the semantic consistency between the generated image and the text description.

We evaluate the proposed dual vision-language matching strategy on two baselines, AttnGAN and DFGAN. The experiments are conducted on two widely-used datasets, Caltech-UCSD Birds 200 (CUB) and Microsoft Common Objects in Context (MSCOCO). The synthesized image quality is evaluated under the popular metrics, IS, FID, R-precision, and the proposed VLMS. The experimental results demonstrate that the proposed method achieves impressive performance improvement over previous methods. The rest of this paper is organized as follows: Section II reviews the related works of text-to-image synthesis briefly. The methodology of our proposed VLMGAN* is introduced in Section III. Then we present and analyze the experimental results in Section IV. Finally, we introduce the conclusion of this paper and the future work of our study in Section V.

II. Related Work

A. Text-to-Image synthesis

Text-to-Image synthesis aims at generating photo-realistic image that is also highly semantic consistent with the text description. For this task, many researchers present various kinds of technical solutions, such as variational inference [32], conditional PiexlCNN [33], and conditional generative adversarial networks [11]. Mansimov et al. [34] introduce a soft attention mechanism into DRAW [23] method to align the text description and the synthetic image. Originally, Reed et al. [33] propose a conditional PixelCNN [32] based approach to synthesize a photo-realistic image from the text description. Generative Adversarial Networks [5] have shown surprising performance in various generative tasks, specifically in image synthesis [6] via adversarial learning. For text-to-image synthesis, GAN also becomes the most popular research direction, such as [11], [8], [15], [9], [10], [12], [21], [22], [35], [36], [17] and so on. For instance, Reed et al. [11] firstly decompose text-to-image synthesis into two subtasks, encoding the text description to a unique representation and then synthesizing images conditioned on this vector by generative adversarial networks, called GAN-INT-CLS. An improved approach, named Generative Adversarial What-Where Network (GAWWN) [37], can focus on the location where objects should be drawn. Nguyen et al. [38] propose Plug and Play Generative Networks (PPGN) to generate images by interpreting activation maximization.

The multi-stage text-to-image synthesis framework, firstly introduced in StackGAN [21], shows remarkable superiority comparing to the one-stage strategy. This critical thought is widely accepted and applied by many kinds of research [22], [8], [10], [15], [9], which gradually improves the image resolution and quality. Specifically, StackGAN [21] adopts two-stage GANs to synthesize high-resolution images gradually: the first generator synthesizes 64x64 pixel images, and then the second generator refines it to 128x128 resolution. Based on StackGAN, a more advanced version StackGAN++ [22] is proposed, which has three generators. Besides, HDGAN [36] introduces a patch-wise adversarial loss into multi-stage generative framework.

For synthesizing image conditioned on fine-grained word-level textual features, AttnGAN [8] adopts the attention mechanism into a multi-stage generative framework. MirrorGAN [12] introduces a mirror procedure in the text-to-image task, which firstly conducts text-to-image generation and then re-describes the synthesized image. Zhu et al. [10] propose a Dynamic Memory Generative Adversarial Network (DM-GAN), which can refine the generated image by a dynamic memory block. SDGAN [20] adopts conditional batch normalization to reinforce the text highly relevant elements in the image features. DFGAN [17] fuses the text information into the hidden visual feature by a novel deep visual-textual fusion block in the image synthesis procedure. It should be noted that DFGAN adopts one-stage framework rather than multi-stage framework for text-to-image synthesis. ControlGAN [16] introduces spatial and channel-wise attentive mechanism and perceptual loss to synthesize high-quality image. LeicaGAN [39] introduces multiple prior knowledge to enforce semantic consistency. RiFeGAN [40] learns rich feature for text-to-image generation from prior knowledge.

Obj-GAN [35] can focus on synthesizing object aware images by object-driven attentive generative network. Besides, its discriminator adopt Fast R-CNN with a binary cross-entropy loss to discriminate the object information of each bounding box. Tobias et al. [41] introduce OPGAN to focus on individual objects in generating. CPGAN [42] adopts Yolo-V3 [43] to design an image content-aware discriminator in the text-to-image framework, which shows remarkable performance. Dong et al. [44] propose a text-to-image synthesis model in an unsupervised learning manner, which does not rely on the human-labeled data. Wang et al. [45] propose an end-to-end framework for text-to-image. Recently, DALLE [46] shows amazing performance on generating image from text with the help of pre-training on tons of data, which also verifies the importance of data volume. However, the images synthesized by DALLE [46] are usually cartoon style, which may be due to the pre-training dataset containing large mount of cartoon picture.

B. Image-text Matching

Cross-modal understanding is an attractive but challenging task, which includes cross-modal retrieval [47], [48], image captioning [3] and semantic grounding [49]. Specifically, image-text matching plays a key role in cross-modal retrieval. An image-text matching model aims to project the visual image and textual description into a semantic shared space using contrastive learning [50], [51], [52]. The heterogeneity gap between various type of data can be bridged by the mapped space. Specifically, visual and textual features are encoded separately into the same subspace, where the similarity values can be directly calculated. Methods can be divided into three kinds: global matching methods, regional matching methods, and multi-level matching methods.

For global matching methods, images and texts are encoded in a global way either with a CNN [53], or an LSTM [54]. The visual features and textual features are then embedded into the subspace, where their global similarity can be computed

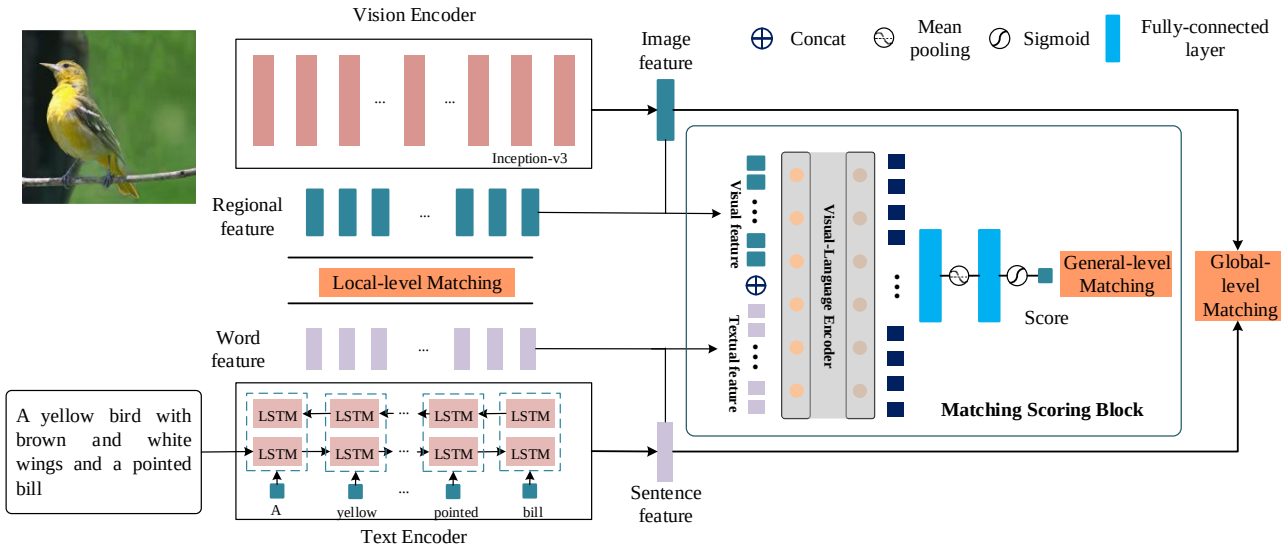


Fig. 2. Illustration of the multi-level vision-language matching model. The Visual-Language matching model contains three sub-models: Vision Encoder, Text Encoder and Matching Scoring Block.

and optimized by a triplet-ranking loss [55]. Since CNNs for image feature extracting are pre-trained on ImageNet [31], for text feature extracting, a pre-trained BERT [56] can be used for more refined features, as did in COOKIE [57].

However, these methods fail to match the concrete objects in the raw image and words in the sentence, which can be solved by regional matching methods. Thus SCAN [58] uses a pre-trained Faster RCNN [59] to detect the concrete objects and designs a stacked cross attention mechanism to align objects and words. Further, VSRN [60] adopts graph convolution [61] to learn the regional relations corresponding to the textual relations. With cross-modal pre-training and transformer-based encoders, the similarity score of image and sentence can be directly learned instead of distance calculation, as did in Uniter [62].

Considering both regional and global cross-modal matching, multi-level matching methods learn the object-word alignment and global semantic alignment simultaneously. GSLS [63] designs a multi-path structure to get both global and local similarities. CRAN [64] designs a multi-path structure for learning the global, local, and relational alignment at the same time. Wen et al. [47] utilized GAT [65] to learn dual relations of image objects and backgrounds in alignment with phrases in sentences. To calculate the similarity between the generated image and the original sentence more comprehensively, we design a multi-level matching model VLM in our method.

III. Visual-Language Matching GAN

A. Vision-Language Matching Model

The vision-language matching (VLM) model learns the multi-level similarity of text and image modality, including local-level matching, global-level matching, and general-level matching. The architecture of the proposed VLM model is shown in Figure 2. The VLM model contains three sub-models: Vision Encoder, Textual Encoder and Matching Scoring Block (MSB). The Vision Encoder and Text Encoder aim at embedding the image and text into semantic aligned vectors, which is a key process for connecting the domains of vision and language. For a fair comparison, we adopt the same backbone (Inception-v3 [66] for image and Long Short-Term Memory [54] for text) with DAMSM [8] to extract the semantic features. Inception-v3 [66] is a widely-used model for visual feature extraction. LSTM [54] can solve long distance memory problem, which is widely-used in natural language processing. The proposed MSB plays a role of generating a matching score for the image and text by a transformer encoder.

Text Encoder. Embedding text language, the simplest approach is adopting bag-of-words (BOW) model, which is widely-used in many tasks, such as cross-modal retrieval [67]. However, BOW does not consider the semantic context of the text description, which is gradually replaced by learning-based model, such as LSTM [54] and Skip-gram [68]. For the i-word of a sentence, an embedding layer embed it into a semantic vector w_i and then feed into the LSTM. Specifically, for a text, the word feature is denoted by the hidden states, and the sentence feature is represented by the last hidden state.

$$\varphi, \overline{\varphi} = F_{\text{Text-Encoder}}(w_1, w_2, ..., w_n), \tag{1}$$

where φ (matrix size:$256 \times T_0$, T_0 is the sentence length) is the word feature matrix and $\overline{\varphi}$ is the sentence feature.

Vision Encoder. The visual feature is extracted by a Convolutional Neural Network, named Inception-v3 [66]. Following previous works [8], the intermediate features of CNNs can present the local regional feature of an image, while the feature of the last layer is the global feature of an image. The

Inception-v3 [66] model is pre-trained on ImageNet [31]. The local-regional features f $(768 \times 17 \times 17)$ are denoted by the output of the $mixed\_6e$ layer and the global features $\overline{f}$ (2048×1) are represented by the $Mixed\_7b$ layer. With reshaping and linear projection, Φ (768×289) denotes the local-regional feature and $\overline{\phi}$ represents the global image feature. The projection is shown as follows,

$$\phi = F_{1\times 1conv}(f), \overline{\phi} = W\overline{f}, \tag{2}$$

where $\phi \in \mathbb{R}^{D\times 289}$ and $\overline{\phi} \in \mathbb{R}^{D}$. D is the dimension of visual and textual feature, which is equal to 256. It should be noted that only the newly added layers are trainable.

The whole process of visual feature extraction can be presented by the following formula,

$$\phi, \overline{\phi} = F_{\text{Vision-Encoder}}(x) \tag{3}$$

Matching Scoring Block. The vision-language matching scoring block aims at producing a general matching score to evaluate the matching degree between the image and the text. Transformer [26] has shown promising performance in various tasks especially in vision-language understanding, such as Bert [56] and Uniter [62]. Self-attention in Transformer [26] can deeply explore the semantic relations between visual feature and textual feature. Therefore, we adopt this mechanism to learn the matching score between image and text. Calculating the general-level matching considers both global feature and local features of the image and sentence feature and word feature of text. Specifically, the vision-language united feature is defined as

$$\psi = F_{cat}(\varphi, \overline{\varphi}, \phi, \overline{\phi}). \tag{4}$$

Where F_{cat} mean the concat operation. Then, the united feature is feed into the Transformer-based vision-language encoder, as follows.

$$\hat{\psi} = F_{Transformer}(\psi). \tag{5}$$

After obtaining the visual-textual latent features, a fully connected layer is chosen to project the features into a hidden space.

$$\hat{\psi} = W_0\hat{\psi} + b_0, \tag{6}$$

where W_0 and b_0 are the learn-able parameters of the fully connected layer. The final feature can be obtained by Mean pooling.

$$\overline{\psi} = F_{mean\_pooling}(\hat{\psi}). \tag{7}$$

Lastly, the vision-language matching score is calculated by a Sigmoid function after a Fully Connected layer, which projects the feature into a 1-dimensional value.

$$Score = F_{Sigmoid}(W_1\overline{\psi} + b_1), \tag{8}$$

where W_1 and b_1 are the learn-able parameters of the second fully connected layer. The whole process of learning the visual-language matching score can be denoted by the following formula,

$$Score = F_{\text{MSB}}(\phi, \overline{\phi}, \varphi, \overline{\varphi}). \tag{9}$$

Local-level matching. The local-level matching considers the semantic consistence between the word features and image local-regional features. Given a specific image with 289 local regional features and a text description with T_0 word features, the cosine similarity for all possible image region and word pairs are calculated by the following formula,

$$s(\phi_i, \varphi_j) = \frac{\phi_i^{\mathrm{T}}\varphi_j}{\|\phi_i\|\|\varphi_j\|}, i \in [1, 2, ..., 289], j \in [1, 2, ..., T_0]. \tag{10}$$

Here, $s(\phi_i, \varphi_j)$ is the similarity between the i-th image region and j-th word. We adopt S be the similarity matrix between word features and image local features. We adopt the popular attention mechanism [26] to learn the fine-grained similarity. The word context with respect to each image region is calculated by a weight sum of image visual feature, as following.

$$c_i = \sum_{j=0}^{288} \alpha_{ij}\phi_j, \tag{11}$$

where

$$\alpha_{ij} = \frac{exp\left(\gamma_1 s_{i,j}\right)}{\sum_{k=0}^{288} exp\left(\gamma_1 s_{i,k}\right)}. \tag{12}$$

Here, γ_1 is the in-versed temperature of the softmax function, set as 4.

Following minimum classification error formulation in speech recognition [69], the local-level matching score between the image and the text description is calculated by the LogSumExp pooling, as following,

$$S_{local}(I, T) = log\left(\sum_{i=1}^{T-1} exp\left(\gamma_2 S\left(c_i, \varphi_i\right)\right)\right)^{\frac{1}{\gamma_2}}, \tag{13}$$

where $S(c_i, \varphi_i)$ is the matching score between the i-th word and the i-th region-context, calculated by cosine similarity, γ_2 is an adjusting factor, set as 5.

$$S(c_i, \varphi_i) = \frac{c_i^{\mathrm{T}}\varphi_i}{\|c_i\|\|\varphi_i\|}. \tag{14}$$

Global-level matching. The global-level matching considers the visual global feature and the global textual feature. Similarly, for global visual feature $\overline{\phi}$ and the sentence feature $\overline{\varphi}$, the matching score is directly calculated by the cosine similarity,

$$S_{global}\left(I, T\right) = \frac{\overline{\varphi}^{\mathrm{T}}\overline{\phi}}{\|\overline{\varphi}\|\|\overline{\phi}\|}. \tag{15}$$

General-level matching. The general-level matching score is produced by the pre-trained vision-language Matching Scoring Block. As follows,

$$S_{general}(I, T) = F_{\text{MSB}}(\phi, \overline{\phi}, \varphi, \overline{\varphi}). \tag{16}$$

Objective function. Triplet loss is a popular ranking objective for matching task, which is widely-used in image-text matching [47], [60], [62]. After obtaining the matching scores of three levels, the hinge-based triplet ranking loss [55] is

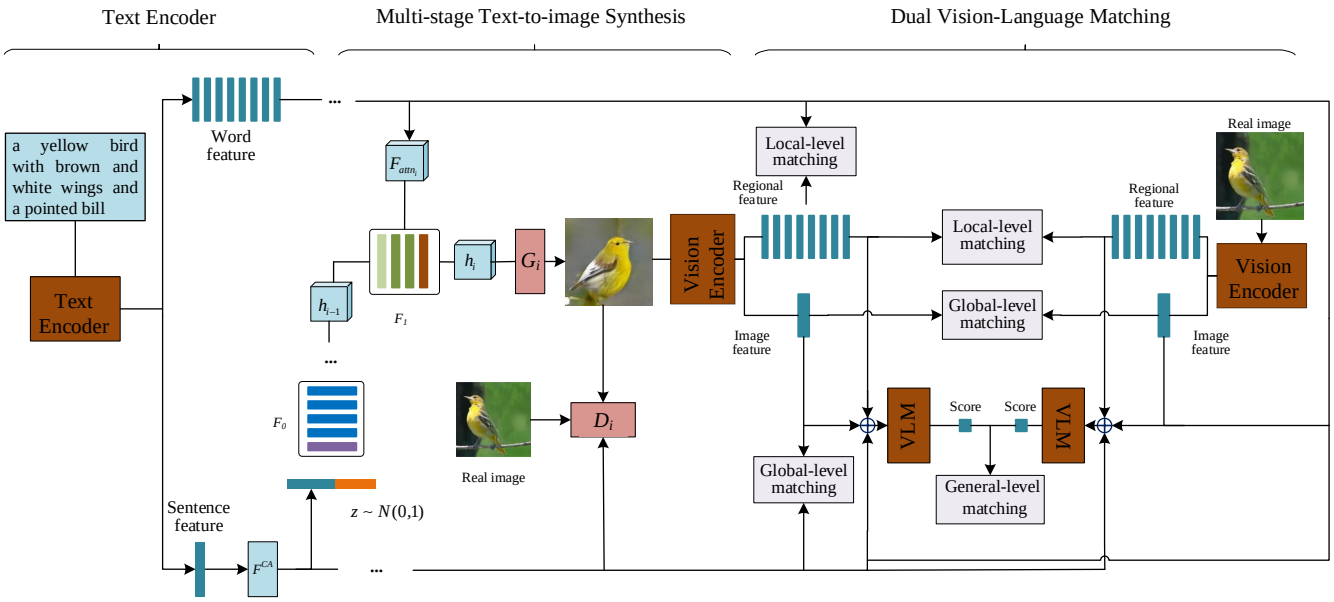


Fig. 3. The text-to-image architecture of the proposed VLMGAN$_{+\text{AttnGAN}}$. The VLMGAN$_{+\text{AttnGAN}}$ also adopts the popular attentive multi-stage strategy to improve the image quality gradually. Besides, dual vision-language matching module provides semantic consistency supervision.

adopted to optimize the vision-language matching model. The optimization loss for general-level matching is defined below.

$$\begin{aligned}\mathcal{L}_{general} =&[\alpha + S_{general}(I^{'}, T) - S_{general}(I, T)]_{+} + \\ &[\alpha + S_{general}(I, T^{'}) - S_{general}(I, T)]_{+},\end{aligned} \tag{17}$$

where $S_{general}(I^{'}, T)$ and $S_{general}(I, T^{'})$ are the general-level matching scores of un-pairing image-text instance, $S_{general}(I, T)$ are the general-level matching scores of pairing image-text instances, and α is a margin. In our experiments, α is set as 0.2. If the image and the text are closer to one another in the joint embedding space than any negatives pairs, by the margin α, the hinge loss is zero. If we substitute the general-level matching score by local-level and global-level matching score, we can obtain the loss of $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$.

Finally, the overall loss function of VLM model is defined as

$$\mathcal{L}_{VLM} = \mathcal{L}_{local} + \mathcal{L}_{global} + \mathcal{L}_{general}. \tag{18}$$

Optimization. The Optimization of VLM contains two types. The first type is training for supervising text-to-image synthesis, which is optimizing on the training dataset. The second type is optimizing the Matching Scoring Block for obtaining the VLMS to evaluate the performance. This type is training on the whole dataset including the testing dataset. The optimizer is Adam and the learning rate is 0.0002. The training is stopped after 200 epochs.

B. Dual Matching-driven Attentive GAN

For a fair comparison and better understanding, we adopt AttnGAN [8] as baseline, which is also chosen by many state-of-the-art methods due to its excellent performance, to implement the dual vision-language matching strategy. As a example, the VLMGAN$_{+\text{AttnGAN}}$ The multi-stage text-to-image synthesis architecture is stacking three generative-adversarial blocks sequentially, as shown in the left part of Figure 3. Given a specific text description, the sentence feature $\overline{\varphi}$ and word features φ are extracted by the text encoder in VLM model. The synthesized image can be obtained by the following procedures.

$$\begin{aligned} h_0 &= F_0(F_{cat}(z, F^{ca}(\overline{\varphi}))), \\ h_i &= F_i(h_{i-1}, F_{attn_i}(h_{i-1}, \varphi)), i \in \{1, 2.\} \\ \hat{x}_i &= G_i(h_i), i \in \{0, 1, 2.\} \end{aligned} \tag{19}$$

where $z \sim N(0, 1)$ is a random noise vector, F^{ca} is conditioning augmentation (CA) [21] process, and F_{attn_i} is attentive mechanism described in AttnGAN [8]. Due to the excellent performance of attentive mechanism, We adopt the word-level attentive mechanism F_{attn_i} to fuse the word features into visual features. The F_{attn_i} has two inputs, the word features φ and previous hidden features h. Mathematically, the word-visual fusion context can be defined as follows,

$$h_i = F_{attn_i}(\varphi, h_{i-1}) = (c_0, c_1, \cdots, c_N) \in \mathbb{R}^{\hat{D} \times N}, \tag{20}$$

where

$$c_j = \sum_{i=0}^{T_0 - 1} \alpha_{j,i} \varphi_i. \tag{21}$$

Here, $\alpha_{j,i}$ is the attention weight between the j-th image region and the i-th word feature, as follows.

$$\alpha_{j,i} = \frac{exp\,(s^{'}{}_{j,i})}{\sum_{k=0}^{T-1} exp\,(s^{'}{}_{j,k})}. \tag{22}$$

where $s'_{j,i} = {h_j}^{\mathrm{T}}\varphi_i$ and $s'_{j,i}$ indicates the similarity between the j-th feature in hidden feature h and the i-th word feature.

Objective function. Each stage contains a generator and discriminator, which are optimized alternately by a generative loss and a discriminating loss. Specifically, the generative loss is

$$\mathcal{L}_{G_i} = -\frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}[log(D_i(\hat{x}_i))] - \frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}[log(D_i(\hat{x}_i, \overline{\varphi}))], \tag{23}$$

where x_i is the synthesized image. In this objective function, the former is the unconditional loss, which determines whether the synthesized image is real or fake. The latter is the conditional loss, which determines whether the image contents are matched with the text description or not. The generative loss function forces the model synthesize photo-realistic and text semantic consistent images.

At the same time, the discriminator is designed to distinguish the generated image is both fake and semantic consistent or not. Therefore, the discriminating loss also consists of unconditional visual realism loss and conditional semantic consistent loss. Mathematically, it is defined as follows.

$$\begin{aligned}\mathcal{L}_{D_i} = &-\frac{1}{2}E_{x_i\sim P_{data}}[log(D_i(x_i))] - \frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}[log(1 - D_i(\hat{x}_i))] + \\ &-\frac{1}{2}E_{x_i\sim P_{data}}[log(D_i(x_i, \overline{\varphi}))] - \frac{1}{2}E_{\hat{x}_i\sim P_{G_i}}[log(1 - D_i(\hat{x}_i, \overline{\varphi}))],\end{aligned} \tag{24}$$

where P_{data} is the real data distribution and P_{G_i} is the generated image data distribution. The first two expressions are unconditioned losses, which focus on distinguishing the synthesized image is real or fake. The later two expressions are conditioned losses, which focus on distinguishing the synthesized image is consistent with the text description or not. However, only adopting the discriminator to force the synthesized image consistence is insufficient. As shown in Equation 24, the discriminator only considers the global text feature instead of both global feature and local feature.

To obtain more semantic consistent image, we introduce an additional supervision part, dual multi-level vision-language matching module. The dual multi-level vision-language matching module contains two parts, textual-visual matching and visual-visual matching. We introduce the loss function $\mathcal{L}_{VLM}$ of VLM model to strengthen the textual-visual consistency between the generated image and the corresponding text. Besides, the visual-visual consistency between the synthesized image and the real image should be considered in the objective function.

Visual-Visual Matching (VVM) also contains three parts, local-level matching, global-level matching, and general-level matching. We adopt the vision encoder to extract the local-level features ϕ_{fake} and global-level feature $\overline{\phi}_{fake}$ of the synthesized image $\hat{x}$, as following.

$$\phi_{fake}, \overline{\phi}_{fake} = F_{\text{Vision-Encoder}}(\hat{x}) \tag{25}$$

Analogously, the local-level features ϕ_{real} and global-level feature $\overline{\phi}_{real}$ of the real image x

$$\overline{\phi}_{real}, \overline{\phi}_{real} = F_{\text{Vision-Encoder}}(x) \tag{26}$$

The visual-visual global-level matching loss aims at maximizing the global matching score between the real image and the synthesized fake image. The loss function is

$$\mathcal{L}_{VG}(\phi_{real}, \overline{\phi}_{fake}) = -\frac{1}{B}\sum_{i=1}^{B} log\frac{e^{(S(\overline{\phi}_{real}, \overline{\phi}^{+}_{fake})/\tau_0)}}{e^{(S(\overline{\phi}_{real}, \overline{\phi}^{+}_{fake})/\tau_0)} + \sum_{j=1}^{B-1} e^{(S(\overline{\phi}_{real}, \overline{\phi}^{-}_{fake})/\tau_0)}}, \tag{27}$$

where B is the batch size and τ_0 is a hyperparameter (set as 0.07). $S(\overline{\phi}_{real}, \overline{\phi}^{+}_{fake})$ is the cosine similarity between the paired real image and synthesized image. $S(\overline{\phi}_{real}, \overline{\phi}^{-}_{fake})$ is the cosine similarity between the unpaired real image and synthesized image. The loss function of Vision-Vision global matching is the sum of fake-to-real and real-to-fake, as following.

$$\mathcal{L}_{VG} = \mathcal{L}_{VG}(\overline{\phi}_{real}, \overline{\phi}_{fake}) + \mathcal{L}_{VG}(\overline{\phi}_{fake}, \overline{\phi}_{real}), \tag{28}$$

The visual-visual local-level matching loss maximizes the similarity between the regional features of real image and synthesized image.

$$\mathcal{L}_{VL}(\phi_{real}, \phi_{fake}) = -\frac{1}{B}\sum_{i=1}^{B} log\frac{e^{(S(\phi_{real}, \phi^{+}_{fake})/\tau_0)}}{e^{(S(\phi_{real}, \phi^{+}_{fake})/\tau_0)} + \sum_{j=1}^{B-1} e^{(S(\phi_{real}, \phi^{-}_{fake})/\tau_0)}}, \tag{29}$$

where $S(\phi_{real}, \phi_{fake})$ is calculated by the following formula.

$$S(\phi_{real}, \phi_{fake}) = log\left(\sum_{i=1}^{N-1} exp\left(\gamma_3 S\left(\varphi_i^{real}, \varphi_i^{fake}\right)\right)\right)^{\frac{1}{\gamma_3}}, \tag{30}$$

Here $S(\cdot)$ is cosine similarity calculation formula and γ_3 is set as 5. The loss function of visual-visual local-level matching is the sum of two parts, as following.

$$\mathcal{L}_{VL} = \mathcal{L}_{VL}(\phi_{real}, \phi_{fake}) + \mathcal{L}_{VL}(\phi_{fake}, \phi_{real}), \tag{31}$$

The visual-visual general-level matching loss is calculated by the pre-trained MSB model, as following.

$$\mathcal{L}_{VGEN} = \| F_{\text{MSB}}(\phi_{real}, \overline{\phi}_{real}, \varphi, \overline{\varphi}) - F_{\text{MSB}}(\phi_{fake}, \overline{\phi}_{fake}, \varphi, \overline{\varphi}) \|_2^2 . \tag{32}$$

The overall loss function of Visual-Visual Matching $\mathcal{L}_{VVM}$ is the sum of local-level matching $\mathcal{L}_{VL}$, global-level matching $\mathcal{L}_{VG}$, and general-level matching $\mathcal{L}_{VGEN}$, as following.

$$\mathcal{L}_{VVM} = \mathcal{L}_{VG} + \mathcal{L}_{VL} + \mathcal{L}_{VGEN}. \tag{33}$$

Textual-Visual Matching (TVM) is presented in Section Vision-Language Matching Model. In text-to-image model, we feed the synthesized image $\hat{x}$ in the VLM model to calculate the textual-visual matching loss $\mathcal{L}_{VLM}$.

The adversarial generative loss, the textual-visual matching loss and the visual-visual loss are combined in the general loss function. Therefore, the objective function of the overall generative model is defined as

$$\mathcal{L}_G = \sum_{i=0}^{2}\mathcal{L}_{G_i} + \lambda_1\mathcal{L}_{VVM} + \lambda_2\mathcal{L}_{VLM}. \tag{34}$$

where λ_1 and λ_2 are two balancing factors.They are set as 5 for VLMGAN$_{+\text{AttnGAN}}$ and set as as 0.25 for VLMGAN$_{+\text{DFGAN}}$. In the training process, the three generators are optimized simultaneously and the three discriminators are optimized independently. The parameters of vision encoder, text encoder, and matching scoring block are not trainable.

IV. Experiments

In this section, extensive experiments are conducted to verify the effectiveness of the proposed method. We firstly introduce the experimental settings and then show the quantitative and qualitative evaluation results. Lastly, we present ablation studies and further discussions.

A. Datasets and Evaluation Metrics

Datasets. Two widely-used benchmarks, CUB [28], and MSCOCO [70], are adopted to demonstrate the capability of the proposed method. The CUB contains 11,788 images belonging to 200 categories, which is divided into two sub-datasets, 8,855 images for training and the remaining 2,933 for testing. For each image, there are ten textual descriptions. The MSCOCO dataset is a larger and more challenging benchmark. It contains 120k images, and each image is described by five texts. We split them into a training set with about 80k images and a testing set with 40k images. The dataset settings are the same as previous works.

Evaluation Metric. We quantify the effectiveness of the proposed method in terms of Fréchet Inception Distance (FID) [30], Inception Score (IS) [29], R-precision [8], and the proposed VLMS.

The IS calculates the Kullback-Leibler (KL) divergence between the class distribution of original image and the class distribution of generated image. The class distribution is calculated by the pre-trained Inception-v3 model. The higher IS suggests that the synthesized images are more realistic and more confident to a specific class. The Inception Score is calculated by the following formula:

$$\text{IS} = exp(E_{X\sim P_G}[D_{KL}(P_{Y|X}(y|x))||P_Y(y)]), \tag{35}$$

where x is the generated fake image, and y is its corresponding semantic label predicted by the pre-trained Inception-v3 model. The distribution $p\,(y\,|x)$ denotes the probability distribution of image x belongs to a specific category y , and $p\,(y)$ is the probability distribution of predicted class.

The FID measures the Fréchet Distance between global semantic feature of synthesized image and real image, which are extracted by the pre-trained Inception-v3 model. Thus, lower value of FID ndicates that the synthesized images are close to the original images. Lower value is better, vice versa. The calculation of FID is as follows.

$$\text{FID} = \|m - m_r\|_2^2 + Tr\left(C + C_r - 2\left(CC_r\right)^{\frac{1}{2}}\right), \tag{36}$$

where (m, C) are mean and variance of the generated data, and (m_r, C_r) are mean and variance of the real data.

The R-precision is to evaluate whether the synthesized image is consistent with the corresponding description by retrieving the text description for a given image. For a fair comparison, we follow the evaluation settings with DMGAN [10] and quote their results. The VLMS is calculated by a pre-trained VLM model, considering both the image quality and the visual-textual semantic consistency.

B. Implementations

The proposed dual vision-language matching strategy on two baselines is implemented with 7700k CPU and 8 NVIDIA GeForce GTX2080ti GPUs. For the vision encoder, text encoder and visual-language matching scoring model, the batch size is set to 64. For the VLMGAN$_{+\text{AttnGAN}}$ model, the batch size is set to 24, and the learning rate is 0.0001 for the generators and 0.0004 for the discriminators. We only apply the dual multi-level vision-language supervision in the last generator (256x256) due to the low-resolution images are not well synthesized. The ADAM optimizer [71] is adopted to optimize the proposed model. The training of VLMGAN$_{+\text{AttnGAN}}$ is stopped at 600 epochs for CUB bird dataset and at 120 epochs for MSCOCO dataset following previous works [8], [10]. The parameters of generative networks and the discriminating networks are optimized alternatively. For VLMGAN$_{+\text{AttnGAN}}$, the training process is shown in Algorithm 1. For another baseline (DFGAN), the settings are similar with those of VLMGAN$_{+\text{AttnGAN}}$.

Algorithm 1 Training procedure of VLMGAN$_{+\text{AttnGAN}}$

Require: Pre-trained models (Vision-Encoder, Text-Encoder, and MSB); Batch size M; Text-image paired instances $\{T, I\}$; Learning rate α.
Ensure: Generators (G_0, G_1, G_2) and discriminators (D_0, D_1, D_2);

repeat
  Sample image-text pairs $\{I_i, T_i\}$ and generate random noise vector z_i;
  Extract representation of text captions by $\varphi, \overline{\varphi} = F_{\text{Text-Encoder}}(w_1, w_2, ..., w_n)$;
  Generate fake images by $(\hat{x}_0, \hat{x}_1, \hat{x}_2) \leftarrow G(\overline{\varphi}, \varphi, z_i)$
  for $i \in [0, 1, 2]$ **do**
    Calculate the discriminative loss $\mathcal{L}_{D_i}$:
    Update parameters of discriminator D_i by Adam optimizer;
  end for
  for $i \in [0, 1, 2]$ **do**
    Calculate the generative loss $\mathcal{L}_{G_i}$;
    if i is equal to 2 **then**
      Calculate textual-visual matching loss $\mathcal{L}_{VLM}$;
      Calculate visual-visual matching loss $\mathcal{L}_{VVM}$;
    end if
  end for
  Calculate total loss $\mathcal{L}_G$:
    $\mathcal{L}_G \leftarrow \mathcal{L}_{G_0} + \mathcal{L}_{G_1} + \mathcal{L}_{G_2} + \mathcal{L}_{VVM} + \mathcal{L}_{VLM}$;
  Update the parameters of the generators (G_0, G_1, G_2) by Adam optimizer;
until VLMGAN$_{+\text{AttnGAN}}$ converges
return G_0, G_1, G_2, D_0, D_1, D_2.

C. Quantitative Evaluation

The proposed dual vision-language matching module can be applied to other text-to-image synthesis architectures. In our experiments, we apply the dual vision-language matching

TABLE I
THE INCEPTION SCORE COMPARISON OF THE PROPOSED VLMGAN* AND THE STATE-OF-THE-ART METHODS ON THE CUB BIRD DATASET AND MSCOCO DATASET. DFGAN* MEANS THE SCORES ARE OBTAINED BY USING THEIR PRE-TRAINED MODEL.

Model	Resolution	CUB	MSCOCO
GAN-INT-CLS [11]	64x64	2.88 ± 0.04	7.88 ± 0.07
GAWWN [37]	256x256	3.62 ± 0.07	-
StackGAN [21]	256x256	3.70 ± 0.04	8.45 ± 0.03
StackGAN++ [22]	256x256	3.82 ± 0.06	-
HDGAN [36]	512x512	4.15 ± 0.05	-
MirrorGAN [12]	256x256	4.56 ± 0.04	26.47 ± 0.41
LeicaGAN [39]	256x256	4.62 ± 0.06	-
DMGAN [10]	256x256	4.75 ± 0.07	30.49 ± 0.57
Bridge-GAN [15]	256x256	4.74 ± 0.04	16.40 ± 0.30
OPGAN [41]	256x256	-	28.57 ± 0.17
C4Synth [72]	256x256	4.07 ± 0.13	-
CGL-GAN [73]	256x256	3.67 ± 0.04	13.62 ± 0.02
KTGAN [19]	256x256	4.85±0.04	31.67 ± 0.36
LD-CGAN [74]	128x128	4.18± 0.06	-
SAMGAN [27]	256x256	4.61 ± 0.03	27.31 ± 0.23
CPGAN* [42]	256x256	-	52.73 ± 0.61
MA-GAN [25]	256x256	4.76 ± 0.05	-
AttnGAN [8]	256x256	4.36 ± 0.04	25.89 ± 0.47
DFGAN [17]	256x256	4.86 ± 0.04	-
DFGAN* [17]	256x256	4.70 ± 0.05	18.70 ± 0.07
VLMGAN$_{+\text{AttnGAN}}$	256x256	4.86 ± 0.06	31.84 ± 0.46
VLMGAN$_{+\text{DFGAN}}$	256x256	4.95 ± 0.04	26.51 ± 0.43

TABLE II
THE FID AND R-PRECISION COMPARISON OF ATTNGAN, DMGAN, DFGAN, VLMGAN$_{+\text{ATTNGAN}}$ AND VLMGAN$_{+\text{DFGAN}}$ ON CUB DATASET AND MSCOCO DATASET. DFGAN* MEANS THE SCORES ARE OBTAINED BY USING THEIR RELEASED PRE-TRAINED MODEL. THE '↓' MEANS THE LOWER, THE BETTER. THE '↑' MEANS THE HIGHER, THE BETTER.

Methods	**CUB**		**MSCOCO**	
	FID ↓	R-precision ↑	FID ↓	R-precision ↑
AttnGAN [8]	23.98	67.82±4.43	35.49	85.47±3.69
DMGAN [10]	16.09	72.31±0.91	32.64	88.56±0.28
DFGAN [17]	19.24	-	28.92	-
DFGAN* [17]	21.85	38.76±0.08	27.39	55.34±0.90
VLMGAN$_{+\text{AttnGAN}}$	15.02	77.75±0.74	31.24	89.45±0.52
VLMGAN$_{+\text{DFGAN}}$	16.04	72.59±0.32	23.62	82.95±0.60

module on two popular baselines, AttnGAN and DFGAN. They are marked as VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{DFGAN}}$ respectively.

The experimental results of Inception Score on CUB and MSCOCO are reported in Table I. Table I shows that the proposed VLMGAN$_{+\text{AttnGAN}}$ achieves 4.86 on the CUB bird dataset, which outperforms the other methods except for DFGAN and VLMGAN$_{+\text{DFGAN}}$. Compared with baseline (AttnGAN) [8], the proposed method VLMGAN$_{+\text{AttnGAN}}$ can improve the Inception Score from 4.36 to 4.86 and 25.89 to 31.84. This suggests that the VLMGAN$_{+\text{AttnGAN}}$ can generate images with better diversity and image quality. It should be noted that the IS value of the proposed VLMGAN$_{+\text{AttnGAN}}$ is lower than CPGAN [42], as a result of CPGAN adopts pre-trained Yolo-v3 [43] as the discriminator. With extra information, CPGAN obtains the highest IS score on MSCOCO. The proposed VLMGAN$_{+\text{AttnGAN}}$ can obtain same results with DFGAN on CUB bird daraset. In addition to AttnGAN, we also choose DFGAN as baseline to implement the dual vision-language matching strategy named VLMGAN$_{+\text{DFGAN}}$. In original paper of DFGAN, the authors do not report the IS and R-precision on MSCOCO dataset. Therefore, we obtain the values by using their public available pre-trained model, which is marked with DFGAN*. The results in Table I indicate that VLMGAN$_{+\text{DFGAN}}$ obtain the best IS score (4.95). However, the IS score on MSCOCO dataset is lower than other methods. This phenomenon is also appeared in Bridge-GAN [15]. The cause may be DFGAN only adopts the sentence feature as condition and the diversity of synthesized images decreases without word features. Compared with its baseline (DFGAN), VLMGAN$_{+\text{DFGAN}}$ can improve the IS from 18.70 to 26.51. In summary, the proposed vision-language matching strategy is beneficial to the performance of two baselines.

The FID and R-precision of AttnGAN, DMGAN, DFGAN, VLMGAN$_{+\text{AttnGAN}}$, and VLMGAN$_{+\text{DFGAN}}$ on CUB dataset and MSCOCO dataset are reported in Table II. DMGAN, DFGAN, and VLMGAN* are also improved from AttnGAN. Compared with AttnGAN, the proposed VLMGAN$_{+\text{AttnGAN}}$ outperforms it by a large margin on both two datasets. Specifically, VLMGAN$_{+\text{AttnGAN}}$ improves R-precision from 67.82 to 77.75 for CUB and from 85.47 to 89.45 for MSCOCO. R-precision evaluates whether the synthesized image is consistent with the text description by retrieving manner. The image feature for retrieval is obtained by the pre-trained Vision-Encoder. The results show that the proposed dual vision-language matching strategy makes a pivotal contribution to improve visual-textual semantic consistency. The comparison of the FID score indicates that the image distribution generated by VLMGAN$_{+\text{AttnGAN}}$ is closer to real image distribution than others on CUB dataset. Comparing to advanced DMGAN and DFGAN, the proposed VLMGAN$_{+\text{AttnGAN}}$ also keeps its superiority,which can obtain competitive results. For another baseline (DFGAN), we can find that VLMGAN$_{+\text{DFGAN}}$ obtains the relatively lower FID on CUB dataset and the lowest FID on MSCOCO dataset. The FID scores of DFGAN shows that DFGAN have excellent performance in synthesizing photo-realistic image. However its semantic consistency is relatively poor in term of R-precision. This phenomenon also verifies that some methods can not be good at both image reality and text semantic consistency. By strengthening the vision-language matching, VLMGAN$_{+\text{DFGAN}}$ can improve the performance on both FID and R-precision.

D. Qualitative Evaluation

Visual comparisons of AttnGAN, DMGAN, DFGAN, VLMGAN$_{+\text{AttnGAN}}$, and VLMGAN$_{+\text{DFGAN}}$ are shown in Figure 4. In this paper, we use 'VLMGAN*' to denote the dual-matching driven methods VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{DFGAN}}$. In general, the images synthesized by VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{DFGAN}}$ are more realistic and highly consistent with the text description because it employs a vision-language matching model and a visual consistency constraint. On the CUB dataset, the proposed

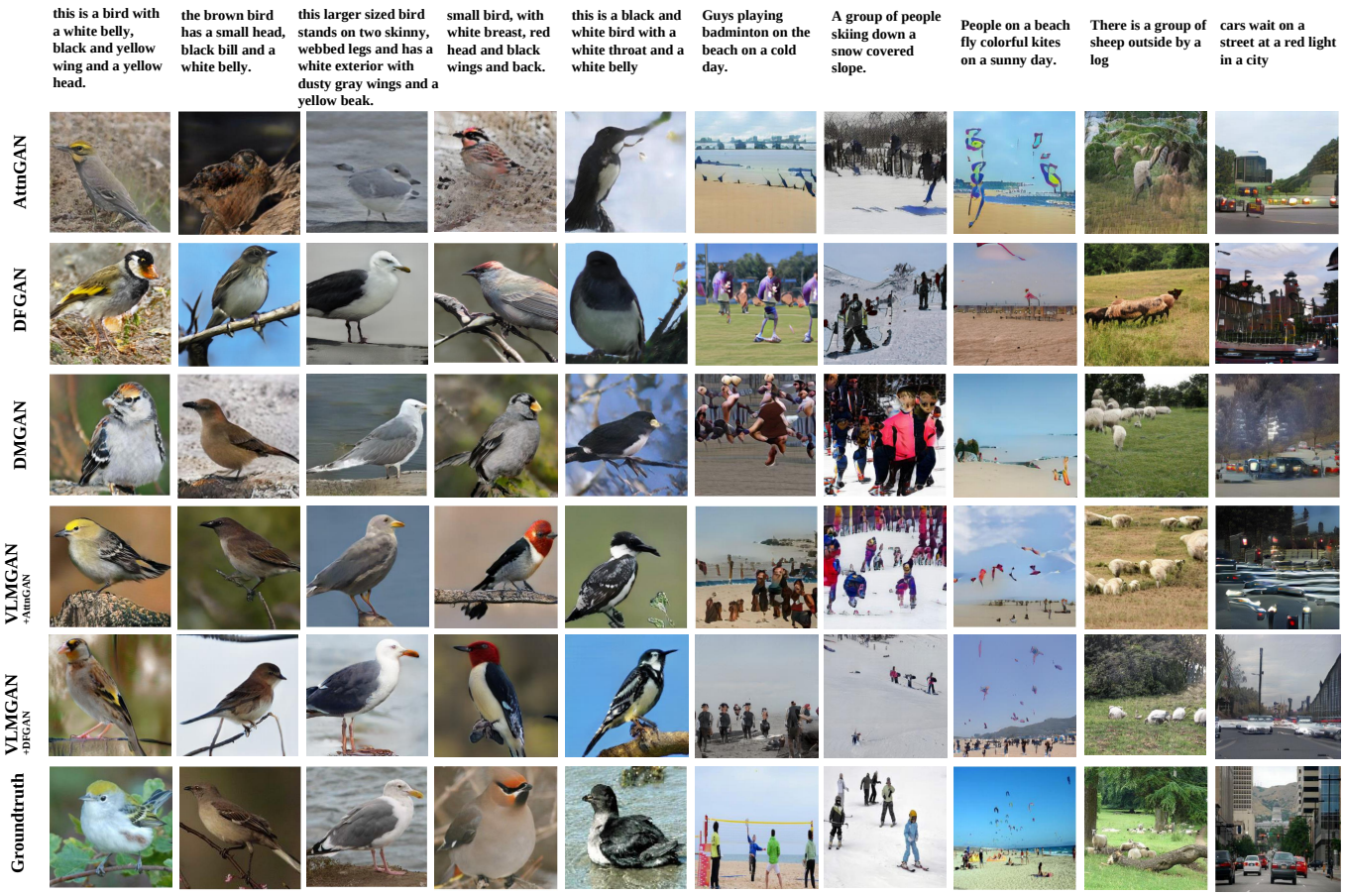


Fig. 4. Examples synthesized by AttnGAN [8], DFGAN [17], DMGAN [10], VLMGAN$_{+\text{AttnGAN}}$, and VLMGAN$_{+\text{DFGAN}}$. The images in the last row are the corresponding ground truth. The images in the same column are conditioned the same description.

TABLE III
THE PERFORMANCE OF DIFFERENT COMPONENTS OF THE PROPOSED VLMGAN ON CUB DATASET.'W/O' MEANS 'WITHOUT'.

Architectures	FID ↓	IS ↑	R-precision ↑
AttnGAN, w/o DAMSM	53.73	3.89 ± 0.04	10.37 ± 5.88
AttnGAN (baseline)	23.89	4.36 ± 0.03	67.82 ± 4.43
VLMGAN$_{+\text{AttnGAN}}$, w/o VLM	33.25	4.20 ± 0.05	46.45 ± 3.36
VLMGAN$_{+\text{AttnGAN}}$, w/o VVM	16.23	4.73 ± 0.07	73.56 ± 0.82
VLMGAN$_{+\text{AttnGAN}}$	15.02	4.86 ± 0.06	77.75 ± 0.74

TABLE IV
ABLATION STUDIES ON DIFFERENT LEVEL MATCHING ON CUB BIRD DATASET. 'W/O' MEANS 'WITHOUT'.

Model Setting	FID ↓	IS ↑	R-precision ↑
Baseline	53.73	3.89 ± 0.04	10.37 ± 5.88
VLMGAN$_{+\text{AttnGAN}}$, w/o General	17.00	4.76 ± 0.07	73.41 ± 0.60
VLMGAN$_{+\text{AttnGAN}}$, w/o Local	23.32	4.47 ± 0.07	57.40 ± 0.88
VLMGAN$_{+\text{AttnGAN}}$, w/o Global	18.75	4.65 ± 0.06	67.76 ± 0.56
VLMGAN$_{+\text{AttnGAN}}$	15.20	4.86 ± 0.06	77.75 ± 0.74

model can better understand the description and synthesize a more clearly structured image. Comparing with MSCOCO, the CUB is more straightforward, so that all of these methods have relatively better performance. In terms of complex image synthesis, the MSCOCO dataset is adopted to verify the proposed method's performance. The models with vision-language matching strategy can precisely understand the text description and generates a well-structured image. For example, VLMGAN* well presents the shape and structure of kites like the ground-truth, while other methods can not. The visual comparison shows that the proposed VLMGAN* has superiority in keeping semantic and visual consistency by using a multi-level vision-language matching model and a visual-consistent constraint.

E. Ablation Study

To thoroughly verify VLMGAN*'s effectiveness, we do ablation studies on VLMGAN* and its variants. VLMGAN* means the dual vision-language matching strategy based GAN for text-to-image synthesis. We conduct this ablation study of on VLMGAN$_{+\text{AttnGAN}}$ on CUB dataset. Several comparative experimental results are reported in Table III. "VLM" means the proposed vision-language matching constraint, and "VVM" means the visual-visual matching consistent constraint between the synthesized image and original real image. The AttnGAN is our baseline. The comparing results show that the vision-language matching model is fundamental in improving the image quality. Without the VLM model, the performance

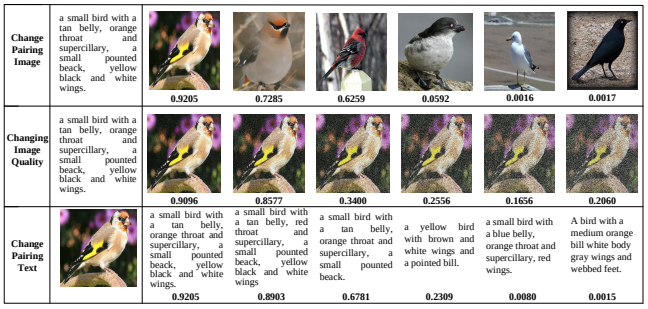


Fig. 5. Some examples of VLMS by modifying the image and text description. The numbers under the images (texts) are the corresponding VLMS value between the left text (image).

TABLE V
THE VLMS VARIATION BY CHANGING THE PAIRING IMAGE QUALITY ON CUB DATASET.

Settings	VLMS ↑
Ground truth	0.77±0.30
Random image	0.12±0.22
$\sigma(0.01)$	0.63±0.29
$\sigma(0.1)$	0.40±0.30
$\sigma(0.3)$	0.23±0.23
$\sigma(0.5)$	0.21±0.21
$\sigma(1)$	0.18±0.20

TABLE VI
THE VLMS VARIATION BY CHANGING THE PAIRING TEXT DESCRIPTION ON CUB DATASET.

Settings	VLMS ↑
Ground truth	0.77±0.30
Random text	0.12±0.22
mask stopwords	0.75±0.24
10%	0.67±0.26
20%	0.56±0.27
50%	0.37±0.26
70%	0.35±0.25
90%	0.10±0.22

TABLE VII
THE VLMS COMPARISON WITH DIFFERENT MODELS ON CUB DATASET.

Methods	CUB	MSCOCO
AttnGAN	0.49±0.20	0.36±0.28
DMGAN	0.52±0.22	0.42±0.30
DFGAN	0.53±0.27	0.44±0.23
VLMGAN$_{+\text{AttnGAN}}$	0.55±0.27	0.47±0.26
VLMGAN$_{+\text{DFGAN}}$	0.56±0.23	0.46±0.23

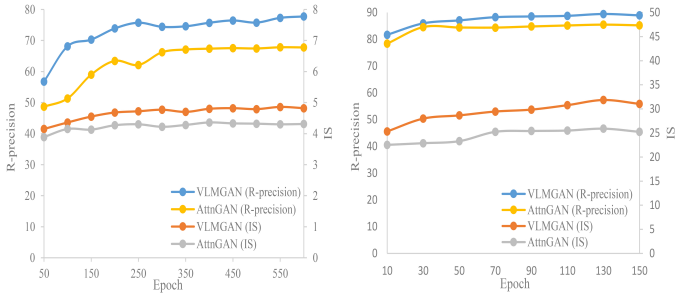
of all evaluation metrics drop rapidly , such as the IS score drops from 4.86 to 4.20, the FID increases from 15.02 to 33.25, the R-precison drops from 77.75 to 46.45. Comparing "VLMGAN$_{+\text{AttnGAN}}$, w/o VVM" to "AttnGAN", the results indicate that the proposed multi-level vision-language model, which can effectively match the image content and the textual semantic information, is better than DAMSM. Besides, the visual consistency constraint between the synthesized and the real image also can improve the scores by a considerable margin. These experimental results show that the components of the proposed method contribute to improving the image quality.

To clarify the contribution of different level matching, we conduct more experiments on CUB dataset. We also adopt the AttnGAN as the baseline. The experimental results are shown in Table IV. We can find that the local-level matching makes the biggest contribution to the performance, which improves the IS from 4.47 to 4.86, FID from 23.32 to 15.20. The three kinds of matching can improve the model performance in terms of different metrics.

The above ablation studies can verify the effectiveness of the proposed dual multi-level vision-language matching strategy.

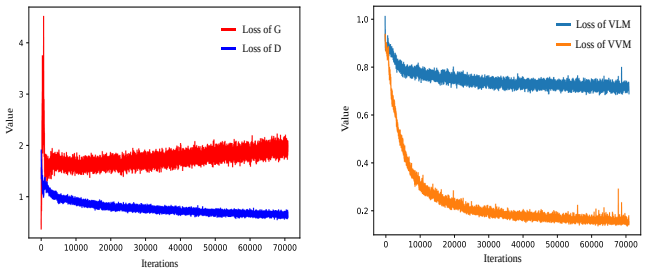
F. Effectiveness of VLMS

In this subsection, we firstly explain the effectiveness and rationality of the proposed novel text-to-image evaluation metric, named Vision-Language Matching Score (VLMS), which directly measures the similarity between the synthesized image and the corresponding text description by the pre-trained Matching Scoring Block of the VLM model. To verify the effectiveness of the proposed VLMS, we conduct experiments with different variants, as shown in Table V and Table VI. In Table V, 'Random text' means that we randomly select a text from the dataset for a specific image, which obtains the lowest VLMS. This result indicates that VLMS is sensitive to the image-text semantic consistency. For changing the image quality, we add different level white gaussian noise (standard deviation: 0.01, 0.1, 0.3, 0.5, 1). We can find that the VLMS decreases rapidly with the increase of noise. This phenomenon indicates that the VLMS is sensitive to the image quality. For changing the text description, we randomly replace or remove words of the whole sentence by different percentage (10%, 30%, 50%, 70%, 90%). For a specific text description, the ground truth obtains the best VLMS score, and the dissimilar images receive relatively low scores. Besides, we add experiments to analyze the influence of these irrelevant words, such as 'the', 'a' and so on. To be specific, we build a stop words dictionary, which contains 'and, this, a, an, there, of'. If the words of the sentence are in the stop words, we mask them in calculating the VLMS. We find that these irrelevant words make little impact on the final results (from 0.77 to 0.75). The comparisons show that the lower the image quality, the lower the score. If we modify the text description, the VLMS scores also decrease. For better understanding, we present some examples to explain this, as shown in Figure 5. Therefore, from above analyses, the proposed VLMS metric is reasonable to measure the image-text matching score, which considers both image quality and semantic consistency. We calculate the mean VLMS score of 30000 generated images. Table VII shows that VLMGAN$_{+\text{DFGAN}}$ obtains the highest value on CUB datasets and VLMGAN$_{+\text{AttnGAN}}$ obtains the highest value on MSCOCO datasets.



(a) CUB (b) MSCOCO

Fig. 6. IS and R-precision comparison between AttnGAN [8] and VLMGAN$_{+\text{AttnGAN}}$ with the training processing.



(a) Loss of Generator and Discriminator (b) Loss of VVM and VLM

Fig. 7. The training loss of VLMGAN$_{+\text{DFGAN}}$.

G. Convergence Analysis

Figure 6 shows the comparison between VLMGAN$_{+\text{AttnGAN}}$ and AttnGAN with the training processing. This figure shows that VLMGAN$_{+\text{AttnGAN}}$ exceeds AttnGAN in the whole training process. Besides, the training losses of VLMGAN$_{+\text{DFGAN}}$ are shown in Figure 7. It should be noted that loss of generator does not include the loss of $\mathcal{L}_{VVM}$ and $\mathcal{L}_{VLM}$. By comparing $\mathcal{L}_{VVM}$ and $\mathcal{L}_{VLM}$, we can find that the decrease of $\mathcal{L}_{VVM}$ is more significant. The model is converged after about 70000 iterations.

H. Generalization Study

In this section, we conduct more experiments to analyze the generalization and robustness of the proposed approach. The first experiment is modifying some attribute words when generate the corresponding image, as shown in the top row of Figure 8. The words in red are some key attributes when describing the bird. We randomly replace these words with other attributes words. The results of modifying some important words show that the synthesized images can keep consistency with the descriptions' variation. The second experiment is generating a series of images by fixing the text description. The results of some examples conditioned the same text of CUB dataset and MSCOCO dataset are presented in the second and three rows of Figure 8, respectively. From the second row, we can find that the attributes of the synthesized birds (black bill, white breast and red feathers) keep consistency

Fig. 8. More synthesized examples of VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{DFGAN}}$.

among different images. In the same time, we can observe that these birds have rich variety of the posts, such as the direction of body. This experimental phenomenon is also appeared on the MSCOCO dataset. Therefore, we can conclude that the proposed method VLMGAN* have excellent generalization and robustness.

V. Conclusion

This paper addresses the text-to-image synthesis by strengthening the semantic and visual matching between the synthesized image and the real data. To this end, the proposed dual multi-level vision-language matching considers both textual-visual matching and visual-visual matching. By introducing this idea into generative architecture, the VLMGAN* successfully exploits this idea and achieves excellent image quality performance. In addition, the VLM can also measure the matching score between the image and text by considering both image quality and image semantic, which is more consistent with our human perception. We implement the proposed dual vision-language matching strategy on two popular baselines, AttnGAN and DFGAN. The experimental results of VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{DFGAN}}$ show that the VLMGAN* achieves state-of-the-art performance on CUB dataset and more challenging MSCOCO dataset. Compare with the baselines, both VLMGAN$_{+\text{AttnGAN}}$ and VLMGAN$_{+\text{AttnGAN}}$ can significantly improve their performance. In the future study, we will try to explore more excellent mechanisms to improve the quality of synthesized image and semantic consistency between the synthesized image and text.

REFERENCES

[1] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected gans converge faster," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[2] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2479–2486.

[3] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[4] C. Deng, Z. Li, X. Gao, and D. Tao, "Deep multi-scale discriminative networks for double jpeg compression forensics," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–20, 2019.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[6] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. 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*, 2018, pp. 1316–1324.

[9] Q. Cheng and X. Gu, "Cross-modal feature alignment based hybrid attentional generative adversarial networks for text-to-image synthesis," *Digital Signal Processing*, p. 102866, 2020.

[10] M. Zhu, P. Pan, W. Chen, and Y. 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*, 2019, pp. 5802–5810.

[11] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[12] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.

[13] B. Zhu and C.-W. Ngo, "Cookgan: Causality based text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] L. Gao, D. Chen, J. Song, X. Xu, D. Zhang, and H. T. Shen, "Perceptual pyramid adversarial networks for text-to-image synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8312–8319.

[15] M. Yuan and Y. Peng, "Bridge-gan: Interpretable representation learning for text-to-image synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[16] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *Advances in Neural Information Processing Systems*, 2019, pp. 2065–2075.

[17] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, and X.-Y. Jing, "Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis," *arXiv preprint arXiv:2008.05865*, 2020.

[18] Q. Cheng and X. Gu, "Hybrid attention driven text-to-image synthesis via generative adversarial networks," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 483–495.

[19] H. Tan, X. Liu, M. Liu, B. Yin, and X. Li, "Kt-gan: Knowledge-transfer generative adversarial network for text-to-image synthesis," *IEEE Transactions on Image Processing*, vol. 30, pp. 1275–1290, 2020.

[20] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2327–2336.

[21] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[22] ——, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[23] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1462–1471.

[24] X. Fan, Y. Yang, C. Deng, J. Xu, and X. Gao, "Compressed multi-scale feature fusion network for single image super-resolution," *Signal processing*, vol. 146, pp. 50–60, 2018.

[25] Y. Yang, L. Wang, D. Xie, C. Deng, and D. Tao, "Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis," *IEEE Transactions on Image Processing*, vol. 30, pp. 2798–2809, 2021.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] D. Peng, W. Yang, C. Liu, and S. Lü, "Sam-gan: Self-attention supporting multi-stage generative adversarial networks for text-to-image synthesis," *Neural Networks*, vol. 138, pp. 57–67, 2021.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[32] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville, "Pixelvae: A latent variable model for natural images," *arXiv preprint arXiv:1611.05013*, 2016.

[33] S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. de Freitas, "Parallel multiscale autoregressive density estimation," in *ICML*, 2017.

[34] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," *arXiv preprint arXiv:1511.02793*, 2015.

[35] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 174–12 182.

[36] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6199–6208.

[37] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Advances in neural information processing systems*, 2016, pp. 217–225.

[38] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4467–4477.

[39] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Learn, imagine and create: Text-to-image generation from prior knowledge," in *Advances in Neural Information Processing Systems*, 2019, pp. 887–897.

[40] J. Cheng, F. Wu, Y. Tian, L. Wang, and D. Tao, "Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 911–10 920.

[41] T. Hinz, S. Heinrich, and S. Wermter, "Semantic object accuracy for generative text-to-image synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[42] J. Liang, W. Pei, and F. Lu, "Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis," in *European Conference on Computer Vision*. Springer, 2020, pp. 491–508.

[43] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[44] Y. Dong, Y. Zhang, L. Ma, Z. Wang, and J. Luo, "Unsupervised text-to-image synthesis," *Pattern Recognition*, p. 107573, 2020.

[45] M. Wang, C. Lang, L. Liang, S. Feng, T. Wang, and Y. Gao, "End-to-end text-to-image synthesis with spatial constrains," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 4, pp. 1–19, 2020.

[46] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[47] K. Wen, X. Gu, and Q. Cheng, "Learning dual semantic relations with graph attention for image-text matching," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020.

[48] Q. Cheng, Z. Tan, K. Wen, C. Chen, and X. Gu, "Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[49] H. Wang, C. Deng, F. Ma, and Y. Yang, "Context modulated dynamic networks for actor and action video segmentation with language queries," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 152–12 159.

[50] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[51] Z. Tan, C. Chen, K. Wen, Y. Qin, and X. Gu, "A unified two-stage group semantics propagation and contrastive learning network for co-saliency detection," *arXiv preprint arXiv:2208.06615*, 2022.

[52] C. Chen, Z. Tan, Q. Cheng, X. Jiang, Q. Liu, Y. Zhu, and X. Gu, "Utc: A unified transformer with inter-task contrastive learning for visual dialog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 103–18 112.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[55] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[57] K. Wen, J. Xia, Y. Huang, L. Li, J. Xu, and J. Shao, "Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 2208–2217.

[58] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[59] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[60] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4654–4662.

[61] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[62] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," *arXiv preprint arXiv:1909.11740*, 2019.

[63] Z. Li, F. Ling, C. Zhang, and H. Ma, "Combining global and local similarity for cross-media retrieval," *IEEE Access*, vol. 8, pp. 21 847–21 856, 2020.

[64] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," *arXiv preprint arXiv:1804.09539*, 2018.

[65] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[66] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[67] Q. Cheng and X. Gu, "Bridging multimedia heterogeneity gap via graph representation learning for cross-modal retrieval," *Neural Networks*, vol. 134, pp. 143–162, 2021.

[68] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[69] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio processing*, vol. 5, no. 3, pp. 257–265, 1997.

[70] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[72] K. Joseph, A. Pal, S. Rajanala, and V. N. Balasubramanian, "C4synth: Cross-caption cycle-consistent text-to-image synthesis," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 358–366.

[73] R. Li, N. Wang, F. Feng, G. Zhang, and X. Wang, "Exploring global and local linguistic representation for text-to-image synthesis," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[74] L. Gao, D. Chen, Z. Zhao, J. Shao, and H. T. Shen, "Lightweight dynamic conditional gan with pyramid attention for text-to-image synthesis," *Pattern Recognition*, vol. 110, p. 107384, 2021.