DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis

Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, *Senior Member, IEEE,* Fei Wu, and Xiao-Yuan Jing, *Member, IEEE,*

***Abstract*—Synthesizing high-resolution realistic images from text descriptions is a challenging task. Almost all existing text-to-image methods employ stacked generative adversarial networks as the backbone, utilize cross-modal attention mechanisms to fuse text and image features, and use extra networks to ensure text-image semantic consistency. The existing text-to-image models have three problems: 1) For the backbone, there are multiple generators and discriminators stacked for generating different scales of images making the training process slow and inefficient. 2) For semantic consistency, the existing models employ extra networks to ensure the semantic consistency increasing the training complexity and bringing an additional computational cost. 3) For the text-image feature fusion method, cross-modal attention is only applied a few times during the generation process due to its computational cost impeding fusing the text and image features deeply. To solve these limitations, we propose 1) a novel simplified text-to-image backbone which is able to synthesize high-quality images directly by one pair of generator and discriminator, 2) a novel regularization method called Matching-Aware zero-centered Gradient Penalty which promotes the generator to synthesize more realistic and text-image semantic consistent images without introducing extra networks, 3) a novel fusion module called Deep Text-Image Fusion Block which can exploit the semantics of text descriptions effectively and fuse text and image features deeply during the generation process. Compared with the previous text-to-image models, our DF-GAN is simpler and more efficient and achieves better performance. Extensive experiments and ablation studies on both Caltech-UCSD Birds 200 and COCO datasets demonstrate the superiority of the proposed model in comparison to state-of-the-art models.**

***Index Terms*—Text-to-image synthesis, generative adversarial network, cross-modal, text-image semantic understanding.**

I. Introduction

THE last few years have witnessed the great success of Generative Adversarial Networks (GANs) [1] for a variety of applications.

Among them, text-to-image synthesis is one of the important applications of GANs. It aims to generate realistic and text-consistent images from given natural language descriptions. Due to its practical applications, text-to-image synthesis has become an active research area recently [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. Most recently proposed text-to-image synthesis models are based on Stacked Generative Adversarial Networks [7] to generate high-resolution images. As shown in Figure 1.(a), there are multiple generators and discriminators stacked in their networks. By conditioning on the low-resolution initial images generated by G_0 and text descriptions, the following generators refine the initial images to high-resolution ones.

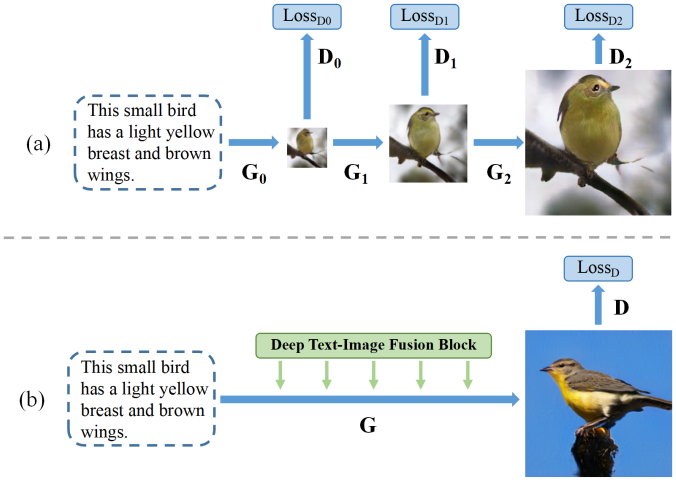

Fig. 1. An illustrative comparison between existing models and our DF-GAN: (a) existing text-to-image models stack multiple generators and discriminators to generate high-resolution images which makes final results look like a simple combination of visual features from different image scales. (b) Our DF-GAN generates high-resolution images directly by Wasserstein distance, and fuses the text information into visual feature maps by Deep text-image fusion Blocks. Compared with the final results generated by stacked GAN, the image generated by DF-GAN is more realistic.

Although impressive results have been presented by stacked text-to-image GANs, there remain two problems for stacked generators and discriminators, respectively. First, the initial images have a great influence on the final refined results, if the initial image is not well synthesized, the following generators can hardly refine it to a satisfactory quality. Second, different discriminators correspond to different image scales, each discriminator predicts a discriminator loss to evaluate the input image. The Stacked GANs family [7], [10], [8], [9], [12], [11], [13], [14], [15], [16], [17], [18] makes a summation of all discriminator losses as the whole discriminator loss for training end to end. But it is hard to balance all discriminator losses while training. This uncertainty makes the final refined images look like a simple combination of fuzzy shape and some details. Figure 1(a) shows the problem with the image

M. Tao, S. Wu and F. Wu are with the School of Automation, Nanjing University of Posts and Telecommunications, 210023 Nanjing, China. (e-mail:mingtao2000@126.com; sswuai@126.com; wufei_8888@njupt.edu.cn)

H. Tang and N. Sebe are with the Department of Information Engineering and Computer Science (DISI), University of Trento, Trento 38123, Italy. (e-mail: hao.tang@unitn.it; sebe@disi.unitn.it)

X. Jing is with the School of Computer, Wuhan University, 430072 Wuhan, China. (e-mail: jingxy_2000@126.com)

M. Tao and H. Tang contributed equally to this work.

generated by Stacked GANs. This image has a coarse shape synthesized by G_0 and G_1 and some details synthesized by G_2. It makes the synthetic images unrealistic.

Another shortcoming of current text-to-image models is the method of fusing text and image information during generation. There are three typical text-image fusion methods: concatenating [2], [7], [8], cross-modal attention [10], [14] and Conditional Batch Normalization [15]. Early researches [2], [7], [8] in text-to-image GANs typically provide the text information to the generator by naively concatenating the sentence vector to the input noise and some intermediate visual feature maps. Naively concatenating cannot make full use of text information. Recently researches compute the word-context features through cross-modal attention mechanisms and concatenate it to intermediate feature maps such as AttnGAN [10]. Cross-modal attention can help the generator synthesize fine-grained details of the image. But it has two problems. First, it is a variant of spatial attention, its computational cost grows rapidly as the image size increases. Second, this cross-modal spatial attention only tries to find the relationship between each pixel and text description. But the human language conveys strong high-level semantics, while one pixel in an image with tens of thousands of pixels is relatively low-level [19], [20]. So it cannot deal with high-level semantics well during the image generation process which brings difficulties to synthesize complex images with multiple objects. SD-GAN [15] introduces Conditional Batch Normalization (CBN) [21], [22] to reinforce the text-related parts in the feature maps. But it does not decompose the effectiveness of Affine Transformation from CBN, and the CBN is only applied a few times during the image generation process. As a summation, these text-image fusion methods cannot fuse the text information into visual feature maps deeply and efficiently.

Furthermore, current text-to-image models always employ DAMSM loss [10], cycle consistency [13] or Siamese network [15] to ensure the text-image semantic consistency. These operations require an extra network to compute the text-image semantic similarity during training which makes the stacked architecture more complicated, increases training complexity, and requires more computational resources and training time.

To address these issues, we propose a novel text-to-image generation method named as Deep Fusion Generative Adversarial Network (DF-GAN). For the first issue, we propose a novel simplified text-to-image backbone. We employ the ResNet [23] to redesign the architecture of the generator and discriminator. Moreover, inspired by SAGAN [24], we introduce the hinge version of the adversarial loss to stabilize the training process.

For the second issue, we propose a Deep text-image Fusion Block (DFBlock) to effectively fuse the text information into visual feature maps during the generation process. The DFBlock consists of several lightweight Affine Transformations. The Affine Transformation manipulates the visual feature maps through channel-wise scaling and shifting. By stacking multiple Affine Transformations and ReLU layers in DFBlock, the text information can be fully fused into visual feature maps many times. Compared with previous methods [7], [10], [15], our DFBlock decomposes the effectiveness of Affine Transformation from CBN and deepens the text-image fusion process which enhances the semantic consistency between generated images and given text descriptions.

For the third issue, we propose a novel Matching-Aware zero-centred Gradient Penalty (MA-GP) to ensure text-image semantic consistency. MA-GP is a regularization strategy on discriminator. By pushing the real and matching inputs to the place where the gradients of the discriminator loss function are zero, the generator will be promoted to synthesize more realistic and text-image semantic consistent images. Our MA-GP proves that a properly designed regularization on the discriminator also can improve the ability of conditional image generation. Compared with previous methods, our MA-GP does not introduce an extra network to compute the text-image semantic similarity. So it does not increase the complexity and training parameters of the text-to-image backbone. Our MA-GP is a simple and effective method to ensure the text-image semantic consistency which shows excellent performance in our experiments.

Moreover, we pointed out that the previous two-ways discriminator is harmful to text-image semantic consistency and slows the convergence process of the generator. To solve this problem, we propose one-way discriminator. As a complement of Matching-Aware zero-centred Gradient Penalty (MA-GP), the one-way discriminator further promotes the effectiveness of MA-GP and accelerates the convergence of the generator.

Extensive experiments are conducted to evaluate the DF-GAN on two challenging datasets, i.e., the CUB-200 bird dataset [25] and the COCO dataset [26]. Moreover, we use Inception Score (IS) [27] and Fréchet Inception Distance (FID) [28] to measure the quality of generated images. The experimental results illustrate that the proposed DF-GAN outperforms state-of-the-art text-to-image synthesis methods. Specifically, our DF-GAN improves the IS from 4.75 to 4.86 on the CUB dataset, and decreases the FID from 32.64 to 28.92 on the COCO dataset.

Overall, the contributions of our paper are summarized as follows:

- We propose a novel simplified text-to-image backbone which can directly generate high-resolution images from text descriptions by one pair of generator and discriminator.
- A novel DFBlock is proposed which fuses text and image features more effectively and deeply. Armed with DFBlock, the generator achieves higher performance in text-to-image generation.
- We propose a novel Matching-Aware zero-centered Gradient Penalty (MA-GP) which significantly improves the image quality and text-image consistency.
- As a complement of MA-GP, we employ the one-way discriminator which promotes the effectiveness of MA-GP and helps the generator converge faster.
- The experimental results on two challenging datasets prove that the proposed DF-GAN outperforms previous state-of-the-art text-to-image models.

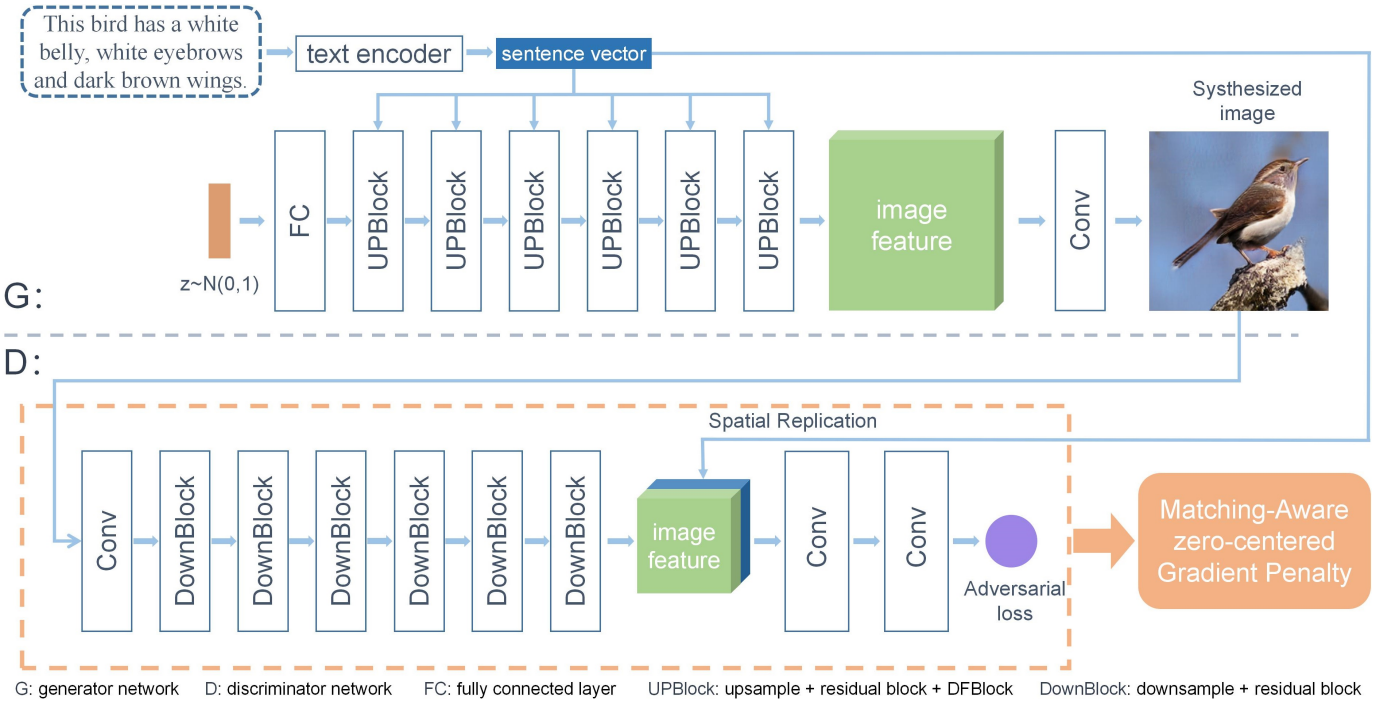

Fig. 2. The architecture of DF-GAN proposed for text-to-image synthesis. Our DF-GAN generates high-resolution images directly by one pair of generator and discriminator.

II. Related Work

A. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [1] are an attractive framework that can be used to mimic complex real-world distributions by solving a min-max optimization problem between generator and discriminator. The generator intends to synthesize visually plausible images to fool the discriminator, while the discriminator attempts to distinguish the synthetic images from real images. With proper adversarial training, the generator can synthesize high-quality images finally [29], [30], [24], [31], [32], [33], [34], [35], [36], [37].

While a lot of successes achieved, GANs are known to be hard to train. A lot of works are proposed to stabilize the training process of GANs [27], [38]. For example, Arjovsky et al. introduce the Wasserstein distance to the GAN framework [39], which solved the gradient vanishing problem in original GAN [1]. It enables Wasserstein GANs (WGANs) [29], [24], [40], [41], [42] can generate better results compared to the original GAN. However, K-Lipschitz is a hard satisfying constraint for WGANs [43], [40]. For solving this problem, some researches propose the Gradient Penalty [44], [45], [46], [35] and Spectral Normalization [40] which frees WGANs from this challenging constraint in unconditional generation tasks.

B. Text-to-Image generation

Reed *et al.* first apply the conditional GAN to generate plausible images from text descriptions [2]. Next, StackGAN [7], [8] generates high-resolution images by stacking multiple generators and discriminators and provides the text information to the generator by naively concatenating text vectors to the input noises and intermediate features. Almost existing text-to-image models are based on StackGAN. Besides, AttnGAN [10] introduces the cross-modal attention mechanism to help the generator synthesize images with more details. MirrorGAN [13] regenerates text descriptions from generated images for text-image semantic consistency [47]. SD-GAN [15] employs the Siamese structure [48], [49] to distill the semantic commons from texts for image generation consistency. Obj-GAN [16] generates complex scenes from pre-generated semantic layouts and text descriptions and employs a Fast R-CNN [50] to compute an object-wise loss. Different from previous text-to-image models, Obj-GANs introduce the bounding box and class label information to help the image generation process. Since Obj-GAN introduces other information besides text description, we did not compare it within our experiment. Finally, DM-GAN [14] introduces the Memory Network [51], [52] to cope with the first aforementioned problem of stacked architecture. But memory network only alleviates the dependence on initial images, it does not solve the problem completely. And adding Memory Network on stacked architecture makes the network require a considerable computational cost.

Our proposed method is much different from previous models. First, our DF-GAN generates high-resolution images directly by one pair of generator and discriminator. Second, our model fuses text and image features more deeply and efficiently through a sequence of DFBlocks. Third, a novel Matching-Aware zero-centered Gradient Penalty (MA-GP) is applied to discriminator to ensure image quality and text-

image semantic consistency. Finally, the one-way discriminator is employed to promote the effectiveness of MA-GP. Compared with previous text-to-image GANs, our DF-GAN is simpler and more effective. The experimental results and ablation studies also demonstrate the superiority of our architecture.

III. Deep Fusion GAN (DF-GAN)

In this paper, we propose a novel cross-modal Generative Adversarial Network named as Deep Fusion GAN (DF-GAN) for text-to-image generation (see Figure 2). We start to present DF-GAN in this section. First, we introduce the overall structure of DF-GAN and also describe the components of the generator and discriminator. Secondly, we illustrate the proposed Matching-Aware zero-centered Gradient Penalty (MA-GP), one-way discriminator and, Deep text-image Fusion Block (DFBlock) in detail.

A. Model Overview

The goal of our proposed DF-GAN is to generate realistic and text-image semantic consistent images from given text descriptions. The entire network is composed of a generator, a discriminator, and pre-trained text encoder [10]. In the following, we describe the structure of the generator and discriminator, respectively.

The generator has two inputs, a sentence vector that is encoded by text encoder and a noise vector sampled from the Gaussian distribution to ensure the diversity of generated images. The noise vector is first fed into a fully connected layer and the output is reshaped to (-1,4,4). We then apply a series of UPBlocks to upsample the image features. The UPBlock is composed of upsample layers, a residual block, and DFBlocks (see Figure 7(e)) to fuse the text and image features during the image generation process. Finally, a convolution layer converts image features into images.

The discriminator is composed of some DownBlocks and convolution layers. First, the discriminator converts images into feature maps and the output is downsampled by a series of DownBlocks. Then the sentence vector will be replicated and concatenated on the image feature. An adversarial loss will be predicted to evaluate the visual realism and semantic consistency of inputs. By distinguishing generated images from real samples, the discriminator promotes the generator to synthesize images with higher quality and text-image semantic consistency.

B. Simplified Text-to-Image Backbone

Previous text-to-image GANs are based on stacked original GANs [1] to generate high-resolution images. Although stacking architecture enables the original GAN to generate high-resolution images, it also makes the generated image look like a simple combination of coarse shape synthesized by G_0 and G_1 and some details synthesized by G_2. It makes the synthetic images unrealistic.

Inspired by recent models proposed for unconditional image generation [38], [24], we propose a novel simplified text-to-image backbone which can synthesize high-resolution images directly by one pair of generator and discriminator. Our proposed simplified text-to-image backbone can even achieve better performance than most well-designed stacked GANs. We employ the hinge loss [38] to stabilize the training process. The formulation of our method is as follows:

$$\begin{aligned} L_D = & -\mathbb{E}_{x\sim\mathbb{P}_r}[min(0,-1+D(x,e))] \\ & -(1/2)\mathbb{E}_{G(z)\sim\mathbb{P}_g}[min(0,-1-D(G(z),e))] \\ & -(1/2)\mathbb{E}_{x\sim\mathbb{P}_{mis}}[min(0,-1-D(x,e))] \\ L_G = & -\mathbb{E}_{G(z)\sim\mathbb{P}_g}D(G(z),e) \end{aligned} \tag{1}$$

where z is the noise vector sampled from Gaussian distribution. e is the sentence vector. $\mathbb{P}_g$, $\mathbb{P}_r$, $\mathbb{P}_{mis}$ denotes the synthetic data distribution, real data distribution and mismatching data distribution, respectively.

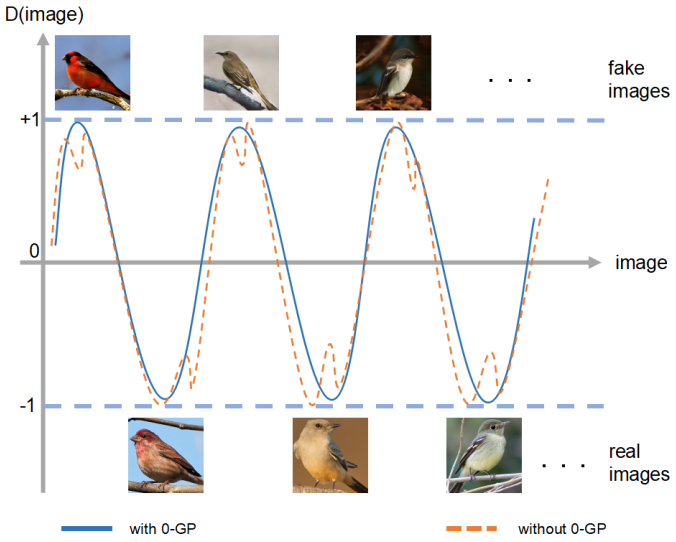

Fig. 3. A comparison of loss function surfaces before and after applying zero-centered gradient penalty(0-GP). 0-GP smooths the discriminator loss surface which is helpful for generator convergence.

C. Matching-Aware Zero-Centered Gradient Penalty

In this work, we also propose a novel conditional Matching-Aware zero-centered Gradient Penalty (MA-GP) to enable the generator to synthesize more realistic and text-image semantic-consistent images. In this subsection, we first show the unconditional zero-centered gradient penalty (0-GP) [46] on real data points from a novel and clear perspective, then extend it to our text-to-image generation task. The unconditional zero-centered gradient penalty on real data is formulated as:

$$L_{0-GP} = k \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[\|\nabla_x D(x)\|^p, \tag{2}$$

where k and p are two hyper-parameters to balance the effectiveness of gradient penalty, P_r are the real data distribution.

As shown in Figure 3, real images correspond to a low discriminator loss and synthetic images correspond to a high discriminator loss. The hinge loss limits the range of discriminator loss between -1 and 1. The zero-centered gradient penalty on real data will reduce the gradient of the real data

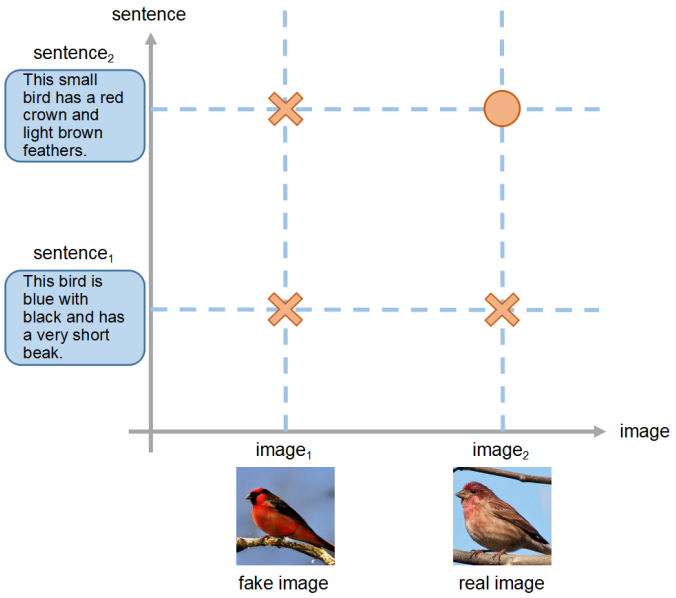

Fig. 4. A diagram for Matching-Aware zero-centered Gradient Penalty (MA-GP). The data point marked by a circle should be applied MA-GP.

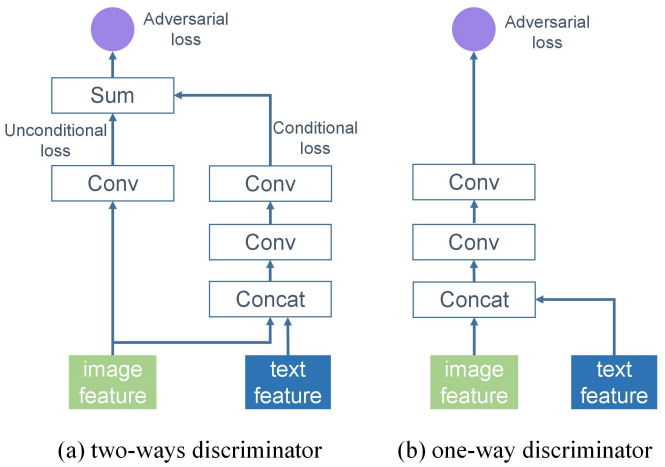

Fig. 5. Comparison between two-ways discriminator and our proposed one-way discriminator.

point and push it to the minimum point of the loss function surface. It makes the loss function surface of the real data point and its vicinity smooth which is helpful for the synthetic data point to converge to the real data point. From this perspective, the generator can synthesize more realistic images through a zero-centered gradient penalty on real data.

From the above analysis, we find that the discriminator should do two things to ensure the quality of synthetic data. First, it should put the real data at the minimum point and put synthetic data at a high point. Second, it should ensure that the loss surface of the real data point and its vicinity are smooth to help the generator converge.

Move the view from unconditional generation to text-to-image generation. Take text-to-image synthesis as an example in Figure 4, the discriminator observes four kinds of inputs: synthetic images with matching text, synthetic images with mismatched text, real images with matching text, real images with mismatched text. To generate text-matching and realistic images from given text descriptions, we should put real and matching data points to the minimum point and put other inputs at high points, and ensure a smooth vicinity of real and matching data points to help the generator converge to the minimum point. Therefore, we propose the Matching-Aware zero-centered Gradient Penalty (MA-GP) which applies the zero-centered gradient penalty on real images with the matching sentences. The whole formulation of our model is as follows:

$$\begin{aligned} L_D = & -\mathbb{E}_{x\sim\mathbb{P}_r}[min(0,-1+D(x,e))] \\ & -(1/2)\mathbb{E}_{G(z)\sim\mathbb{P}_g}[min(0,-1-D(G(z),e))] \\ & -(1/2)\mathbb{E}_{x\sim\mathbb{P}_{mis}}[min(0,-1-D(x,e))] \\ & +k\mathbb{E}_{x\sim\mathbb{P}_r}[(\|\nabla_x D(x,e)\|+\|\nabla_e D(x,e)\|)^p] \\ L_G = & -\mathbb{E}_{G(z)\sim\mathbb{P}_g} D(G(z),e) \end{aligned} \tag{3}$$

where k and p are two hyper parameters. We set the $k=2$ and $p=6$ in our network.

Compared with previous methods for ensuring the image quality and semantic consistency, our proposed Matching-Aware zero-centered Gradient Penalty (MA-GP) does not employ extra networks to compute the text-image semantic similarity. We consider that the discriminator itself is a sufficiently strong network. The addition of extra networks is not necessary. But it is very important to construct a proper discriminator loss function surface to meet expectations. In the text-to-image generation task, we hope the generator synthesizes real and text-image semantic consistent images. And our proposed MA-GP can ensure the real and matching data points are at the minimum points of the discriminator loss function surface and the vicinity of real and matching data points is smooth. It enables the generator to synthesize more realistic and text-image semantic consistent images.

D. One-way discriminator

In previous text-to-image GANs[7], [8], [10], the discriminator first extracts the image feature through a series of downsampling operations. Then the image feature will be used in two ways. As shown in Figure 5, one way determines whether the image is real or fake, another way concatenates the image feature and sentence vector to evaluate text-image consistency. So there are two kinds of loss computed, the unconditional loss and the conditional loss.

We found that the two-ways discriminator slows the convergence of the generator network and weakens the effectiveness of MA-GP. As depicted in Figure 4, the discriminator should put wrong data at high points and right data at minimum points. And the MA-GP ensures a smooth loss surface to help the generator converge to the expected data (minimum points) through gradient descending. But the two-ways discriminator decomposes the adversarial loss into conditional loss and unconditional loss. The conditional loss gives a gradient γ pointing to the real and matching inputs after backpropagation. But the unconditional loss gives a gradient β only pointing to

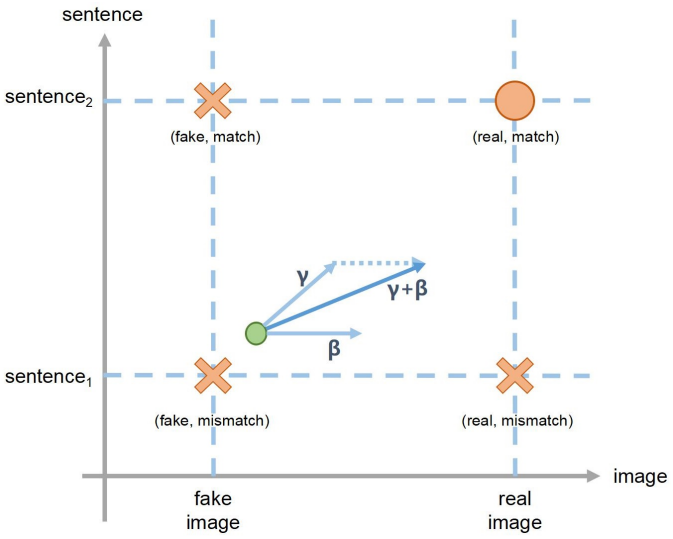

Fig. 6. A diagram shows the problem of two-ways discriminator.

the real images. As shown in Figure 6, the final gradient $\gamma+\beta$ does not point to the real and matching data points directly. This will make the convergence of the generator deviate from the right data point(real and match) slightly. Although the two-ways discriminator can also make the result of the generator approach the right point gradually, the generator needs more iterations due to the deviation during convergence.

Therefor, we propose the one-way discriminator for text-to-image synthesis. As shown in Figure 5(b), our discriminator concatenates the image feature and sentence vector, then outputs one adversarial loss through two convolution layers. The one-way discriminator only gives one gradient pointing to the real and match data points, it gives a more clear convergence goal to the generator. Armed with MA-GP, the one-way discriminator will guide the generator to synthesize more realistic images with better text-image semantic consistency. Our experiments and ablation studies also demonstrate that the one-way discriminator can further promote the effectiveness of MA-GP and accelerate the convergence process of the generator.

E. Deep Text-Image Fusion Block

We then propose a novel Deep text-image Fusion Block (DFBlock) (see Figure 7(e)) which fuses text and visual information more effectively and deeply during the generation process. The DFBlock is composed of a series of Affine Transformations, ReLU layers, and convolution layers. In the following, we introduce Affine Transformation firstly and then show the architecture of Deep Fusion Block (DFBlock).

The Affine Transformation on feature maps is widely used in normalization methods. In Batch Normalization (BN) [53], the feature map is normalized firstly, then the Affine Transformation is applied as denormalization, but the affine parameters in BN is unconditional. They are just trainable parameters learned through gradient descent. However, Conditional Batch Normalization (CBN) [21] employs additional network to predict affine parameters from conditions, it utilizes the information in conditions to modulate the information in feature maps as shown in Figure 7(b). CBN has been widely used in recently proposed conditional image generation models [22], [15], [29].

In our work, we decompose the effectiveness of Affine Transformation from CBN. We consider that it is not essential to normalize the feature maps during the conditional image generation process. In fact, we find that normalization even slightly reduces the efficiency of the text-image fusion process. The Figure 7(a) shows a typical upsample Residual Block in generator [46], [29], there are two Fusion Blocks stacked in Residual Block. There are many implementations of Fusion Block, the Fusion Block with CBN is implemented as Figure 7(b) [22], [29]. BN normalizes the feature maps and the Affine Transformation manipulates the output by scaling and shifting parameters predicted from conditions. We consider that BN in Fusion Block2 will reduce the effectiveness of Affine Transformation in Fusion Block1. Since BN transforms the feature maps into a normal distribution, it can be regarded as a reverse operation of Affine Transformation and reduces the distance between each feature map in a batch. It is not beneficial for the conditional generation process.

So we extract the Affine Transformation from CBN. We only employ Affine Transformation to manipulates visual feature maps conditioned on natural language descriptions (see Figure 7). As shown in Figure 8, we adopt two one-hidden-layer MLPs to predict the language-conditioned channel-wise scaling parameters $\boldsymbol{\gamma}$ and shifting parameters $\boldsymbol{\beta}$ from sentence vector $\boldsymbol{e}$, respectively:

$$\boldsymbol{\gamma} = MLP_1(\boldsymbol{e}), \qquad \boldsymbol{\beta} = MLP_2(\boldsymbol{e}). \tag{4}$$

If the input is feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, the output of MLP will be a vector of size C. We first conduct the channel-wise scaling operation on X with the scaling parameter $\boldsymbol{\gamma}$, then we apply the channel-wise shifting operation on X with the shifting parameter $\boldsymbol{\beta}$. This process can be formally expressed as follows:

$$AFF(\boldsymbol{x_i}|\boldsymbol{e}) = \gamma_i \cdot \boldsymbol{x_i} + \beta_i, \tag{5}$$

where AFF is Affine Transformation, $\boldsymbol{x_i}$ is the i^{th} channel of visual feature maps, $\boldsymbol{e}$ is the sentence vector, γ_i and β_i is the scaling parameter and shifting parameter for the i^{th} channel of visual feature maps. Through channel-wise scaling and shifting, the generator can capture the semantic information in text description and synthesize realistic images matching with given text descriptions. We note this module which fuses text and image features through one Affine Transformation as AFFBlock (see Figure 7(d)).

Furthermore, after decomposing the effectiveness of Affine Transformation from CBN, we can fuse image features and conditions in a more free way. So we propose the Deep text-image Fusion Block (DFBlock) which stacks multiple Affine Transformations and ReLU layers in Fusion Block. As shown in Figure 7(e), there are two Affine Transformations and ReLU layers stacked sequentially in one Fusion Block. The DFBlock deepens the depth of the text-image fusion process. For neural

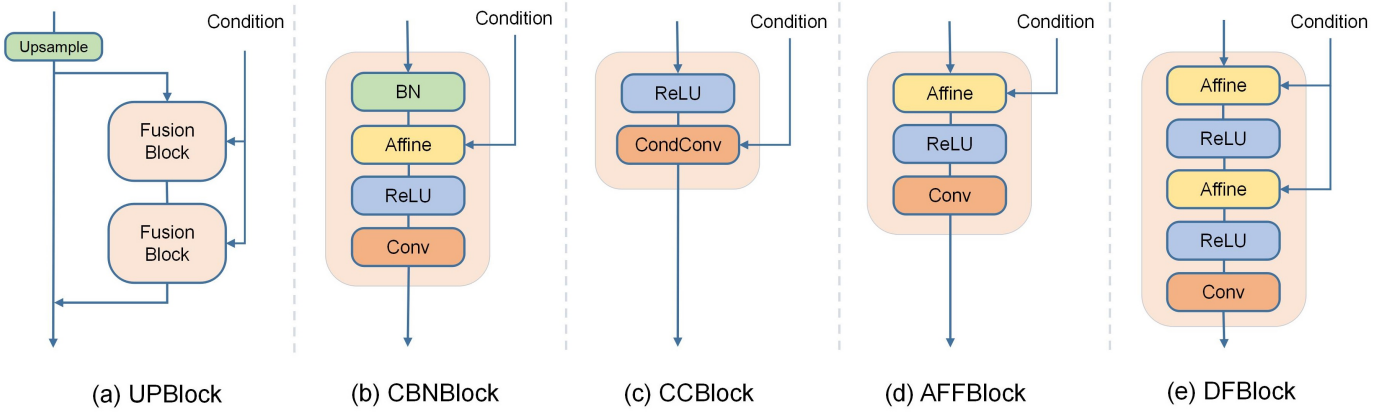

Fig. 7. We redesign the architecture of the Fusion Block and compare our proposed AFFBlock and DFBlock with CBNBlock and CCBlock. (a) A typical UPBlock in the generator network. The UPBlock upsamples the image features and fuses text and image features by two Fusion Blocks. (b) The CBNBlock is a Fusion Block which employs the Conditional Batch Normalization to fuse text and image features. (c) The CCBlock employs the Conditional Convolution layer which modulates the kernel parameters according to conditions in the convolution network. (d) AFFBlock is a simplified version of CBNBlock which removes the Batch Normalization layer. (e) The DFBlock is an enhanced version of AFFBlock, it deepens the text-image fusion process by stacking multiple affine transformations.

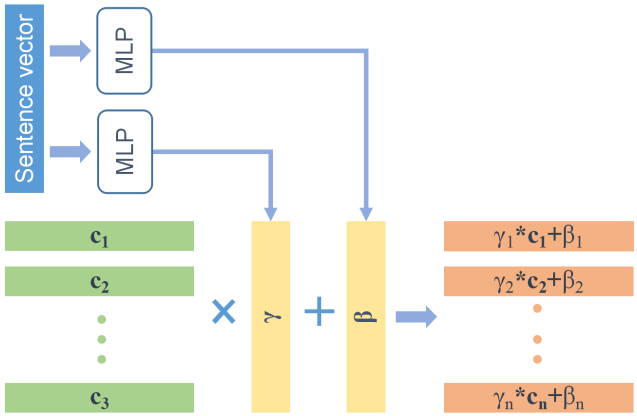

Fig. 8. Illustration of the Affine Transformation.

networks, a deeper network always means a stronger ability. We consider that deepening the fusion process can bring three main benefits for text-to-image generation: First, it gives the generator more chances to fuse text and image features, so that the text information can be fully exploited. Second, deepening the fusion process makes the fusion network have more nonlinearities, which is beneficial to generate semantic consistent images from different text descriptions. Third, stacking multiple Affine Transformations can achieve a more complex and effective fusion process.

Our DFBlock is partly inspired by Conditional Batch Normalization (CBN) [21], [22], but has two main differences: 1) our DFBlock does not normalize the feature map [53] before the scale-and-shift operation. Since Batch Normalization normalizes the features by the mean and variance computed within a batch, this operation reduces differences between image features during generation and decreases the efficiency of the previous Affine Transformations damaging the text-image consistency. So we remove it in our module. 2) we stack multiple Affine Transformations and ReLU layers in DFBlock which deepens the text-image fusion process and significantly improves the ability to fuse the text and image features.

We conduct extensive experiments to compare the effectiveness between CBNBlock (Figure 7(b)) [22], AFFBlock (Figure 7(d)) and DFBlock (Figure 7(e)) in the ablation studies. Moreover, we also compared the recently proposed conditional convolution layer (Figure 7(c)) [54] which manipulates the convolution kernel parameters according to conditions. Experimental results show that our DFBlock is more effective to fuse the text and image features, which proves its superiority and efficiency.

IV. EXPERIMENTS

In this section, we first introduce the datasets, training details, and evaluation metrics used in our experiments, then evaluate the DF-GAN and its variants quantitatively and qualitatively.

Datasets. We evaluate the proposed model on two challenging datasets, i.e., CUB bird [25] and COCO [26]. The CUB bird dataset is commonly used in text-to-image generation, which contains 11788 images belonging to 200 bird species. Each bird image has 10 language descriptions. According to previous works [10], [14], we split 150 bird species with 8855 images as the training set and 50 bird species with 2933 images as the test set. The COCO dataset contains 80k images for training and 40k images for testing. Each image in COCO has 5 language descriptions. Compared with the CUB dataset, the images in COCO are more complex, which makes it more challenging for text-to-image generation tasks.

Training Details. We optimize our network using Adam [55] with $\beta_1 = 0.0$ and $\beta_2 = 0.9$. The learning rate is set to 0.0001 for generator and 0.0004 for discriminator according to Two Timescale Update Rule (TTUR) [28]. The training is performed for 600 epochs for the CUB birds dataset and 120 epochs for the COCO dataset as previous work [10], [14]. To

compare with previous works, we use the same pre-trained sentence encoder and fix its parameters during training.

Evaluation Details. Following previous works [10], [14], we choose the Inception Score (IS) [27] and Fréchet Inception Distance (FID) [28] to evaluate the performance of our network. The Inception Score is formulated as:

$$I = \exp(\mathbb{E}_{\boldsymbol{x}} D_{KL}(p(y|\boldsymbol{x}) \,||\, p(y))), \tag{6}$$

where $\boldsymbol{x}$ is a generated image and y is the image label predicted by a pretrained Inception v3 network [56]. The IS computes the Kullback-Leibler (KL) divergence between conditional distribution $p(y|x)$ and marginal distribution $p(y)$. If the model could generate diverse and realistic images, the KL divergence between $p(y)$ and $p(y|x)$ should be large. Higher IS means higher quality of the generated images and each image clearly belongs to a specific class.

We note that the classes in CUB dataset used for training and testing are disjoint, the model has never seen the text descriptions in the test dataset before. But the Inception v3 network is pre-trained on the test dataset, it enables the IS on the CUB test dataset can evaluate text-image semantic consistency since only the images synthesized for testing meet the text descriptions in test dataset can get a higher IS.

The FID is another assessment which computes the Fréchet distance between the distribution of the synthetic images and real-world images in the feature space of a pre-trained Inception v3 network. The FID is formulated as:

$$F(r, g) = ||\mu_r - \mu_g||^2 + trace(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}), \tag{7}$$

where r is real data, g is generated data, μ_r, μ_g and Σ_r, Σ_g are the respective means and covariance of real data distribution and generated data distribution, respectively. Contrary to IS, More realistic images have a lower FID. To compute both IS and FID, each model generates 30000 images (256×256 resolution) from text descriptions randomly selected from the test dataset.

It should be pointed out that, as observed in Obj-GAN paper [16], we also found that IS on the COCO dataset completely fails in evaluating the synthesized image quality of text-to-image models. Thus, we do not compare the IS on the COCO dataset. While the FID is more robust and aligns human qualitative evaluation on the COCO dataset.

A. Quantitative Results

We compare the proposed method with several state-of-the-art methods including StackGAN++ [8], AttnGAN [10] and DM-GAN [14] which have achieved the remarkable success of text-to-image synthesis by using stacked structures.

As shown in Table I, our proposed DF-GAN achieves the highest Inception Scores (IS) compared with the state-of-the-art models on the CUB dataset. Higher IS on the test set of CUB means higher quality and text-image semantic consistency. Compared with SD-GAN which employs Siamese Network to ensure text-image semantic consistency, our DF-GAN improves the IS from 4.67 to 4.86. Compared with DM-GAN which introduces Memory Network to refine fuzzy image contents, our model also improves the IS from 4.75 to 4.86. The quantitative comparisons of Inception Score (IS) show that our proposed DF-GAN is able to synthesize more realistic images with better text-image semantic consistency.

TABLE I
THE INCEPTION SCORE (IS) OF OUR PROPOSED DF-GAN COMPARED WITH THE STATE-OF-THE ARTS ON THE TEST SET OF CUB.

Methods	Inception Score ↑
StackGAN++ [8]	4.04 ± 0.06
AttnGAN [10]	4.36 ± 0.03
MirrorGAN [13]	4.56 ± 0.05
SD-GAN [15]	4.67 ± 0.09
DM-GAN [14]	4.75 ± 0.07
DF-GAN (Ours)	**4.86** ± **0.04**

TABLE II
THE FID OF ATTNGAN, DM-GAN AND DF-GAN ON THE TEST SET OF CUB AND COCO.

Methods	CUB-FID↓	COCO-FID↓
AttnGAN [10]	23.98	35.49
DM-GAN [14]	**16.09**	32.64
DF-GAN (Ours)	19.24	**28.92**

As shown in Table II, compared with AttnGAN on FID, our DF-GAN decreases the FID from 23.98 to 19.24 (19.77% reduction) on the CUB dataset and from 35.49 to 28.92 (18.51% reduction) on the COCO dataset. Compared with DM-GAN, our model decreases the FID from 32.64 to 28.92 (11.39% reduction). Compared with CUB dataset, the COCO dataset is more challenging since there are always multiple objects in COCO images. It is still hard for current text-to-image models to synthesize images with multiple objects. But our DF-GAN decreases the FID significantly and achieves the best result on the COCO dataset. It demonstrates that our DF-GAN is able to synthesize more complex images with multiple objects.

The extensive quantitative evaluation results demonstrate the superiority and effectiveness of our proposed DF-GAN which is able to generate high-quality images with better semantic consistency and more complex images with multiple objects.

B. Qualitative Results

In this subsection, we compare the images synthesized by StackGAN++, AttnGAN, DM-GAN, and our DF-GAN. We first compare the quality of the synthesized images and text-image semantic consistency on the CUB dataset, then compare the results on a more challenging COCO dataset.

It can be seen that images synthesized by StackGAN++ and AttnGAN in Figure 9 look like a simple combination of fuzzy shape and some visual details (1^{st}, 2^{nd}, 4^{th}, 6^{th} and 7^{th} column). The reason is the employment of their stacked architecture and cross-modal spatial attention. Although DM-GAN introduces Memory Network to alleviate this problem, it is not completely solved. As shown in the 4^{th}, 6^{th} and 7^{th} column, the birds synthesized by StackGAN++, AttnGAN and DM-GAN have wrong shapes. Since we remove the stacked architecture and employ a novel text-to-image backbone and

Fig. 9. Examples of images synthesized by StackGAN++ [8], AttnGAN [10], DM-GAN [14] and our proposed DF-GAN conditioned on text descriptions from the test set of the CUB-200 birds.

Deep text-image Fusion Block (DFBlock), the images synthesized by our DF-GAN are more realistic, rather than a simple combination of shapes and details. Besides, the posture of the bird in our DF-GAN result is also more natural.

Comparing the text-image semantic consistency with other models, we find that our DF-GAN is also able to capture more fine-grained details in text descriptions. For example, as the result shown in 1^{st}, 5^{th}, 8^{th} column in Figure 9, other models cannot synthesize the "the bill is short and pointed", "a black stripe by its eyes" and "white wingbars" described in the text well, but our DF-GAN can synthesize them more correctly.

The superiority is more obvious on the challenging COCO dataset. It is hard for AttnGAN and DM-GAN to synthesize images with multiple objects since they employ the cross-modal spatial attention to fuse text and image features. As shown in the 1^{st}, 2^{nd} and 4^{th} column of Figure 10, the AttnGAN, and DM-GAN almost fail to synthesize semantically meaningful images from text descriptions. However, armed with Deep text-image Fusion Block (DFBlock) and Macthin-Aware Zero-Centered Gradient Penalty (MA-GP), our proposed DF-GAN is able to synthesize more realistic and semantically meaningful images with multiple objects in (see 1^{st}, 2^{nd}, 4^{th}, 5^{th} and 6^{th} column). The objects in the images synthesized by our proposed DF-GAN have clearer shapes and details. Moreover, the relationships between each object are also well presented according to given text descriptions in our DF-GAN results (see 1^{st}, 4^{th}, 5^{th}, 6^{th} and 8^{th} column).

As a complement of quantitative comparison, the qualitative results on the test set of CUB show that our DF-GAN is able to synthesize more realistic and text-image semantic images. In addition, the results on the coco dataset also prove the superiority of our DF-GAN in synthesizing more complex images with multiple objects.

C. Ablation Study

In this section, we conduct ablation studies on the testing set of the CUB dataset to verify the effectiveness of each component in the proposed DF-GAN. The components include a Novel Simplified text-to-image Backbone (NS-B), Matching-Aware zero-centered Gradient Penalty (MA-GP), one-way Discriminator (one-way D) and Deep text-image Fusion Block (DFBlock). Our baseline is a stacked text-to-image GAN which employs two-ways discriminator. In baseline, the sentence vector is naively concatenated to the input noise and intermediate feature maps. We first evaluate the effectiveness of NS-B, MA-GP, and one-way Discriminator. The results of Inception Score (IS) on the CUB dataset are shown in Table III and Figure 11.

Compared with baseline, our proposed Novel Simplified text-to-image Backbone (NS-B) improves the IS from 3.96 to 4.11. It proves that our simplified backbone is simpler and more effective than stacked architecture. Armed with MA-GP, the model improves the IS from 4.11 to 4.46 significantly which demonstrates that the MA-GP can further promote the generator to synthesize more realistic and text-image semantic consistent images. The one-ways discriminator also

Fig. 10. Examples of images synthesized by AttnGAN [10], DM-GAN [14] and our proposed DF-GAN conditioned on text descriptions from the test set of the COCO dataset.

TABLE III
THE PERFORMANCE OF DIFFERENT COMPONENTS OF OUR MODEL ON THE TEST SET OF CUB.

Architecture	Inception Score ↑
Baseline	3.96±0.05
NS-B	4.11±0.04
(NS-B) + (MA-GP)	4.46±0.04
(NS-B) + (MA-GP) + (one-way D)	**4.57±0.04**

TABLE IV
THE PERFORMANCE OF MA-GP GAN WITH DIFFERENT FUSION BLOCKS ON THE TEST SET OF CUB.

Architecture	Inception Score ↑
MA-GP GAN	4.57±0.04
+ CBNBlock	4.70±0.05
+ AFFBlock	4.73±0.05
+ CCBlock	4.75±0.04
+ DFBlock	**4.86±0.04**

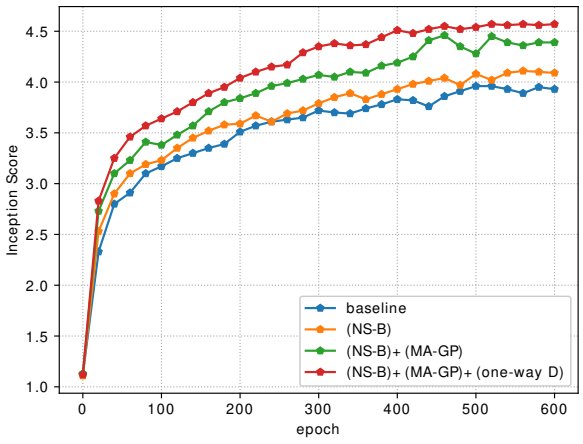

Fig. 11. Inception Scores of different variants at different epochs on the test set of the CUB dataset.

improves the IS from 4.46 to 4.57. It means that the one-way Discriminator is more effective than a two-ways discriminator in the text-to-image generation task. The result in Figure 11 also shows that the MA-GP and one-way discriminator improves the training stability and accelerates the convergence significantly.

To prove the superiority of DFBlock, we compare the DFBlock with CBNBlock, CCBlock, AFFBlock which employs conditional Batch Normalization, Conditional Convolution layer, one Affine Transformation layer to fuse text and image features, respectively. Their differences are shown in Figure 7. There IS are shown in Table IV and Figure 12. MA-GP GAN is the model that employs Novel Simplified text-to-image Backbone, Matching-Aware zero-centered Gradient Penalty, and one-way Discriminator. MA-GP GAN fuses the text and image features by naively concatenating.

From the results in Table IV and Figure 12, we find that compared with other fusion methods, concatenating cannot efficiently fuse text and image features. The comparison between CBNBlock and AFFBlock proves that Batch Normalization is not essential in Fusion Block, and removing normalization even slightly improves the results. The comparison between DFBlock and AFFBlock demonstrates the effectiveness of deepening the text-image fusion process. Our proposed DFBlock is also more effective than the recently proposed CCBlock. This further proves the superiority of our DFBlock. Compared with other text-image fusion methods, our proposed DFBlock achieves the best result.

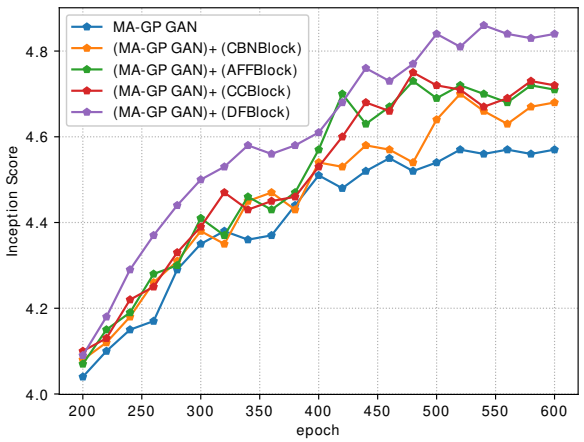

Fig. 12. Inception Scores of MA-GP GAN with different Fusion Blocks at different epochs on the test set of CUB. In order to show the comparison results more clearly, we start to compare the results of the 200th epoch.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel Deep Fusion Generative Adversarial Networks (DF-GAN) for text-to-image generation tasks. Compared with previous models, the proposed model is able to directly synthesize more realistic and text-image semantic consistent images without stacking architecture and extra networks. Moreover, we propose a novel Matching-Aware zero-centered Gradient Penalty to ensure the text-image semantic consistency and promote the generator converge to real data distribution. Besides, we decompose the effectiveness of Affine Transformation from CBN and present a Deep text-image Fusion Block to fuse text and image features more effectively. In addition, we propose the one-way discriminator which stabilizes the training process and accelerates the convergence of the generator network. Extensive experiment results show that our proposed DF-GAN significantly outperforms state-of-the-art models on the CUB dataset and more challenging COCO dataset.

In the future, we will try to add attention mechanisms and memory networks to DF-GAN for further improving the quality of generated images.

REFERENCES

[1] 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. 1, 3, 4

[2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1060–1069. 1, 2, 3

[3] 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. 1

[4] M. Yuan and Y. Peng, "Ckd: Cross-task knowledge distillation for text-to-image synthesis," *IEEE Transactions on Multimedia*, 2019. 1

[5] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7986–7994. 1

[6] 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*, 2020. 1

[7] 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. 1, 2, 3, 5

[8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE TPAMI*, vol. 41, no. 8, pp. 1947–1962, 2018. 1, 2, 3, 5, 8, 9

[9] 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. 1

[10] 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. 1, 2, 3, 4, 5, 7, 8, 9, 10

[11] 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. 1

[12] 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. 1

[13] 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. 1, 2, 3, 8

[14] 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. 1, 2, 3, 7, 8, 9, 10

[15] 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. 1, 2, 3, 6, 8

[16] 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. 1, 3, 8

[17] Y. Gou, Q. Wu, M. Li, B. Gong, and M. Han, "Segattngan: Text to image generation with segmentation attention," *arXiv preprint arXiv:2005.12444*, 2020. 1

[18] 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. 1

[19] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667. 2

[20] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4709–4717. 2

[21] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Advances in Neural Information Processing Systems*, 2017, pp. 6594–6604. 2, 6, 7

[22] T. Miyato and M. Koyama, "cgans with projection discriminator," *arXiv preprint arXiv:1802.05637*, 2018. 2, 6, 7

[23] 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. 2

[24] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363. 2, 3, 4

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011. 2, 7

[26] 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. 2, 7

[27] 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. 2, 3, 8
[28] 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. 2, 7, 8
[29] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019. 3, 6
[30] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in *CVPR*, 2019. 3
[31] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "Gesturegan for hand gesture-to-gesture translation in the wild," in *ACM MM*, 2018. 3
[32] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410. 3
[33] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in *CVPR*, 2020. 3
[34] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan, "Cycle in cycle generative adversarial networks for keypoint-guided image generation," in *ACM MM*, 2019. 3
[35] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119. 3
[36] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *IJCNN*, 2019. 3
[37] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *ECCV*, 2020. 3
[38] J. H. Lim and J. C. Ye, "Geometric gan," *arXiv preprint arXiv:1705.02894*, 2017. 3, 4
[39] S. Martin Arjovsky and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia*, 2017. 3
[40] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018. 3
[41] A. Karnewar and R. S. Iyengar, "Msg-gan: Multi-scale gradients gan for more stable and synchronized multi-scale image synthesis," *arXiv preprint arXiv:1903.06048*, 2019. 3
[42] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. 3
[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777. 3
[44] H. Thanh-Tung, T. Tran, and S. Venkatesh, "Improving generalization and stability of generative adversarial networks," in *International Conference on Learning Representations*, 2019. 3
[45] J. Wu, Z. Huang, J. Thoma, D. Acharya, and L. Van Gool, "Wasserstein divergence for gans," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 653–668. 3
[46] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International Conference on Machine Learning*, 2018, pp. 3481–3490. 3, 4, 6
[47] 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. 3
[48] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European conference on computer vision*. Springer, 2016, pp. 791–808. 3
[49] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European conference on computer vision*. Springer, 2016, pp. 135–153. 3
[50] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448. 3
[51] C. Gulcehre, S. Chandar, K. Cho, and Y. Bengio, "Dynamic neural turing machine with continuous and discrete addressing schemes," *Neural computation*, vol. 30, no. 4, pp. 857–884, 2018. 3
[52] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *International Conference on Learning Representations*, 2015. 3
[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015. 6, 7
[54] M.-C. Sagong, Y.-G. Shin, Y.-J. Yeo, S. Park, and S.-J. Ko, "cgans with conditional convolution layer," *arXiv preprint arXiv:1906.00709*, 2019. 7
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015. 7
[56] 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. 8

Ming Tao recieved the B.E. degree from Sanjiang University, China, in 2018. He is currently pursuing the Master degree at college of automation, Nanjing University Of Posts And Telecommunications. His research interests include text-to-image synthesis, generative adversarial network, cross-modal retrieval and natural language processing.

Hao Tang is a Ph.D. candidate in the Department of Information Engineering and Computer Science and a member of Multimedia and Human Understanding Group (MHUG) at the University of Trento. He received the Master degree in computer application technology in 2016 at the School of Electronics and Computer Engineering, Peking University, China. His research interests are machine learning, human-computer interaction, (deep) representation learning, robotics and their applications to computer vision.

Songsong Wu received the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, China, in 2012. He was a Visiting Scientist with Texas State University, San Marcos, TX, USA, from 2018 to 2019. He is currently an Assistant Professor with the School of Computer, Guangdong University of Petrochemical Technology, and an Assistant Professor with the School of Automation, Nanjing University of Posts and Telecommunications. His research interests include deep learning, fault diagnosis, and image recognition.

Nicu Sebe is Professor with the University of Trento, Italy, leading the research in the areas of multimedia information retrieval and human behavior understanding. He was the General Co-Chair of ACM Multimedia 2013, and the Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010, ACM Multimedia 2007 and 2011, ICCV 2017 and ECCV 2016. He is a Program Chair of ICPR 2020. He is a fellow of the International Association for Pattern Recognition.

Fei Wu received the Ph.D. degree in computer science from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2016. He is currently with the College of Automation in NJUPT. He has authored over forty scientific papers. His research interests include pattern recognition, artificial intelligence, and computer vision.

Xiao-Yuan Jing received the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, China, in 1998. He was a Professor with the Department of Computer, Shenzhen Research Student School, Harbin Institute of Technology, in 2005. He is currently a Professor and the Dean of the School of Computer, Guangdong University of Petrochemical Technology, and a Professor with the School of Computer, Wuhan University, China. He has published more than 100 scientific papers in international journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Information Forensics and Security, the IEEE Transactions on Neural Networks and Learning Systems, TCB, TR, CVPR, AAAI, and IJCAI. His research interests include pattern recognition, machine learning, artificial intelligence, and fault diagnosis.