KT-GAN: Knowledge-Transfer Generative Adversarial Network for Text-to-Image Synthesis

Hongchen Tan, Xiuping Liu, Meng Liu, *Member, IEEE* , Baocai Yin, and Xin Li, *Senior Member, IEEE*

***Abstract*— This paper presents a new framework, Knowledge-Transfer Generative Adversarial Network (KT-GAN), for fine-grained text-to-image generation. We introduce two novel mechanisms: an Alternate Attention-Transfer Mechanism (AATM) and a Semantic Distillation Mechanism (SDM), to help generator better bridge the cross-domain gap between text and image. The AATM updates word attention weights and attention weights of image sub-regions alternately, to progressively highlight important word information and enrich details of synthesized images. The SDM uses the image encoder trained in the Image-to-Image task to guide training of the text encoder in the Text-to-Image task, for generating better text features and higher-quality images. With extensive experimental validation on two public datasets, our KT-GAN outperforms the baseline method significantly, and also achieves the competive results over different evaluation metrics.**

***Index Terms*— Generative adversarial network, knowledge distillation, Text-to-Image Generation, alternate attention-transfer mechanism.**

I. Introduction

PHOTOGRAPHIC Text-to-Image (T2I) synthesis aims to generate a realistic image that is semantically consistent with a given text description, by learning a mapping between the semantic text space and the complex RGB image space [25], [36]. A key challenge in synthesizing realistic objects with semantic details is the heterogeneous gap between high-level concepts in text descriptions and pixel-level contents of synthetic images. Building an effective synthesizer to bridge this domain gap is difficult.

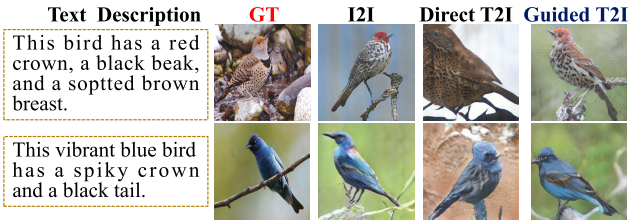


Fig. 1. Images generated by I2I task, Direct T2I task, Guided T2I task (Use I2I to guide T2I to better encode text feature and synthesize images) and the corresponding ground truth (GT).

Many approaches [11], [16], [24], [25], [41], [42] based on Generative Adversarial Networks (GANs) [9] bridge the domain gap by utilizing a discriminator to distinguish the synthesized text-image pair and the ground-truth pair. However, such a discriminator alone is usually insufficient to model underlying semantic consistency between text and image [23], and consequently, results in semantic or structural errors in synthesized images (see Figure 1, the "Direct T2I" column). Recently, the attention mechanism [13], [22], [35] has been exploited to address this problem, which guides the generator to better match certain visual words with corresponding image subregions. But using word-level attention alone does not ensure global semantic consistency due to the diversity between text and image modalities [23]. Thus, MirrorGAN [23] models Text-to-Image and Image-to-Text together to enhance global cross-domain semantic consistency. However, the Image-to-Text in MirrorGAN [23] is still a cross-domain generation, which is not easier than homogeneous generation task such as I2I task. Thus, the problem of semantic inconsistency between heterogeneous information still remains. SEGAN [13] introduces a new contrastive loss and a Semantic Consistency Module (SCM) to better align the synthesized image and the ground truth in feature space. But still due to the heterogeneous semantic inconsistency, SEGAN cannot extract effective text features that can guide the synthesis of realistic and detailed images.

Our observation is that Image-to-Image (I2I) synthesis belongs to a homogeneous generation task, whose information

Manuscript received March 1, 2020; revised August 14, 2020; accepted September 19, 2020. Date of publication October 1, 2020; date of current version December 23, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61976040; in part by the Ministry of Science and Technology of the People's Republic of China 2018AAA0102003; in part by the National Science Foundation of USA OIA-1946231; and in part by the Science and Technology Foundation of Dalian 2018J11CY010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jiaying Liu. *(Corresponding authors: Xiuping Liu; Xin Li.)*

Hongchen Tan and Xiuping Liu are with the School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China (e-mail: tanhongchenphd@mail.dlut.edu.cn; xpliu@dlut.edu.cn).

Meng Liu is with the School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China (e-mail: mengliu.sdu@gmail.com).

Baocai Yin is with the Department of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: ybc@dlut.edu.cn).

Xin Li is with the School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA 70803 USA, and also with the Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803 USA (e-mail: xinli@cct.lsu.edu).

Digital Object Identifier 10.1109/TIP.2020.3026728

1057-7149 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

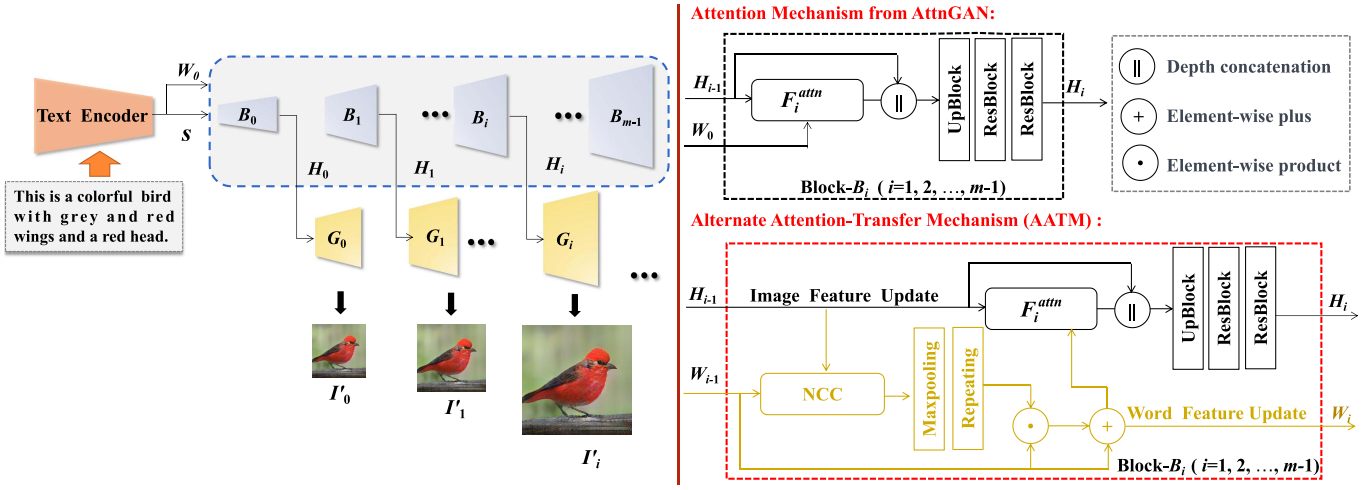


Fig. 2. The left part is cascade-attention generation networks in AttnGAN [35]. The black dotted box is the attention block in AttnGAN. The red dotted box is the attention block in AATM. NCC denotes the Normalized Cross Correlation.

gap is much smaller than that between heterogeneous gap. I2I can generate a synthesized image that has much stronger semantic consistency with the ground truth image. Thus, I2I may effectively guide T2I to better encode text features and synthesize images. As shown in Figure 1, T2I guided by I2I (last column) produces much better results than direct T2I synthesis.

In addition, the recent AttnGAN [35] and its subsequent improvements [10], [13], [17], [22], [31] adopt the word-level attention mechanism to enhance the local semantic matching between word features and local image features. In these algorithms, during the process of image synthesis, the weights of word features are fixed. Consequently, if the attention mechanism does not produce accurate weight estimation in one pass, then important words will be neglected, and image details will be missing. Thus, our second technical development is an *attention update mechanism* so that attention module can gradually focus on important words in a progressive way during the process of image synthesis.

Based on the above observations, we propose a new Knowledge-Transfer Generative Adversarial Network (KT-GAN) for T2I synthesis with two new components: (1) a Semantic Distillation Mechanism (SDM) that uses the I2I task to guide the T2I task for both text feature encoding and image generation, and (2) an Alternate Attention-Transfer Mechanism (AATM) to better identify important words in text. The main contributions of this paper are as follows:

(i) We designed a Semantic Distillation Mechanism with a novel distillation loss function, to use I2I to guide T2I for better text feature extraction and image synthesis.

(ii) We designed an Alternate Attention-Transfer Mechanism to progressively refine word-level attention weights and enrich details of the synthesized image.

(iii) We validated our KT-GAN on two datasets: CUB-Bird [32] and large-scale MS-COCO [19]. Extensive experimental results and analysis demonstrate the effectiveness of KT-GAN and significantly improved performance compared against most previous most state-of-the-art methods on all four evaluation metrics.

II. Related Work

A. Semantic Attention Mechanism

Attention models have been extensively exploited in computer vision and natural language processing, for instance in object detection [40], image/video captioning [39], person Re-identification [33] and visual question answering [29]. In T2I synthesis, recently, AttnGAN [35] introduces the word-level visual attention mechanism for T2I synthesis, it enhances the synthesis of fine-grained details at different image regions. The following work, obj-GAN [17] proposes a object-driven attention mechanism to further improve the detail synthesis, and produces finer images. A limitation of both AttnGAN [35] and obj-GAN [17] is that weight of words are fixed, which sometimes results in attention mechanism neglecting many important words. Recently, SEGAN [13] introduce the attention regularization term to filter out unimportant words and highlight the important words. But it is difficult to find the adaptive threshold in the regularization term.

Recently, [43] propose the attention update mechanism to transfer the pose of a given person to a target pose. The mechanism can effectively and dynamically utilize pose and appearance features to smoothly guide the pose transfer process. Inspired by [43], we propose the suitable attention update mechanism to progressively focus on important words.

B. Knowledge Distillation

Knowledge Distillation (KD) [15], [34] with neural networks was pioneered by Hinton [8], which is a transfer learning method that aims to improve the training of a student network by relying on knowledge borrowed from a powerful teacher network. This has also been addressed for model compression [2], [7], cross-domain task [1], [20] and continual learning tasks [21], [6]. Reference [2] propose two distinct teacher-student frameworks based on knowledge distillation

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

mechanism for person detection. Reference [20] apply domain-invariant feature distillation mechanism for cross-domain sentiment classification. Lifelong GAN [21] employs knowledge distillation to address catastrophic forgetting for conditional generative continual learning tasks. However, above works belongs to homologous information distillation mechanism, because both teacher task and student task belong to the same or similar tasks. In KT-GAN, we solve the more challenging heterogeneous information distillation task. In our KT-GAN, teacher task (I2I) is quite different from student task (T2I).

III. Proposed Method

We first revisit the attention mechanism in AttnGAN (Sec. III-A), then introduce our two new mechanisms: the Alternate Attention-Transfer Mechanism (AATM) (Sec. III-B) and Semantic Distillation Mechanism (SDM) (Sec. III-C). By integrating the AATM and SDM into the AttnGAN, we get our proposed KT-GAN.

A. Attention Generator in AttnGAN [35] Revisit

The AttnGAN [35] introduced an Attentional Generative Network (AGN) to guide the synthesis of different sub-regions in the image following their most relevant words. The input text description is transformed into the sentence feature s and word features W_0 by a pre-trained bi-directional LSTM text encoder in [35].

As shown in left part of Figure 2, the AGN has m blocks ($B_0, B_1, \cdots, B_{m-1}$) and the corresponding m generators ($G_0, G_1, \cdots, G_{m-1}$). The generators take the hidden states ($H_0, H_1, \cdots, H_{m-1}$) as input and generate images of small-to-large scales ($I'_0, I'_1, \cdots, I'_{m-1}$):

$$\begin{aligned} B_0 : H_0 &= F_0(z, F^{ca}(s)); \\ B_i : H_i &= F_i(H_{i-1} || F_i^{attn}(W_0, H_{i-1})), i = 1, 2, \cdots, m-1; \\ G_i : I'_i &= G_i(H_i). \end{aligned} \tag{1}$$

Here, $z \sim N(0, 1)$. F^{ca} is a conditioning augmentation module [41] that converts a sentence feature s to a conditioning feature for the generator. F^{ca}, F_i^{attn}, F_i, and G_i are modeled as neural networks. Here, $\|$ denotes the concatenation of two maps along depth axis.

Consider the i-th block B_i, the black dotted box in the middle row of Figure 2: the core of B_i is $F_i^{attn}(W_0, H_{i-1})$, which can update where to draw or highlight the details of the image according to word feature W_0. $F_i^{attn}(W_0, H_{i-1})$ has two inputs: the word features $W_0 \in \mathbb{R}^{D \times T}$ (T is the number of words, D is the dimension of word features) and the image features from the previous hidden layer $H_{i-1} \in \mathbb{R}^{\hat{D} \times N}$. F_i^{attn} is computed in three steps: (1) Normalized Cross Correlation (NCC) between W_0 and H_{i-1} is computed as the attention weights to words; (2) A word-context matrix $F_i^{attn}(W_0, H_{i-1})$ for image feature is computed; and (3) the image feature H_i is updated by: $H_i = F_i(H_{i-1} \parallel F^{attn}(W_0, H_{i-1}))$. For more details please refer to AttnGAN [35].

B. AATM

We can observe that the input of word feature is always W_0 in the block $B_i (i = 1, 2, 3, \cdots, m-1)$ of AttnGAN [35]. A problem for the attention mechanism in AttnGAN is that word weights calculated in a single attentional block are not guaranteed to be correct. As a result, some visually important words could be neglected and their semantics are not reflected in the synthesized image. In order to tackle this problem, we design an Alternate Attention-Transfer Mechanism (AATM) to iteratively and progressively identify visually important words in the sentence. We construct the AATM by introducing a Word Feature Update module into the block $B_i (i = 1, 2, 3, \cdots, m-1)$, as illustrated in the red dotted box in Figure 2.

Each block B_i in AATM contains two modules: Word Feature Update module (components in yellow in Fig. 2) and Image Feature Update module (components in black in Fig. 2). Firstly, the Word Feature Update module updates the weight of word feature based on the image feature and word feature from the last block. With these blocks, important words will gradually aggregate their weights and they will get highlighted. Secondly, the image features should also be updated according to such change, i.e., image features are synchronized to indicate where to draw the detail on the image according to the updated word features.

1) Word Feature Update Module: In the i-th block B_i, it takes in image features H_{i-1} and word features W_{i-1} from the B_{i-1} block, and outputs the updated word features W_i through a three-step procedure.

(Step 1) Calculate an NCC Matrix $R^* \in \mathbb{R}^{T \times N}$ between $H_{i-1} \in \mathbb{R}^{\hat{D} \times N}$ and $W_{i-1} \in \mathbb{R}^{D \times T}$: (i) We map word features to the same latent semantic space of the image features by $W'_{i-1} = U W_{i-1}$, $W'_{i-1} = \{w'^{j}_{i-1} \in \mathbb{R}^{\hat{D}} | j = 1, 2, \cdots, T\}$, where $U \in \mathbb{R}^{\hat{D} \times D}$ is a perceptual layer. Each column of $H_{i-1} = \{h^j_{i-1} \in \mathbb{R}^{\hat{D}} | j = 1, 2, \cdots, N\}$ (hidden features) is a feature vector of an image's sub-region. (ii) The word-image Cross Correlation Matrix is $R = (r_{i,j}) = W'^{T}_{i-1} H_{i-1} \in \mathbb{R}^{T \times N}$. Here, $r_{i,j}$ encodes the dot-product similarity between the i^{th} word in the sentence and j^{th} sub-region in the image. (iii) The NCC Matrix R^* is $R^* = (r^*_{i,j}) = \frac{exp(r_{i,j})}{\sum_{k=1}^{T} exp(r_{k,j})}$.

(Step 2) Calculate attentional weight mask $\hat{R}$ of words: (i) We reshape $R^* \in \mathbb{R}^{T \times N}$ to $R' \in \mathbb{R}^{T}$ by maxpooling. Each element in R' represents the maximum similarity of a word to all image sub-regions, which is regarded as this word's weight. (ii) In order to match R' with the word feature matrix $W_{i-1} \in \mathbb{R}^{T \times D}$, we reshape R' to attention weight mask $\hat{R} \in \mathbb{R}^{T \times D}$ by repeating the column of R' for D times.

(Step 3) Update the word feature W_i by

$$W_i = \alpha \cdot [\hat{R} \odot W_{i-1}] \oplus \beta \cdot [W_{i-1}], \quad (i = 1, 2, 3, \ldots m-1) \tag{2}$$

where $\odot$ denotes element-wise product, α is the retention factor of current word feature information, β is the attenuation parameter of the word feature information in the previous stage, which reduces the interference of non-important words to some extent. By multiplying the transformed word features

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

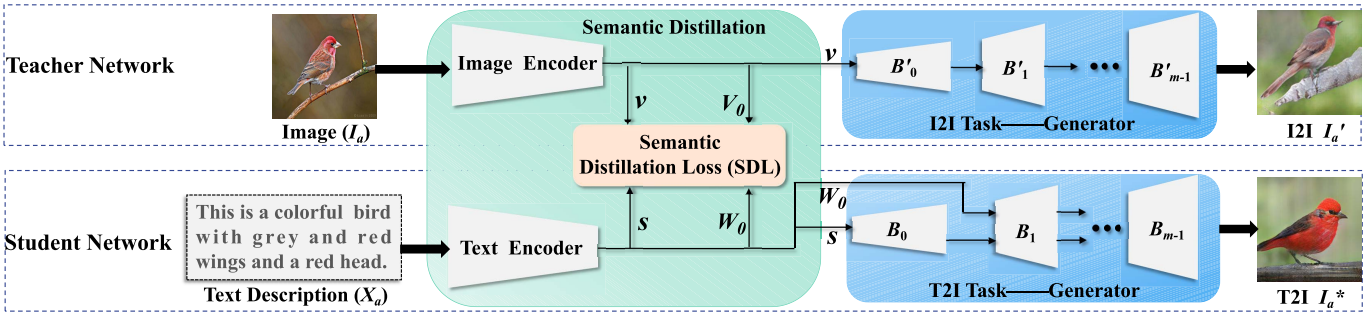


Fig. 3. The Framework of the proposed KT-GAN. KT-GAN contains two novel strategies: Semantic Distillation Mechanism (SDM) and Alternate Attention-Transfer Mechanism (AATM) in Figure2. The AATM is embedded in the generator of the student task. The SDM contains three main components: the teacher task (Image-to-Image task, I2I), the student task (Text-to-Image task, I2I), and the Semantic Distillation Loss (SDL).

W_{i-1} with the weight $\hat{R}$, features W_i of certain words are either preserved or suppressed. The **residual connection** $\oplus$ and **the progressive feedback from discriminators of previous blocks** together help important words to aggregate their weights quickly.

2) Image Feature Update Module: We use the whole operations (Eq. 1) F_i in AttnGAN as the Image Feature Update module in our each block B_i. However, in AttnGAN the input word feature of F_i is always W_0, but in our AATM, the word features and image features need to be updated alternately, thus, in our each B_i, we need to rewrite $H_i = F_i(H_{i-1}||F_i^{attn}(W_0, H_{i-1}))$ in Eq. 1 as $H_i = F_i(H_{i-1}||F_i^{attn}(W_i, H_{i-1}))$. With this modification, our AATM can be implemented correctly.

C. SDM

Besides introducing a better attention mechanism (AATM) into the T2I, we also design a new Semantic Distillation Mechanism (SDM) to guide the text encoder to provide better input features for the T2I. As illustrated in the green box in Figure 3, the SDM trains a student task (T2I) under the guidance of a trained teacher task (I2I), which performs the supervised cross-task semantic transfer. Our SDM contains three main components: the teacher task (I2I), the student task (T2I), and the Semantic Distillation Loss (SDL).

In the teacher task, we train an I2I task to obtain a good image encoder and a good image generator. We design an SDL to train a good text encoder, *GTE* (Guided Text Encoder), that matches with the trained image encoder in I2I. In the student task, we train a T2I task to get a good image generator. The implementation steps are as follows.

Step 1: Train networks in the teacher task (I2I).

Step 2: Train text encoder by SDL.

Step 3: Train networks in the student task (T2I).

In **Step 1**, I2I transforms an input image into a global image feature $v \in \mathbb{R}^D$ by an image encoder. This image encoder is initialized using a pre-trained Inception-V3 [28] on ImageNet [26], and is then fine-tuned during the I2I training. We modify AttnGAN [35] from an T2I synthesizer to an I2I synthesizer. Note that choosing the structure of AttnGAN [35] as the teacher's network is appropriate, because it aligns well with the T2I task. The modification of AttnGAN mainly includes two operations. **(i)** Remove the attention mechanism in the generator, use global image feature v as the input of the generator. **(ii)** Introduce the perceptual loss ($\mathcal{L}_{per}$) [4] into the I2I to improve the quality of synthesized image. More details of these modifications can be found in Appendix-A.

In **Step 2**, we use use the trained image encoder in I2I task to guide training of an effective text encoder for the T2I task.

In **Step 3**, we use the text encoder trained by SDL in **Step 2** to generate text feature as the input to T2I. Meanwhile, we use the generator and discriminator trained in **Step 1** as the initial generator and discriminator. Rather than training the generator and discriminator from scratch, this inheritance greatly improves the generator's performance.

1) Semantic Distillation Loss (SDL): The core of SDM is SDL. In this distillation, image encoder needs to be fixed, because it was already trained in I2I, and can provide effective feature template for text encoder in T2I to follow. In training process, we should design the SDL to (1) globally, push the sentence feature s to match the fixed global image feature v, and (2) locally, push the word feature W_0 to match the fixed image feature's sub-regions V_0. The DAMSM loss in [35] is a widely used function to match image features and text features. We made *two main modifications* on DAMSM Loss. (1) First, unlike AttnGAN [35] that trains both image encoder and text encoder, we modify DAMSM loss to fix image encoder and only train text encoder. (2) Second, DAMSM can not effectively deal with imbalanced easy/difficult data samples in training process. Inspired by [18], Focal Loss can better balance the easy and hard samples. Thus, we further revise the construction of SDL following the design of Focal Loss.

The SDL is composed of $\mathcal{L}_{S*}$ and $\mathcal{L}_{W*}$. The distillation loss $\mathcal{L}_{S*}$ ("S*" stands for "Sentence") is designed for matching sentence feature s and global image feature v. The distillation loss $\mathcal{L}_{W*}$ between Word feature W_0 and local image feature V_0 ("W*" stands for "Words") is designed similarly.

2) The Distillation Loss $\mathcal{L}_{S*}$: For a batch of image-sentence pairs $\{(v_i, s_i)\}_{i=1}^M$, firstly we define the probability of matching between sentence feature s_i and fixed image feature v_i^* (trained in I2I) as $P(s_i, v_i^*)$:

$$P(s_i, v_i^*) = \frac{2 \cdot exp(d(v_i^*, s_i))}{\sum_{j=1}^M exp(d(v_j^*, s_i)) + \sum_{j=1}^M exp(d(v_i^*, s_j))}. \quad (3)$$

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

Here, $d(x, y)$ is cosine distance between x and y. Then, we define the L_{S^*} as the negative log matching probability as

$$\mathcal{L}_{S^*} = -\sum_{i=1}^{M}(1 - P(s_i, v_i^*))^{\gamma} log P(s_i, v_i^*). \quad (4)$$

Here, γ is tunable focusing parameter $\gamma \geq 0$ in [18].

3) The Distillation Loss $\mathcal{L}_{W^*}$: In $\mathcal{L}_{W^*}$, we aim to make the word feature $W_0 = \{w_0^j \in \mathbb{R}^D | j = 1, 2, \cdots, T\}$ (output of the text encoder) to align with the **fixed** image feature's sub-regions $V_0^* = \{v_0^{*j} \in \mathbb{R}^D | j = 1, 2, \cdots, N\}$ (output of the fixed image encoder).

(1) Following [14], [17], [35], we define a **normalized image-text similarity matrix** $R^+ \in \mathbb{R}^{T \times N}$ through two operations:

(1.a) Let $R = (r_{i,j}) = W_0^T V_0^* \in \mathbb{R}^{T \times N}$ encode the dot-product similarity between the i-th word in the sentence and j-th sub-region in the image.

(1.b) This R is normalized to $R^* = (r_{i,j}^*) = \frac{exp(r_{i,j})}{\sum_{k=1}^{T} exp(r_{k,j})}$, and then normalized to $R^+ = (r_{i,j}^+) = \frac{exp(r_{i,j}^*)}{\sum_{k=1}^{N} exp(r_{i,k}^*)}$.

(2) Same as [14], [17], [35], we also define the **attention-driven word-image matching score**:

(2.a) We define the *dynamic representation* of the image with respect to the i-th word using a vector $O_i = \sum_{j=1}^{N} r_{i,j}^+ v_0^{*j}$.

(2.b) We can define the **attention-driven word-image matching score** between image v^* and sentence s,

$$R_W(v^*, s) = log(\sum_{i=1}^{T} exp(\beta_0 O_i \cdot w_0^i))^{\frac{1}{\beta_0}}, \quad (5)$$

where β_0 is a factor weighing the importance of the most relevant word-to-region pair. As described in [35], when $\beta_0 \to \infty$, $R_W(v^*, s) \to max_{i=1}^{T} O_i \cdot w_0^i$.

(3) Based on the **attention-driven word-image matching score**, for a batch of image-sentence pairs $\{(v_i^*, s_i)\}_{i=1}^{M}$, we define the probability of matching between sentence feature s_i and fixed image feature v_i^* (trained in I2I) as $P(s_i, v_i^*)$:

$$P(s_i, v_i^*) = \frac{2 \cdot exp(R_W(v_i^*, s_i))}{\sum_{j=1}^{M} exp(R_W(v_i^*, s_j)) + \sum_{j=1}^{M} exp(R_W(v_j^*, s_i))}, \quad (6)$$

Finally, follow the design of Focal Loss [18], we define the Distillation Loss $\mathcal{L}_{W^*}$ by introducing a modulating factor $(1 - P(s_i, v_i^*))^{\gamma}$ with a tunable focusing parameter $\gamma \geq 0$:

$$\mathcal{L}_{W^*} = -\sum_{i=1}^{M}(1 - P(s_i, v_i^*))^{\gamma} log P(s_i, v_i^*). \quad (7)$$

Here, the value of γ in $\mathcal{L}_{W^*}$ is the same as the value of γ in $\mathcal{L}_{S^*}$

Finally, the Semantic Distillation Loss is defined as

$$\mathcal{L}_{SDL} = \lambda_1 \mathcal{L}_{S^*} + \lambda_2 \mathcal{L}_{W^*}. \quad (8)$$

Here, the subscript SDL stands for "Semantic Distillation Loss".

4) Generative and Discriminative Loss in Student Task: At the Block-B_i, the Generative loss $\mathcal{L}_{G_i}$ and Discriminative loss $\mathcal{L}_{D_i}$ are defined as

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}[\mathbb{E}_{I_i' \sim P_{G_i}} log D_i(I_i')}_{\text{unconditional loss}} + \underbrace{\mathbb{E}_{I_i' \sim P_{G_i}} log D_i(I_i', s)]}_{\text{conditional loss}}, \quad (9)$$

where the unconditional loss is trained to generate images towards the true image distribution to fool the discriminator, and the conditional loss is trained to generate images to match text descriptions.

The discriminator D_i is trained to classify the input into the class of real or fake images by minimizing the cross-entropy loss

$$\mathcal{L}_{D_i} = \underbrace{-\frac{1}{2}[\mathbb{E}_{I_i \sim P_{data_i}} log D_i(I_i) + \mathbb{E}_{I_i' \sim P_{G_i}} log(1 - D_i(I_i'))}_{\text{unconditional loss}} \underbrace{+\mathbb{E}_{I_i \sim P_{data_i}} log D_i(I_i, s) + \mathbb{E}_{I_i' \sim P_{G_i}} log(1 - D_i(I_i', s))]}_{\text{conditional loss}}, \quad (10)$$

where I_i is from the true image distribution p_{data} at the i^{th} scale, and I_i' is from distribution p_{G_i} of the generative images at the same scale.

To generate realistic images, the final loss function of the generator and discriminator are defined as

$$\mathcal{L}_G = \mathcal{L}_G + \lambda_3 \mathcal{L}_{DAMSM}, \mathcal{L}_D = \sum_{i=0}^{m-1} \mathcal{L}_{D_i}, \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}. \quad (11)$$

Here, we utilize the DAMSM loss [35] to make generated images better conditioned on text descriptions.

IV. Experimental Results

A. Experiment Settings

1) Implementation Details: The resolution of the images participating in the evaluation are 256×256. All hyper-parameter values are listed in the Appendix-B. We find the suitable values for these hyper-parameters by a series of ablation studies in the Appendix-C. In the training stage: (i) Train I2I task; (ii) Train SDM; (iii) Train T2I task. In the testing stage: Only input features of sentence and words to the generator of the Student Network.

2) Datasets: Two widely used datasets are used. The CUB-Bird dataset [32] contains 11, 788 bird images belonging to 200 categories, and 10 visual description sentences for each image. We pre-process and split the images following the same pipeline in [25], [41]. The MS-COCO dataset [19] contains $80k$ training images and $40k$ test images, and each image has 5 text annotations.

3) Evaluation: We compare KT-GAN and other state-of-the-art algorithms using four measures: (1) Inception Score (**IS**) [27]; (2) Fréchet Inception Distance (**FID**) [12]; (3) We also compare the **Rank-1** score in text-to-image retrieval [5]; (4) **Human perceptual test**.

IS uses fine-tuned inception models provided by [41] to compute the KL-divergence between the conditional class

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

TABLE I

IS↑, FID ↓ AND RANK-1 ↑ PRODUCED BY COMBINING DIFFERENT COMPONENTS OF THE KT-GAN ON CUB-BIRD AND MS-COCO TEST SETS. KT-GAN=ATTNGAN+AATM+SDM

Method	CUB-Bird			MS-COCO		
	IS	FID	Rank-1	IS	FID	Rank-1
AttnGAN [35]	4.36 ± 0.03	23.98	27.9%	25.89 ± 0.47	35.49	22.9%
AttnGAN+SDM	4.76 ± 0.02	18.21	32.6%	29.02 ± 0.17	31.86	24.4%
AttnGAN+AATM	4.74 ± 0.05	20.40	29.4%	28.54 ± 0.38	32.54	23.7%
KT-GAN	4.85 ± 0.04	17.32	32.9%	31.67 ± 0.36	30.73	24.5%

distribution and the marginal class distribution. A larger **IS** (better) means that a T2I generator can synthesize a higher diversity of images for all classes, where each image belongs more clearly to a specific class. **FID** computes the Fréchet distance between synthetic and realistic images based on the extracted features from a pre-trained Inception-V3 network [28]. A lower **FID** (better) implies a closer distance between generated image distribution and real-world image distribution. The **Rank-1** score denotes the most relevant synthesized images for each text sentence in text-to-image retrieval. A bigger **Rank-1** score (better) implies the synthesized image has better consistency with the given text. The trained retrieval model provided by [5] was used to calculate the **Rank-1**. The **Human perceptual test** aims to judge whether the generated images are well-conditioned on the text descriptions from human subjective perception. In order to conduct the **Human perceptual test**, we randomly select 1000 text descriptions in the CUB-Bird test set and 2000 text descriptions in the MS-COCO test set. Given the same text description, 30 volunteers (not including any author) are asked to rank the images generated by different methods. The average ratio ranked as the best by human users are calculated to evaluate the compared methods. A bigger score (better) in the **Human perceptual test** implies the synthesized image has better consistency with the given text description.

B. Effectiveness of New Modules

We evaluate the effectiveness of two new components, AATM and SDM, in terms of three measures. The results are documented in Table I. (I) We introduce AATM to replace the attention mechanism in AttnGAN [35], i.e. AttnGAN+AATM. As shown in Table I, AttnGAN+AATM leads to 8.7% and 10.2% improvement of **IS**, 24.1% and 10.2% improvement of **FID**, and 4.7% and 1.5% improvement of **Rank-1**, on CUB-Bird and MS-COCO test datasets respectively. (II) If we introduce the SDM to AttnGAN (AttnGAN+SDM), we obtain 9.2% and 12.1% improvement over the AttnGAN in **IS**, 14.9% and 8.3% improvement over the AttnGAN in **FID**, and 1.5% and 0.8% improvement over the AttnGAN in **Rank-1**, on CUB-Bird and MS-COCO datasets respectively. (III) KT-GAN: If we introduce the SDM and AATM into AttnGAN, we obtain 11.2% and 22.3% improvement over the AttnGAN in **IS**, 27.8% and 15.5% improvement over the AttnGAN in **FID**, and 5.0% and 1.6% improvement over the AttnGAN in **Rank-1**, on CUB-Bird and MS-COCO datasets respectively.

In all, Table I shows that both components contribute to the KT-GAN's performance improvement. The **IS** of KT-GAN is 4.85 on CUB-Bird and 31.67 on the MS-COCO test dataset. The **FID** of KT-GAN is 17.32 on CUB-Bird and 30.73 on the MS-COCO test dataset. The **Rank-1** of KT-GAN is 32.9% on CUB-Bird and 24.5% on the MS-COCO test dataset.

C. Component Analysis of AATM

We compare **IS**, **FID** and **Rank-1** of different designs in the Word Feature Update Module (Eq. 2) on the CUB-Bird dataset. Due to GPU memory constraints, we did not try the KT-GAN with more than three blocks. And AATM is employed over the last two blocks. Thus, AATM can only be applied to blocks B_1 and/or B_2, and there are three possible variants: (**I**) AATM (B_1) indicates that AATM is only implemented in block B_1. As shown in Table II, compared with AttnGAN, the performance of AttnGAN+AATM (B_1) gains the moderate improvements in these three measures. (**II**) AttnGAN+AATM indicates that AATM is implemented in both blocks B_1 and B_2 in this paper. Compared with AttnGAN+AATM (B_1), the performance of AttnGAN+AATM (B_1) further gains the moderate improvements in these three measures. It demonstrates that progressively adding the AATM into the block (B_i ($i = 1, 2, \ldots, m-1$)) can effectively improve the performance of generator. (**III**) In order to show the effectiveness of the residual connection, we drop out residual connection in Eq. 2, i.e. $W_i = \alpha \cdot [\hat{R} \odot W_{i-1}]$. The performance of "AttnGAN+AATM w/o Res" is degraded in these three measures. It indicates that the accumulation of word information from previous blocks also play an important role in synthesizing process.

Figure 4 compares AttnGAN and AttnGAN+AATM by visualizing the iterative update on word weights and their corresponding attention maps. Weights for the top-5 words are listed. The attention maps are plotted on synthesized images: each word's relevant region is brighter. In Figure 4, the AttnGAN generally can not effectively accumulate important word information and improve attention maps. In the CUB-Bird example (left column), word weights and attention in Block-B_2 are not better than in Block-B_1 and remain incorrect (e.g., "a" "the" "short" and "beak"). Similarly, in the MS-COCO example (right column), the AttnGAN's attention always focuses on certain words and misses important visual information from some other words, and eventually leads to bad synthesis. In contrast, our AATM progressively aggregates

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

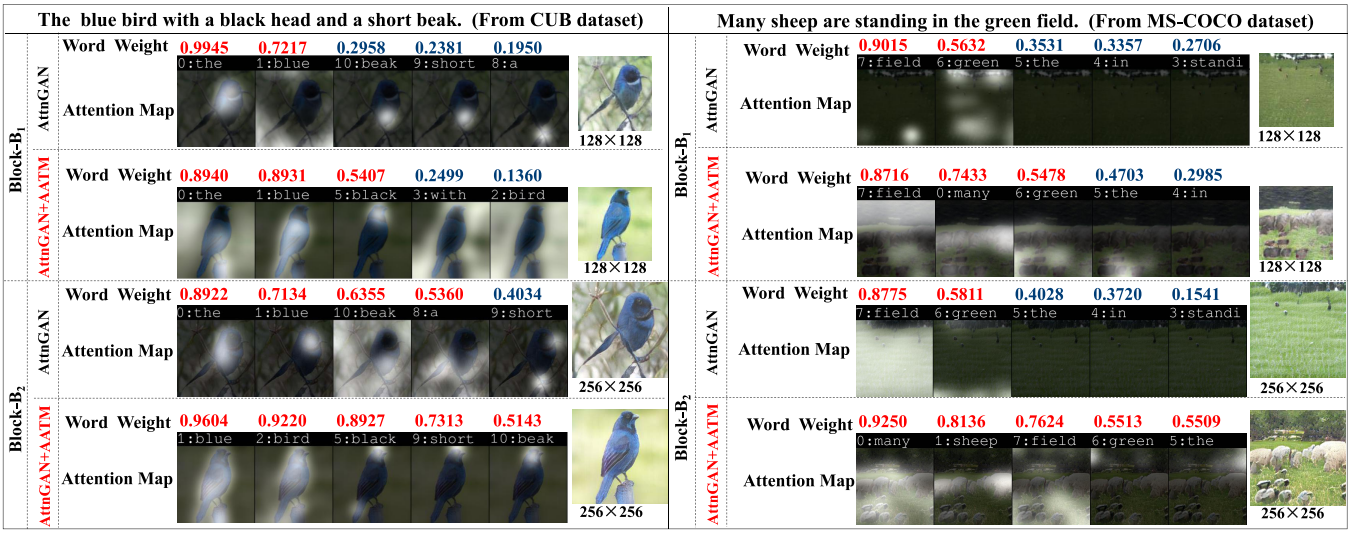


Fig. 4. The top-5 word weights and synthesized images' attention maps from AttnGAN and AttnGAN+AATM. The red mark indicates that the weight > 0.5.

TABLE II
IS↑, **FID** ↓ AND **RANK-1** ↑ ON CUB-BIRD TESTING DATA ABOUT VARIANTS OF EQ. 2 IN WORD FEATURE UPDATE MODULE

Method	IS	FID	Rank-1
AttnGAN [35]	4.36 ± 0.02	23.98	27.9%
AttnGAN+AATM (B_1)	4.70 ± 0.03	20.76	29.0%
AttnGAN+AATM	4.74 ± 0.05	20.40	29.4%
AttnGAN+AATM w/o Res	4.49 ± 0.04	22.17	28.3%

the weights of important words and enhances the accuracy of their attention. We can clearly see such improvements in Block-B_2 over Block-B_1 (e.g., left column: "beak" and "short", right column: "many", "sheep" and "field"). As for the update module for the Image Feature, the module from [35] is already effective and we simply followed that. So no ablation study on that is discussed here.

D. Component Analysis of SDM

We use ablation study to verify the two designs in our SDM: (1) I2I can guides T2I to get better text encoder, and (2) I2I generators are good initial generators in T2I.

For (1), we use the Guided Text Encoder (GTE) trained by SDL to train AttnGAN+AATM from scratch, i.e. AttnGAN+AATM+GTE. As shown in Table III, compared with AttnGAN+AATM without distillation: the **IS** rises from 4.74 to 4.80 on CUB-Bird dataset, and from 28.54 to 30.03 on MS-COCO dataset; the **FID** declines from 20.40 to 18.13 on CUB-Bird dataset, and from 32.54 to 31.22 on MS-COCO dataset; the **Rank-1** rises from 29.4% to 32.0% on CUB-Bird dataset, and from 23.7% to 24.2% on MS-COCO dataset. For (2), if we use the trained generator and discriminator in I2I as the initial generator and discriminator in T2I, i.e. AttnGAN+AATM*. As shown in Table III, compared with AttnGAN+AATM, the performance of AttnGAN+AATM* also gains the moderate improvements on the CUB-bird dataset and MS-COCO dataset over three measures respectively. Finally, we incorporate both designs into the AttnGAN+AATM, i.e. KT-GAN. Compared with AttnGAN+AATM: the **IS** rises from 4.74 to 4.85 on CUB-Bird dataset, and from 28.54 to 31.67 on MS-COCO dataset; The **FID** declines from 20.40 to 17.32 on CUB-Bird, and from 32.54 to 30.73 on MS-COCO dataset; The **Rank-1** rises from 29.4% to 32.9% on CUB-Bird dataset, and from 23.7% to 24.5% on MS-COCO dataset.

Thus, the effectiveness of the two designs can be demonstrated by these ablation studies. And it indicates that the SDM can help the T2I task bridge the domain gap.

E. Comparison With State-of-the-Art GAN Models

We compare our KT-GAN with state-of-the-art GAN models for text-to-image synthesis on CUB-Bird and MS-COCO test datasets. The **IS** for our proposed KT-GAN and other compared methods are listed in Table IV. On the CUB-Bird dataset, our KT-GAN (4.85) leads to the highest **IS** scores. On the MS-COCO dataset, the KT-GAN (31.67) also performs better than most existing approaches except for the SD-GAN [10]. However, SD-GAN requires multiple text sentences to train the generator. If the given database only contains images with single sentence description (which is common in some practical tasks such as Story Visualization [38] and Text-to-Video [37]), SD-GAN can not be trained. In contrast, KT-GAN and AttnGAN [35] only need one sentence per image and they can be trained normally. Besides, the SD-GAN contains many Siamese branches, which is much more complex than KT-GAN. Thus, SD-GAN require much more powerful hardware devices for training.

In Table V, we compare the different models' performance using **FID** and **Rank-1**. The KT-GAN greatly improves the baseline AttnGAN [35] in terms of **FID** and **Rank-1**, on CUB-Bird and MS-COCO dataset respectively. And KT-GAN achieves the best score in terms of **Rank-1** on the two standard datasets. However, the **FID** of KT-GAN is lower than that

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

TABLE III

IS↑, FID ↓ AND RANK-1 ↑ OF DIFFERENT VARIANTS OF SDM ON THE CUB-BIRD AND MS-COCO DATASETS

Method	CUB-Bird			MS-COCO		
	IS	FID	Rank-1	IS	FID	Rank-1
AttnGAN+AATM	4.74 ± 0.05	20.40	29.4%	28.54 ± 0.38	32.54	23.7%
AttnGAN+AATM+GTE	4.80 ± 0.07	18.13	32.0%	30.03 ± 0.44	31.22	24.2%
AttnGAN+AATM*	4.81 ± 0.06	17.45	32.8%	29.28 ± 0.60	30.89	24.5%
KT-GAN	4.85 ± 0.04	17.32	32.9%	31.67 ± 0.36	30.73	24.5%

TABLE IV

IS↑ BY STATE-OF-THE-ART GAN MODELS AND OUR KT-GAN ON CUB-BIRD AND MS-COCO TEST DATASETS. ATTNGAN IS OUR BASELINE MODEL. THE FIRST, SECOND AND THIRD SCORES ARE SHOWN IN RED, GREEN AND BLUE RESPECTIVELY

Methods	CUB-Bird	MS-COCO
GAN-INT-CLS [25]	2.88 ± 0.04	7.88 ± 0.07
StackGAN [41]	3.70 ± 0.04	8.45 ± 0.03
StackGAN++ [11]	3.84 ± 0.06	8.30 ± 0.10
HDGAN [42]	4.15 ± 0.05	11.86 ± 0.18
AttnGAN [35]	4.36 ± 0.02	25.89 ± 0.19
AttnGAN+O.P.*[31]	-	24.76 ± 0.43
Obj-GAN [17]	-	30.29 ± 0.33
MirrorGAN [23]	4.56 ± 0.05	26.47 ± 0.41
ControlGAN [3]	4.58 ± 0.09	24.06 ± 0.60
LeicaGAN [30]	4.62 ± 0.06	-
SEGAN [13]	4.67 ± 0.04	27.86 ± 0.31
SD-GAN [10]	4.67 ± 0.09	35.69 ± 0.50
DM-GAN [22]	4.75 ± 0.07	30.49 ± 0.57
KT-GAN (Proposed)	4.85 ± 0.04	31.67 ± 0.36

TABLE V

FID ↓ AND RANK-1 ↑ BY SOME GAN MODELS AND OUR KT-GAN ON CUB-BIRD AND MS-COCO TEST DATASETS. ATTNGAN IS OUR BASELINE MODEL. THE FIRST, SECOND AND THIRD SCORES ARE SHOWN IN RED, GREEN AND BLUE RESPECTIVELY

Method	CUB-Bird		MS-COCO	
	FID	Rank-1	FID	Rank-1
StackGAN [41]	51.89	22.8%	74.05	-
HDGAN [42]	25.17	24.6%	71.27	19.9%
AttnGAN [35]	23.98	27.9%	35.49	22.9%
SEGAN [13]	18.17	30.2%	32.28	23.3%
Obj-GAN [17]	-	-	25.64	24.1%
DM-GAN [22]	16.09	31.7%	32.64	23.6%
KT-GAN (Proposed)	17.32	32.9%	30.73	24.5%

of DMGAN. Compared with the text descriptions in MS-COCO dataset, the text descriptions in CUB-Bird datasets is more detailed and localized. The Dynamic Memory Module in DMGAN is a kind of local attention mechanism. Our KT-GAN combines the global knowledge distillation strategy with the local word attention enhancement strategy. Compared with our KT-GAN, the DMGAN pays more attention on the details generation indeed. Besides, the Dynamic Memory Module in DMGAN contains too many learning parameters. The good Dynamic Memory Module with more learning parameters can better drive the generator to learn the real data distribution. So, in the CUB-Bird dataset, the **FID** of DMGAN is better than that of our KT-GAN. In the semantic consistent aspect, our KT-GAN enhances the semantic consistent from global and local aspects. Thus, the "Rank-1" of our KT-GAN are better than that of DMGAN. Besides, because the Dynamic Memory Module in DMGAN contains too many learning parameters. So, the DMGAN's generator is more complex than our KT-GAN. In all, the results of **FID** demonstrates that KT-GAN performs better in capturing the feature distribution of more complex real images. The results of **Rank-1** demonstrates the KT-GAN leads to better semantic consistency between synthesized images and its text description.

TABLE VI

HUMAN PERCEPTUAL TEST RESULTS ↑ OF KT-GAN COMPARING WITH ATTNGAN [35] AND DM-GAN [22]. THE BOLD IS THE BEST RESULT

Method	CUB-Bird	MS-COCO
AttnGAN [35]	21.46%	19.27%
DMGAN [22]	33.74%	35.10%
KT-GAN	**44.80%**	**45.63%**

In Table VI, we compared our KT-GAN with AttnGAN [35] and DM-GAN [22] using **Human perceptual test**. After the volunteers finished the experiment, we counted the votes for each method in the two datasets. The results of subjective test shows that KT-GAN is more effective in terms of semantic consistency. These results demonstrate the superiority of KT-GAN for generating visually realistic and semantically consistent images.

Visualization. For qualitative evaluation, Figure 5 shows text-to-image synthesis examples generated by AttnGAN [35], SEGAN [13], DM-GAN [22] and KT-GAN (Ours). Observing the samples generated on the CUB-Bird dataset in the left four columns of Figure 5, images synthesized by AttnGAN [35] and SEGAN [13] are prone to semantic structure ambiguity. The quality of images synthesized by DM-GAN [22] is higher than that of AttnGAN [35] and SEGAN [13], but not as good as that of our KT-GAN. In contrast, our KT-GAN model better highlights the main object with detail, and its contrast with the background. In terms of multi-subjects image generation, e.g., the MS-COCO data (see the right four columns of Figure 5), it is more challenging to generate photo-realistic images when text descriptions are more complicated and contain multiple objects. Because KT-GAN can better bridge the domain gap between text and image, it can better capture the major objects and arrange contents

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

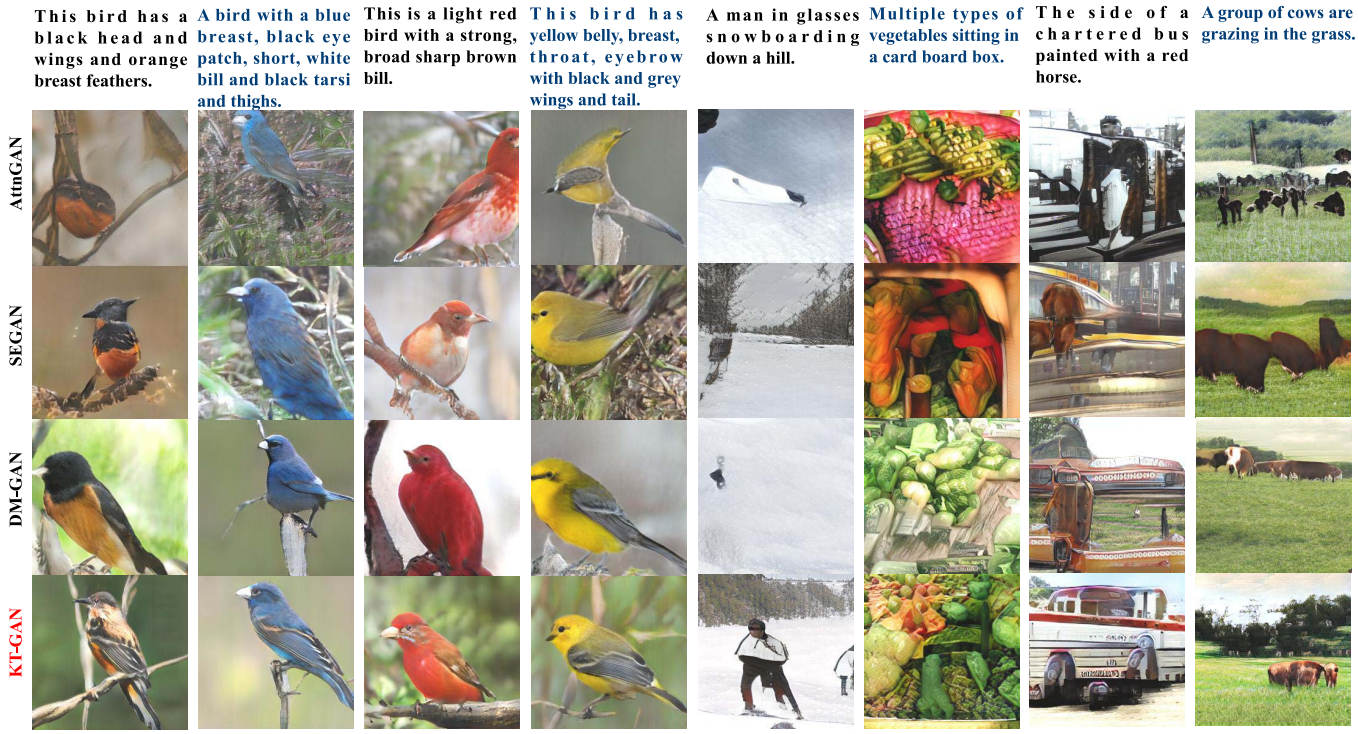


Fig. 5. Images of 256 × 256 resolution are generated by AttnGAN [35], SEGAN [13], DM-GAN [22] and KT-GAN (Ours) conditioned on text descriptions. Texts in the left four columns are from CUB-Bird [32] test datasets. Texts in the right four columns are from MS-COCO [19] test datasets.

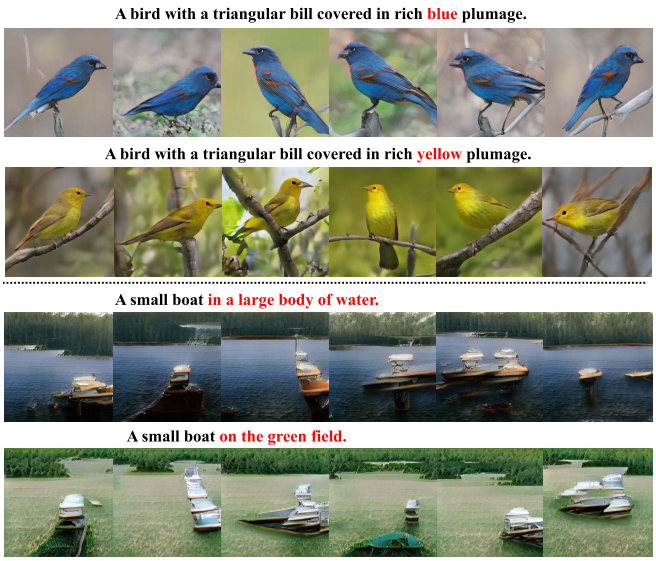


Fig. 6. Examples of KT-GAN on the ability of catching subtle changes (phrase in red) of the text descriptions on CUB-Bird (top) and MS-COCO (bottom) test sets.

in a more meaningful way. Eventually, these lead to images with better global structure. More visualizations are given in Appendix-D.

Besides, we further evaluate the sensitivity of the proposed KT-GAN by changing just one word or phrase in the input sentence. As shown in Figure 6, the synthesized images are modified according to the changes of the input sentence, e.g., bird color ("blue" versus "yellow") and image scene

Fig. 7. Failure Cases are generated by our KT-GAN on the CUB-Bird test set (top row) and on the MS-COCO test set (bottom row).

("in a large body of water" versus "on the green field"). It demonstrates that our KT-GAN has the ability to catch subtle changes of the text and retains the semantic diversities and details from text.

V. Limitation and Discussion

Although our proposed KT-GAN shows superiority in generating visually realistic and semantically consistent images, some limitations and discussion must be taken into consideration.

A. In Terms of Model Design

First, I2I task, SDM and T2I task are not jointly optimized with complete end-to-end training due to limited computational resources. Second, we use AATM to refine the word embeddings for generator, which could be further improved. In the future sudies, we can further use some sentence parser to extract the informations of objects or details.

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

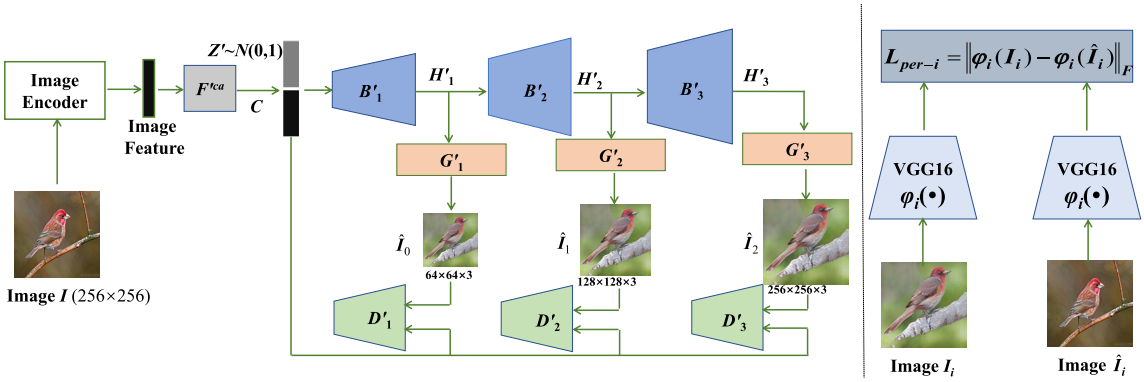


Fig. 8. The Structure of the Teacher Network. The Teacher Network includes three main module: the Image Encoder module, the Generator, and Discriminator. The right part of this figure is the structure of the perceptual loss in the Teacher Network.

B. In Terms of Image Visualization

We show some failure images synthesized by our KT-GAN on the CUB-Bird test set (the first row of Figure 7) and the MS-COCO test set (the second row of Figure 7).

On the CUB-Bird dataset (the first row of Figure 7): in the first two images, the body parts of the bird are missing; in the third and the fourth images, the bill of the bird is very strange; in the fifth image, the background is too noisy; in the sixth images, our KT-GAN creates the bird with two heads.

On the challenging COCO datasets (the second row of Figure 7): like all existing approaches, our KT-GAN also cannot effectively extract correlated structural and semantic information to support realistic synthesis; in the first two images, KT-GAN tends to place objects corresponding to specific features at many locations throughout the image; in the last four pictures, the big problem of current methods and our KT-GAN is that objects in the images can not be correctly synthesized by the generator.

We think the above issues are caused by: **(i)** The generator's generation capacity is not strong enough; so, it is necessary to design a stronger generator to better synthesize the objects or details; **(ii)** A single sentence contains very little semantics; Based on the limited semantic information, it is difficult for the generator to synthesize complex images, especially the complex scenes and objects on the MS-COCO dataset; Thus, it is necessary to explore more valid semantic information from more text descriptions to help the generator synthesize high-quality images.

VI. Conclusion

In this paper, we propose a novel Attention-Transfer Mechanism (AATM) and a Semantic Distillation Mechanism (SDM), and build a Knowledge-Transfer Generative Adversarial Network (KT-GAN) for Text-to-Image (T2I) synthesis. The SDM uses Image-to-Image synthesis to guide the T2I synthesis to better encode text feature and synthesize photographic image. The AATM helps the generator progressively identify important words and enrich the details of synthesized image. SDM and AATM successfully bridge the heterogeneous gap and help the generator synthesize high quality images. Extensive experimental results and analysis demonstrate the effectiveness of KT-GAN and significantly improved performance compared against previous most state-of-the-art methods.

Appendix A
Structure of Teacher Task (I2I Task)

In the Figure 8, the Teacher Network includes three main module: Image Encoder module, Generator, and Discriminator. In our paper, we describe that the modification of AttnGAN mainly includes two operations. **(i)** Remove the attention mechanism in the generator, use global image feature v as the input of the generator. **(ii)** Introduce the perceptual loss ($\mathcal{L}_{per}$) [4] into the I2I to improve the quality of synthesized image.

In addition the above two main modifications, compared with AttnGAN [35], the input feature to the synthesizer is image features instead of text features. The corresponding loss functions need to be simply modified. Thus, in the following, we describe the details of loss function in the Teacher Network.

Before describing the loss functions, we simply introduce the pipeline of synthesizer in Figure 8 in order to easily understand the following definition of the loss functions.

A. Terminologies for Teacher Network Pipeline

In Section III-C, we explain the student network. Here we elaborate the teacher's network. They are similar. we use symbols B', G', and H' *et al.* to represent corresonding B, G, and H *et al.* in the student network (T2I task).

The input image is transformed into global image feature $v \in \mathbb{R}^D$ by a pre-trained Inception-V3 [28] Image Encoder on ImageNet [26]. As shown in left part of Figure 8, the teacher network has m blocks ($B'_0, B'_1, \cdots, B'_{m-1}$) and the corresponding m generators ($G'_0, G'_1, \cdots, G'_{m-1}$). The generators take the hidden states ($H'_0, H'_1, \cdots, H'_{m-1}$) as input and generate images of small-to-large scales ($\hat{I}_0, \hat{I}_1, \cdots, \hat{I}_{m-1}$):

$$\begin{aligned} B'_0 : H'_0 &= F'_0(z', F'^{ca}(v)); \\ B'_i : H'_i &= F'_i(H'_{i-1} \parallel H'_{i-1}), i = 1, 2, \cdots, m-1; \\ G'_i : \hat{I}_i &= G'_i(H'_i). \end{aligned} \tag{12}$$

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

Here, $z' \sim N(0,1)$. F'^{ca} is a conditioning augmentation module [41] that converts a sentence feature v to a conditioning feature for the generator. F'^{ca}, F'_i, and G'_i are modeled as neural networks. Here, $\|$ denotes the concatenation of two maps along depth axis. **In our paper, based on the modification (i) about AttnGAN [41], we modify the** $H_i = F_i(H_{i-1}||F_i^{attn}(W_0, H_{i-1}))$ **in Eq. 1 to** $H'_i = F'_i(H'_{i-1} \| H'_{i-1})$. In the subsection A-B, we describe the details of the loss functions in I2I task (Teacher Network).

B. Generative and Discriminative Loss

Combining the above modules together, at the i-th stage of the teacher network, the Generative loss $\mathcal{L}_{G'_i}$ and Discriminative loss $\mathcal{L}_{D'_i}$ are defined as

$$\mathcal{L}_{G'_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G'_i}}[log D'_i(\hat{I}_i)]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G'_i}}[log D'_i(\hat{I}_i, v)]}_{\text{conditional loss}}, \tag{13}$$

where the unconditional loss is trained to generate images towards the true data distribution to fool the discriminator, and the conditional loss is trained to generate samples to match the real image feature embedding.

The discriminator D'_i is trained to classify the input into the class of real or fake images by minimizing the cross-entropy loss (14), as shown at the bottom of the page, where I_i is from the realistic image distribution p_{data} at the i-th scale, and $\hat{I}_i$ is from distribution $p_{G'_i}$ of the generative images at the same scale.

The final objective function of the generative network and discriminative network in the Teacher Network are defined as

$$\mathcal{L}_{G^*} = \sum_{i=0}^{m-1} \mathcal{L}_{G'_i} + \sum_{i=0}^{m-1} \eta_i \mathcal{L}_{per_i}, \mathcal{L}_{D^*} = \sum_{i=0}^{m-1} \mathcal{L}_{D'_i}. \tag{15}$$

Here, we train $\mathcal{L}_{G^*}$ and $\mathcal{L}_{D^*}$ using the entire training dataset.

APPENDIX B
NETWORK/ALGORITHM PARAMETERS

Based on experiments on a held-out validation set, we list the hyperparameters in Table VII. Due to GPU memory constraints, we did not try the KT-GAN with more than three blocks. Thus, we set the parameter $m = 3$ in this paper.

APPENDIX C
NETWORK/ALGORITHM PARAMETERS DISCUSSION

In this subsection, we mainly find the suitable values for these hyper-parameters in Table VII by a series of ablation studies.

TABLE VII
PARAMETER VALUES OF OUR KT-GAN

Hyperparameter	Value
Dimension of output in **Text and Image Encoders**	256
α β in Eq. 2	0.85, 0.9
M in Eq. 4 and Eq. 7	50
γ in Eq. 4 and Eq. 7	2
β_0 in Eq. 5	5
λ_1, λ_2 in Eq. 8	50, 50
λ_3 in Eq. 11	5
η_0, η_1, η_2 in Eq. 15	10^{-3}, 10^{-2}, 10^{-1}

TABLE VIII
INCEPTION SCORE (**IS** ↑), **FID** ↓ AND **RANK-1** ↑ OF DIFFERENT VALUES OF THE HYPER-PARAMETERS α, β ON CUB-BIRD DATASET

Method	IS	FID	Rank-1
$\alpha = 1.0$, $\beta = 1.0$	4.60 ± 0.07	21.16	29.1%
$\alpha = 0.85$, $\beta = 0.9$	4.74 ± 0.05	20.40	29.4%
$\alpha = 0.9$, $\beta = 0.85$	4.74 ± 0.02	20.33	29.4%
$\alpha = 0.5$, $\beta = 1.0$	4.64 ± 0.04	21.53	29.3%
$\alpha = 1.0$, $\beta = 0.5$	4.66 ± 0.07	20.89	29.4%
$\alpha = 0.5$, $\beta = 0.5$	4.52 ± 0.04	21.76	28.7%

A. α, β

The α, β is the important hyper-parameters in the KT-GAN. The hyper-parameters α and β belong to the AATM module. We use the AttnGAN+AATM to discuss the sensitivity of the hyper-parameters α, β. As shown in Table VIII, when the $\alpha = 0.85$, $\beta = 0.9$ or $\alpha = 0.9$, $\beta = 0.85$, the AttnGAN+AATM achieves the best performance in the three measures.

B. γ

In the SDM, we introduce the focal loss into the cross-modal knowledge distillation stage. We hope the focal loss can better balance the hard samples and easy samples in the cross-modal knowledge matching process. Here, the γ is the important hyper-parameter in the focal loss [18]. In this subsection, we mainly find the suitable value for γ in the KT-GAN. In Table IX, when the hyper-parameter $\gamma = 2.0$, the KT-GAN gains the best performance on the CUB-Bird test set over these three measures.

C. β_0

β_0 is a factor that determines how much to magnify the importance of the most relevant word-to-region context pair. When $\beta_0 \to \infty$, $R_W(v^*, s) \to max_{i=1}^{T} O_i \cdot w_0^i$. So, in this subsection, we find the suitable value for the β_0. As shown in Table X, when $\beta_0 = 1, 2, 5$, the performance of the

$$\mathcal{L}_{D'_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{I_i \sim P_{data_i}}[log D'_i(I_i)] - \frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G'_i}}[log(1 - D'_i(\hat{I}_i)]}_{\text{unconditional loss}}$$
$$+ \underbrace{-\frac{1}{2}\mathbb{E}_{I_i \sim P_{data_i}}[log D'_i(I_i, v)] - \frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G'_i}}[log(1 - D'_i(\hat{I}_i, v)]}_{\text{conditional loss}}, \tag{14}$$

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

Fig. 9. Images of 256 × 256 resolution are generated by our KT-GAN, DM-GAN [22], SEGAN [13], and AttnGAN [35] (baseline) conditioned on text descriptions from MS-COCO test datasets.

TABLE IX

IS ↑, **FID** ↓ AND **RANK-1** ↑ OF DIFFERENT VALUE OF THE HYPER-PARAMETER γ ON CUB-BIRD DATASET

Method	IS	FID	Rank-1
$\gamma = 0$	4.82 ± 0.08	17.22	31.0%
$\gamma = 0.5$	4.82 ± 0.02	17.39	31.8%
$\gamma = 1.0$	4.84 ± 0.02	17.36	32.2%
$\gamma = 2.0$	4.85 ± 0.04	17.32	32.9%
$\gamma = 5.0$	4.84 ± 0.09	17.45	32.0%

TABLE X

IS ↑, **FID** ↓ AND **RANK-1** ↑ OF DIFFERENT VALUE OF THE HYPER-PARAMETER β_0 ON CUB-BIRD DATASET

Method	IS	FID	Rank-1
$\beta_0 = 1$	4.79 ± 0.05	17.75	32.7%
$\beta_0 = 2$	4.85 ± 0.03	17.34	32.9%
$\beta_0 = 5$	4.85 ± 0.04	17.32	32.9%
$\beta_0 = 50$	4.63 ± 0.06	18.62	31.5%

KT-GAN is stable. When $\beta_0 = 2, 5$, KT-GAN achieves the best performance on the CUB-Bird data set over these three measures.

D. λ_3

In the training stage of KT-GAN, we also utilize the DAMSM loss [35] to make generated images better conditioned on text descriptions. Same as AttnGAN [35], we also set $\lambda_3 = 5$ in our KT-GAN. Besides, we further show the performance of KT-GAN based on different value of the hyper-parameter λ_3 in Table XI. When $\lambda_3 = 0, 5, 10$, the performance of KT-GAN is stable.

TABLE XI

IS ↑, **FID** ↓ AND **RANK-1** ↑ OF DIFFERENT VALUE OF THE HYPER-PARAMETER λ_3 ON CUB-BIRD DATASET

Method	IS	FID	Rank-1
$\lambda_3 = 0$	4.82 ± 0.07	17.44	32.9%
$\lambda_3 = 5$	4.85 ± 0.04	17.32	32.9%
$\lambda_3 = 10$	4.87 ± 0.06	17.29	33.1%
$\lambda_3 = 50$	4.60 ± 0.04	19.64	31.0%

TABLE XII

IS ↑, **FID** ↓ AND **RANK-1** ↑ OF DIFFERENT VALUES OF THE HYPER-PARAMETERS η_0, η_1 AND η_2 ON CUB-BIRD DATASET

Method	IS	FID	Rank-1
$\eta_0 = 0$, $\eta_1 = 0$, $\eta_2 = 0$	4.82 ± 0.02	17.84	32.0%
$\eta_0 = 10^{-3}$, $\eta_1 = 10^{-2}$, $\eta_2 = 10^{-1}$	4.85 ± 0.04	17.32	32.9%
$\eta_0 = 10^{-1}$, $\eta_1 = 10^{-2}$, $\eta_2 = 10^{-3}$	4.85 ± 0.07	17.51	32.4%
$\eta_0 = 10^{-1}$, $\eta_1 = 10^{-1}$, $\eta_2 = 10^{-1}$	4.84 ± 0.05	17.36	33.3%
$\eta_0 = 1.0$, $\eta_1 = 1.0$, $\eta_2 = 1.0$	4.72 ± 0.05	19.94	30.9%

E. η_0, η_1, η_2

η_0, η_1, η_2 is hyper-parameters in the teacher network of KT-GAN. The η_0, η_1, η_2 balance the learning weights of the three scale perceptual losses in the teacher network. Table XII shown the main results on the CUB-Bird test set. As shown in Table XII, when $\eta_0 = 10^{-3}$, $\eta_1 = 10^{-2}$, $\eta_2 = 10^{-1}$, KT-GAN achieves the best performance in the CUB-Bird dataset over three measures.

APPENDIX D
MORE VISUAL COMPARISON RESULTS

In this section, we show more visual comparison results between our KT-GAN, DM-GAN [22], SEGAN [13], and

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

Fig. 10. Images of 256 × 256 resolution are generated by our KT-GAN, DM-GAN [22], SEGAN [13], and AttnGAN [35] (baseline) conditioned on text descriptions from MS-COCO test datasets.

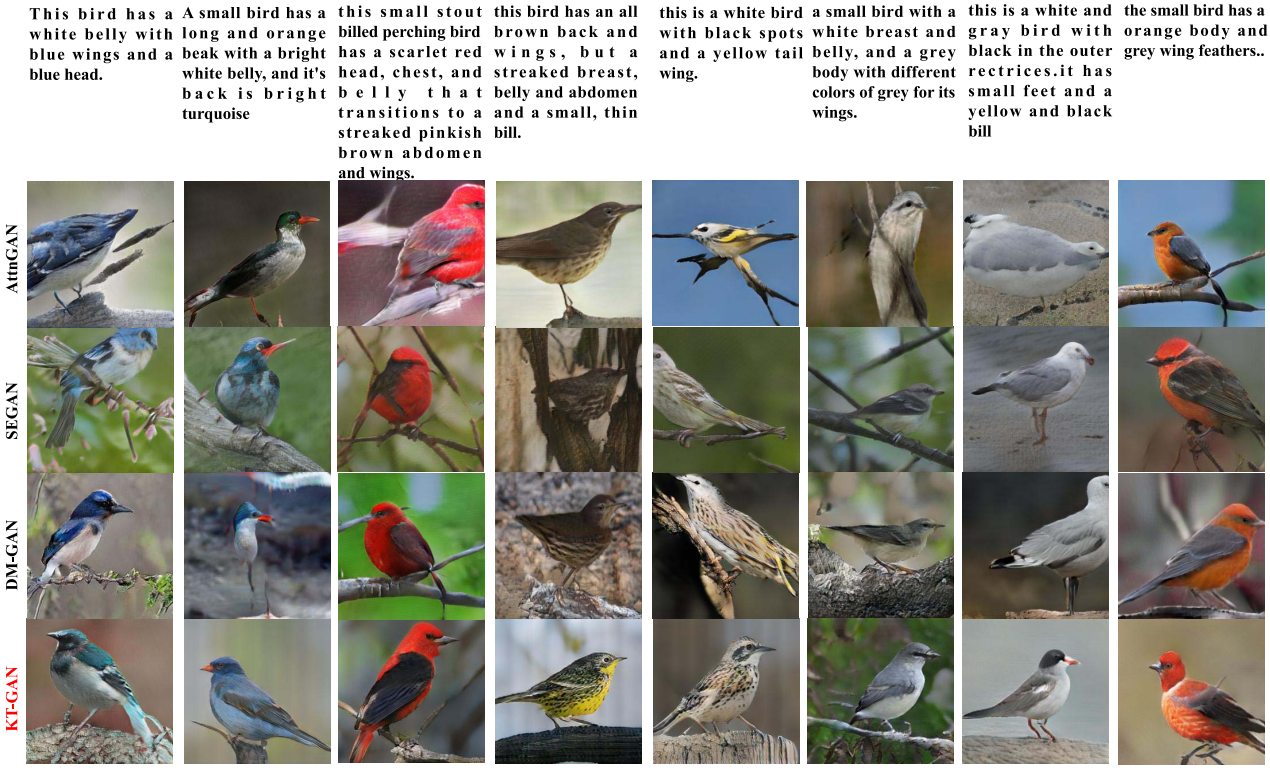


Fig. 11. Images of 256 × 256 resolution are generated by our KT-GAN, DM-GAN [22], SEGAN [13], and AttnGAN [35] (baseline) conditioned on text descriptions from CUB-Bird test datasets.

AttnGAN [35] (baseline) on the CUB-Bird and MS-COCO dataset in Figure 11, Figure 12, Figure 9, and Figure 10. These visual comparison results further demonstrate the generalization ability of the KT-GAN. Besides, we show more generated results on CUB-Bird dataset and MS-COCO dataset. As shown in Figure 13, we further

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

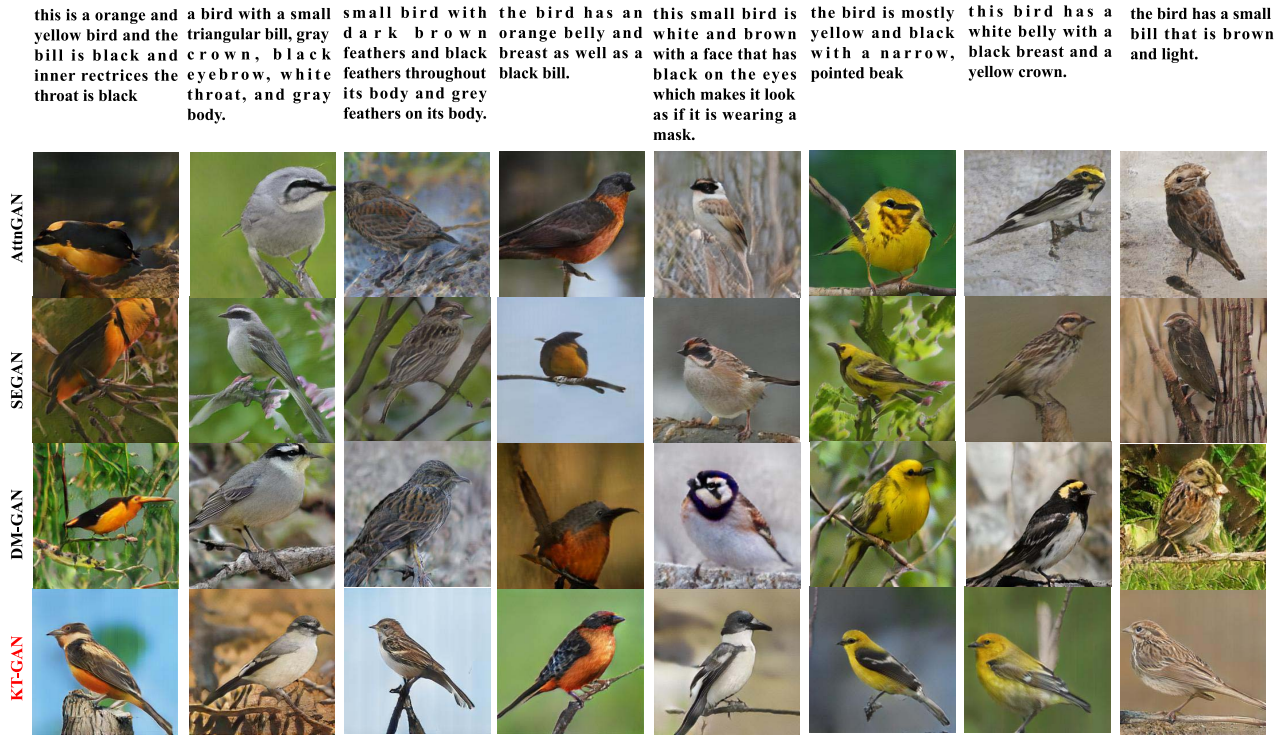
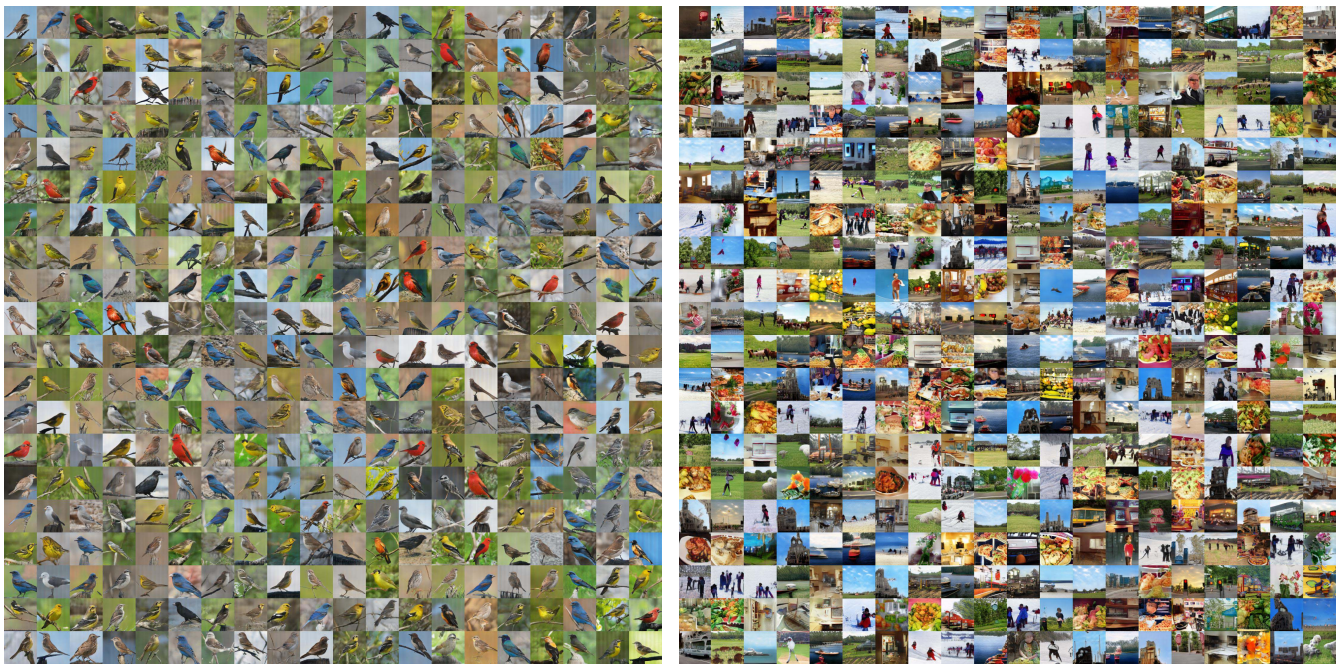


Fig. 12. Images of 256 × 256 resolution are generated by our KT-GAN, DM-GAN [22], SEGAN [13], and AttnGAN [35] (baseline) conditioned on text descriptions from CUB-Bird test datasets.



(a) More generated images of our KT-GAN from CUB-Bird test set.

(b) More generated images of our KT-GAN from MS-COCO test set.

Fig. 13. More generated images of our KT-GAN on the CUB-Bird test set and on the MS-COCO test set.

show 400 images for each dataset. Since the limited size of the Appendix, you can down these figures from https://pan.baidu.com/s/11QCfAcCfWi41B2DHiWGemA, password: qqx1, and view more details.

ACKNOWLEDGMENT

No conflict of interest: Hongchen Tan, Xiuping Liu, Meng Liu, Baocai Yin, and Xin Li declare that they have no conflict of interest.

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] A. RoyChowdhury *et al.*, "Automatic adaptation of object detectors to new domains using self-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 780–790.

[2] M. Bharti, G. Fabio, and A. Sikandar, "Knowledge distillation for end-to-end person search," in *Proc. BMVC*, 2019, pp. 1–16.

[3] L. Bowen, Q. Xiaojuan, L. Thomas, and H. S. T. Philip, "Controllable text-to-image generation," in *Proc. NeurIPS*, 2019, pp. 2065–2075.

[4] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[5] F. Faghri, J. David Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improved visual-semantic embeddings," in *Proc. BMVC*, 2018, pp. 1–9.

[6] M. Nicolás Guil Francisco Castro and J. Manuel Marín-Jiménez, "End-to-end incremental learning," in *Proc. ECCV*, Sep. 2018, pp. 233–248.

[7] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1365–1374.

[8] H. Geoffrey, V. Oriol, and D. Jeff, "Distilling the knowledge in a neural network," in *Proc. NeurIPS Workshop*, 2015, pp. 1–9.

[9] J. Ian Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[10] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2327–2336.

[11] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.

[12] 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 *Proc. NeurIPS*, 2017, pp. 6626–6637.

[13] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, "Semantics-enhanced adversarial nets for text-to-image synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10501–10510.

[14] W. Huang, Y. Xu, and I. Oppermann, "Realistic image generation using region-phrase attention," 2019, *arXiv:1902.05395*. [Online]. Available: http://arxiv.org/abs/1902.05395

[15] J. Li, K. Fu, S. Zhao, and S. Ge, "Spatiotemporal knowledge distillation for efficient estimation of aerial video saliency," *IEEE Trans. Image Process.*, vol. 29, pp. 1902–1914, 2020.

[16] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1219–1228.

[17] W. Li *et al.*, "Object-driven text-to-image synthesis via adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12174–12182.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[19] T. Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[20] M. Hu, Y. Wu, S. Zhao, H. Guo, R. Cheng, and Z. Su, "Domain-invariant feature distillation for cross-domain sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Nov. 2019, pp. 5562–5571.

[21] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong GAN: Continual learning for conditional image generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2759–2768.

[22] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5802–5810.

[23] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.

[24] L. Qicheng, H. Mohammad, P. Ahmad, D. Francis, D. J. Lisa, and F. Thomas, "Dual adversarial inference for text-to-image synthesis," in *Proc. ICCV*, Oct. 2019, pp. 7567–7576.

[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. ICML*, 2016, pp. 1060–1069.

[26] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and C. Xi, "Improved techniques for training gans," in *Proc. NeurIPS*, 2016, pp. 2234–2242.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[29] T. Yu, J. Yu, Z. Yu, and D. Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Trans. Image Process.*, vol. 29, pp. 1204–1218, 2020.

[30] Q. Tingting, Z. Jing, X. Duanqing, and T. Dacheng, "Learn, imagine and create: Text-to-image generation from prior knowledge," in *Proc. NeurIPS*, 2019, pp. 887–897.

[31] H. Tobias, H. Stefan, and W. Stefan, "Generating multiple objects at spatially distinct locations," in *Proc. ICLR*, 2019, pp. 1–23.

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Comput. Neural Syst., Tech. Rep. CNS-TR-2011-001.

[33] W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, and Q. Tian, "A multi-scale spatial-temporal attention model for person re-identification in videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3365–3373, 2020.

[34] X. Han, X. Song, Y. Yao, X.-S. Xu, and L. Nie, "Neural compatibility modeling with probabilistic knowledge distillation," *IEEE Trans. Image Process.*, vol. 29, pp. 871–882, 2020.

[35] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[36] L. Yang, L. Jing, and M. K. Ng, "Robust and non-negative collective matrix factorization for Text-to-Image transfer learning," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4701–4714, Dec. 2015.

[37] L. Yitong, M. M. Renqiang, S. Dinghan, C. David, and C. Lawrence, "Video generation from text," in *Proc. AAAI*, 2018, p. 5.

[38] Y. Li *et al.*, "StoryGAN: A sequential conditional GAN for story visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6329–6338.

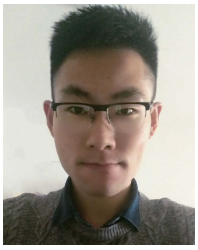
[39] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[40] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention CoupleNet: Fully convolutional attention coupling network for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2019.

[41] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[42] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.

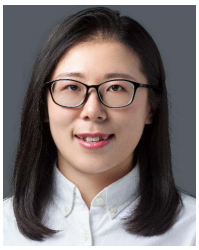
[43] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2347–2356.



Hongchen Tan is currently pursuing the Ph.D. degree in mathematical sciences with the Dalian University of Technology. His research interests are person re-identification and text-to-image synthesis. Various parts of his work have been published in top forums and journals, such as TIP, ICCV, and Neurocomputing.

Xiuping Liu received the Ph.D. degree in computational mathematics from the Dalian University of Technology, China. She is currently a Professor with the School of Mathematical Sciences, Dalian University of Technology. Her research interests include shape modeling and analyzing, and computer vision.

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.



Meng Liu (Member, IEEE) received the M.S. degree in computational mathematics from the Dalian University of Technology, China, in 2016. She is currently a Professor with the School of Computer Science and Technology, Shandong Jianzhu University. Her research interests include multimedia computing and information retrieval. Various parts of her work have been published in top forums and journals, such as SIGIR, MM, and IEEE TIP. She has served as a Reviewer and Subreviewer for various conferences and journals, such as MMM 2018, ACM MM 2018/2019, JVCI, and INS.

Baocai Yin received the B.S. and Ph.D. degrees in computer science from the Dalian University of Technology, Dalian, China. He is currently a Professor of computer science with the Dalian University of Technology and the Dean of the Faculty of Electronic Information and Electrical Engineering. His research interests include digital multimedia and computer vision.

Xin Li (Senior Member, IEEE) received the B.E. degree in computer science from the University of Science and Technology of China, in 2003, and the M.S. and Ph.D. degrees in computer science from The State University of New York at Stony Brook, in 2005 and 2008, respectively. He is currently a Professor with the Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA, USA. His research interests are in geometric and visual data computing, processing, and understanding, computer vision, and virtual reality.

Authorized licensed use limited to: Louisiana State University. Downloaded on March 25,2022 at 14:25:34 UTC from IEEE Xplore. Restrictions apply.