FuseGAN: A Global Cross-Modal Fusion Baseline for Text to Image Synthesis

Anonymous submission

Abstract

It's a challenging task to realize cross-modal image synthesis conditioned on given text descriptions. Compared with previous stacked architectures, the single-stage synthesis method has recently progressed due to its lightweight and efficiency. However, a closer look reveals three flaws in these works. First, the methods usually don't achieve sufficient global cross-modal text-image information fusion both in spatial and channel, which severely limits the capability of these networks. Second, existing text encoders often can't reflect attention between different words in a text description. Third, previous single-stage works tend to employ single text-adaptive discriminator to provide weak feedback for generator. To address these issues, we propose FuseGAN, a global cross-modal fusion baseline for text to image synthesis. Specifically, (i) we build a new single-stage backbone network and propose a novel global cross-modal fusion block (FuseBlock), to achieve global cross-modal information fusion both in spatial and channel with slight computational cost, and (ii) we propose an attention-based text encoder that embodies the difference of each word in a text description, which can be a general component for text to image synthesis, (iii) we incorporate image contrastive loss and semantic contrastive loss to improve the fidelity and semantic consistency of generated images. Extensive experiments demonstrate that FuseGAN achieves state-of-the-art performance. On CUB datasets, we reach a new state-of-the-art FID 10.16. On COCO datasets, compared with the current state-of-the-art model Lafite, we achieve comparable performance (FID 11.92 *vs.* 8.12) only with 20% model parameters.

1 Introduction

Text to image synthesis aims to generate realistic images based on given text descriptions, which is a difficult task due to its cross-modal nature. Recently, it has gained widespread attention due to the potential value in various fields (such as computer-aided design, virtual scene generation, photo editing). To reach this, many methods have been proposed, including: Generative Adversarial Networks (Goodfellow et al. 2014), Diffusion Models (Gu et al. 2022), variational auto-encoders (Kingma and Welling 2014), *etc.* Among them, generative adversarial networks have achieved notable success and recently achieved promising results (Zhang et al. 2021; Zhou et al. 2022; Xia et al. 2021; Tao et al. 2022).

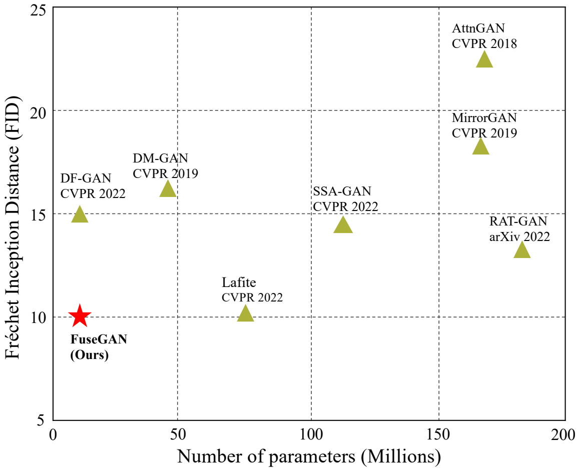


Figure 1: Performance of different text to image synthesis methods on CUB datasets. Lower FID means better. Our proposed FuseGAN achieves a new state-of-the-art FID with fewer model parameters.

Due to the instability of GANs, previous methods mainly adopt stacked architecture, employing multiple generator-discriminator pairs to gradually complete image synthesis from coarse to fine (Zhang et al. 2017, 2018; Xu et al. 2018; Li et al. 2019; Ruan et al. 2021). However, there are two problems with the stacked methods. First, the stacked architecture introduces entanglement, which means the generated images look like a combination of visual features from previous stages. Second, the stacked structure often utilizes cross-modal attention for text-image fusion, which imposes a huge training burden due to its high computational cost. To address these issues, single-stage architecture has been proposed (Tao et al. 2020) and received enormous attention. Compared with stacked architecture, the single-stage architecture utilizes single generator-discriminator pair for training. In this work, we follow the single-stage model architecture.

However, a careful study of the existing methods reveals the following three issues. First, the existing methods usually don't achieve sufficient global cross-modal text-image information fusion both in spatial and channel, which severely limits the capabilities of these networks. In the past, there were three methods to achieve cross-modal text-image fusion: concatenation, cross-modal attention and Affine operation.

The method of concatenation (Reed et al. 2016b) directly concatenate textual features and visual features together. Obviously, this method can't achieve effective text-image fusion. Although the previously adopted cross-modal attention (Xu et al. 2018) extracts long-range relationships and achieves spatial fusion well, it can't achieve cross-modal fusion between channels and brings high computational cost. Moreover, the Affine operation separately fuses textual features into each channel of visual feature maps, which has shown its advantages in recent research (Ye, Liu, and Tan 2022; Liao et al. 2022). However, it ignores the fusion on spatial which treats spatial pixels equally. More importantly, these works all employ convolutional neural networks as backbone network. Convolutional networks can capture local visual features well, but it may not be effective for global cross-modal text-image fusion (Guo et al. 2022; Wu et al. 2021).

Second, existing text encoders often can't reflect attention between different words in a text description (Radford et al. 2021; Schuster and Paliwal 1997). Most of previous works utilizes bidirectional LSTM to encode text description. Compared with the original fully-connected layer encoder (Reed et al. 2016b), the LSTM-based text encoder can better capture the context relationship. But it ignores the difference of different words in a text description, which will cause some semantically irrelevant words to impose an impact on model, resulting in poor performance. Third, previous singel-stage works prefer to employ single text-adapative discriminator to provide weak feedback for generator (Ye, Liu, and Tan 2022; Tao et al. 2022). It was insufficient to supervise generator to synthesis photo-realistic and text-matching images. We argue that more powerful supervision should be introduced in order to generate desired images.

To solve these, we propose FuseGAN, which aims to achieve global cross-modal text-image fusion both in spatial and channel. For the first issue, we propose a novel global cross-modal fusion block (FuseBlock), to achieve global cross-modal information fusion both in spatial and channel with slight computational cost. Furthermore, inspired by StyleGAN (Karras, Laine, and Aila 2019) and ResNet (He et al. 2016), we build a new single-stage backbone network (Details in Fig. 2(a)). FuseGAN unlocks the ability to achieve global cross-modal text-image fusion only with slight computational cost. For the second issue, we propose a novel attention-based text encoder. Before extracting global sentence vectors using a bidirectional LSTM, we first refine the word embedding using a vanilla Transformer encoder in order to give each word different weight. By doing so, our attention-based text encoder can not only learn the context relationships but also distinguish the difference between different words. Besides, our attention-based text encoder can be a general component for other text to image synthesis works. For the third issue, we introduce image contrastive loss and semantic contrastive loss to improve the fidelity and semantic consistency of generated images. Specifically, we additionally introduce LPIPS loss and DAMSM loss to supervise generator. What's important, we can regard SAFM as spatial mixer and CAFM as channel mixer. They all can be replaced by the components with same role to further improve model capability. From this perspective, we hope FuseGAN can become a new global cross-modal fusion baseline to inspire future research.

In summary, our contributions are as follows:

- We propose FuseGAN by building a new single-stage backbone network and introducing a novel cross-modal fusion block (FuseBlock), which can achieve global cross-modal text-image fusion more effectively and deeply both in spatial and channel with slight computational cost.
- We propose an attention-based text encoder that embodies the difference of different words in description, which can be a general component for text to image synthesis. Our attention-based text encoder can not only learn context relationships in descriptions, but also can distinguish the difference between different words.
- We introduce image contrastive loss and semantic contrastive loss to improve the fidelity and semantic consistency of generated images.
- Extensive experiments demonstrate that our proposed FuseGAN achieves the state-of-the-art results. On CUB dataset, we reach a new state-of-the-art FID 10.16. On COCO dataset, compared with current state-of-the-art model Lafite (Zhou et al. 2022), we achieve comparable performance (FID 11.92 *vs.* 8.12) only with 20% model parameters.

2 Related Work

Reed *et al.* first proposed employing conditional generative adversarial networks to generate images under text conditions in 2016 (Reed et al. 2016a), which opened the door to text to image synthesis. To further improve the quality of generated images, Zhang *et al.* proposed stacking multiple generator-discriminator pairs to gradually generate high-quality images from coarse to fine under text conditions (Zhang et al. 2017). During training, multiple generator-discriminator pairs are required to coordinate to generate high-quality images. After that, Xu *et al.* proposed AttnGAN (Xu et al. 2018) to achieve word-level fine-grained generation by introducing a word-level attention mechanism. In addition, AttnGAN also proposes DAMSM to supervise generator to synthesis images that are semantically consistent with the corresponding texts. Zhu *et al.* proposed using a dynamic memory network to purify the initial image. Li *et al.* proposed ControlGAN (Li et al. 2019) based on the word-level spatial attention and channel attention, in order to generate controllable and more realistic images. For a period of time, the stacked architecture has become the basic method for text to image synthesis.

To overcome the limitations of stacked architecture, Ming *et al.* proposed DF-GAN (Tao et al. 2020), which aims to employ single generator to complete text to image synthesis. In addition, he also proposed utilizing MA-GP to generate text-matching images. The following SSA-GAN (Liao et al. 2022) and RAT-GAN (Ye, Liu, and Tan 2022) also adopted single-stage architecture. SSA-GAN proposes semantic-spatial condition batch normalization, which employs mask map to overcome the problem of insufficient spatial fusion in DF-GAN. RAT-GAN proposes Recurrent Affine Transformation to model long-range dependencies between fusion blocks.

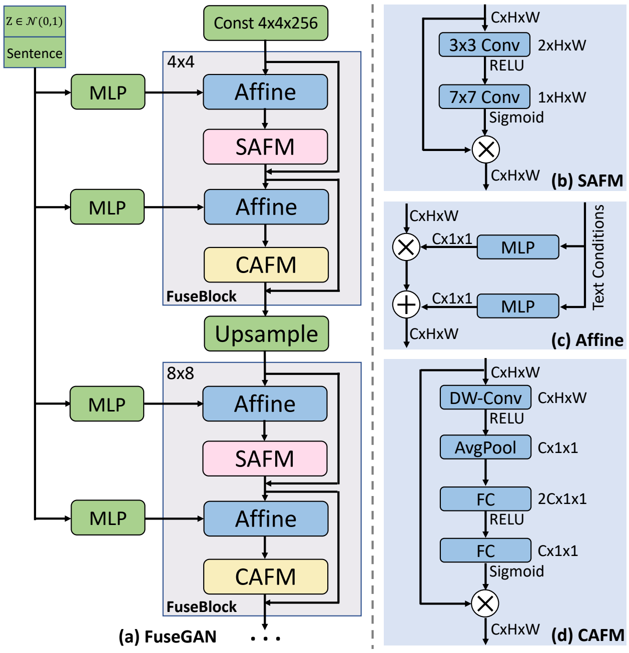


Figure 2: The architectures of FuseGAN. Our proposed FuseGAN builds a new single-stage backbone network and consists of 7 FuseBlocks from 4^2 to 256^2, which can achieve global cross-modal text-image fusion both in spatial and channel with slight computational cost. The FuseBlock consists of three modules: Spatial Attention Fusion Module (SAFM), Channel Attention Fusion Module (CAFM), and Affine module. The "Sentence" stands for sentence vector encoded by our proposed attention-based text encoder.

Our proposed FuseGAN also adopts the single-stage architecture. Different from all previous works, we build a new single-stage backbone network and propose a novel cross-modal fusion block (FuseBlock), unlocking the ability to achieve global cross-modal text-image fusion both in spatial and channel with slight computational cost.

3 Method

In this paper, we propose FuseGAN, to achieve global cross-modal text-image fusion for text-to-image synthesis. We will introduce FuseGAN in this section.

3.1 Model Overview

The architecture of generator is shown in Fig. 2(a). We follow the single-stage architecture (Tao et al. 2020). First, inspired by StyleGAN (Karras, Laine, and Aila 2019), we replace random noise with a learnable constant as the input to the first layer. Second, we build an innovative basic computational block (called FuseBlock) to unlock the ability of global cross-modal text-image fusion with slight computational cost. FuseBlock includes three modules: Affine, SAFM, and CAFM. Given the widespread success of ResNet (He et al. 2016), FuseBlock adopts the residual design. Thus, unlike previous works, we build a new single-stage backbone network, which we called style-based backbone network.

We first generate a 100-dimensional random noise sampled from Gaussian distribution. Then, the text descriptions are encoded to get sentence vector by a pretrained attention-based text encoder (Details in Fig. 3(b)). We concatenate random noise with sentence vector to get the input to model. From 4^2 to 256^2, FuseGAN consists of 7 FuseBlocks. The mathematical form of FuseBlock is shown below:

$$\begin{aligned} \hat{h}^l &= \mathrm{SAFM}(\mathrm{AFF}(h^{l-1}, s)) + h^{l-1}, \\ h^l &= \mathrm{CAFM}(\mathrm{AFF}(\hat{h}^l, s)) + \hat{h}^l, \end{aligned} \tag{1}$$

where h^l is the output feature map of FuseBlock l, AFF is Affine module, SAFM and CAFM will be introduced in next subsection, respectively. After SAFM and CAFM, we employ a convolution network to align channels of feature maps.

For the Affine module (See Fig. 2(c)), following previous works (Tao et al. 2022; Liao et al. 2022), we stack two MLPs (Tolstikhin et al. 2021) to predict channel-wise scaling parameters and shifting parameters. The mathematical form is shown below:

$$\begin{aligned} \gamma &= \mathrm{MLP}(s), \\ \beta &= \mathrm{MLP}(s), \\ \hat{x}_i &= \gamma_i \cdot x_i + \beta_i, \end{aligned} \tag{2}$$

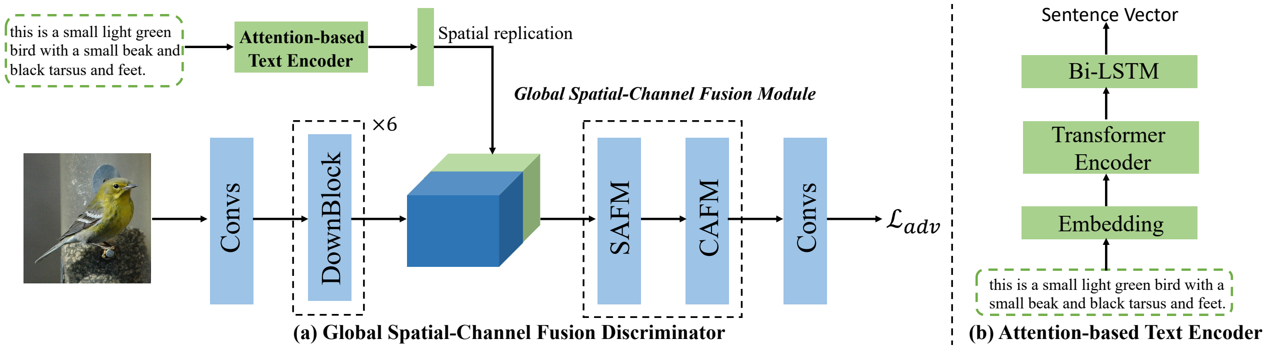
where s is the input of text conditions, γ is the scaling parameter, β is the shifting parameter, $\hat{x}_i$ and x_i represent the i-th channel of output and input feature maps, respectively.

3.2 Global Spatial-Channel Text-Image Fusion

The ability of cross-modal fusion can significantly affect model performance, which hasn't received deserved attention in previous works. To solve it, we propose FuseGAN to perform global cross-modal text-image fusion. In this subsection, we detail three novel components of FuseGAN: Spatial Attention Fusion Module, Channel Attention Fusion Module, and Global Spatial-Channel Fusion Discriminator.

Spatial Attention Fusion Module The Spatial Attention Fusion Module aims to promote the spatial fusion of cross-modal information and enhance long-range modeling ability in spatial. Most previous works solely employed convolution operations to establish local receptive fields, ignoring the long-range information dependencies. Besides, the previously employed cross-modal attention mechanism brings a heavy training burden due to its expensive computational cost. To overcome those, we propose a simple but effective lightweight spatial fusion component, Spatial Attention Fusion Module (Details in Fig. 2(b)).

First, we employ a 3×3 convolution network with output channels being 3, followed by a ReLU function. Compared with the max pooling and average pooling operations used by CBAM (Woo et al. 2018), the convolution network with output channels being 3 has better adaptability. Then, stacking a 7×7 convolution network with padding being 3, stride being 1, and output channel being 1. Compared with the 3×3 convolution kernel, the convolution kernel of 7×7 has a larger local receptive field which can capture longer dependencies. Followed sigmoid functions scale all values between (0,1). The obtained result is multiplied pixel-wisely with original feature map to complete spatial fusion. Compared with


this is a small light green bird with a small beak and black tarsus and feet.
Attention-based Text Encoder
Spatial replication
Global Spatial-Channel Fusion Module
×6
Convs
DownBlock
SAFM
CAFM
Convs
$\mathcal{L}_{adv}$
Sentence Vector
Bi-LSTM
Transformer Encoder
Embedding
this is a small light green bird with a small beak and black tarsus and feet.

(a) Global Spatial-Channel Fusion Discriminator

(b) Attention-based Text Encoder

Figure 3: (a) Global Spatial-Channel Fusion Discriminator: we achieve global cross-modal text-image fusion in discriminator. (b) Attention-based Text Encoder: before Bi-LSTM, we stack a vanilla Transformer encoder (Vaswani et al. 2017) to reflect attention between words.

cross-modal attention mechanism, we achieve global spatial text-image information fusion only with slight computational cost. Furthermore, to verify the novelty of SAFM, we conduct ablation study with spatial attention module (Woo et al. 2018) and self-attention (Vaswani et al. 2017) (See Table 2).

Channel Attention Fusion Module The Channel Attention Fusion Module aims to facilitate cross-modal fusion and explicitly model the inter-dependencies between channels. The previous Affine-based method is to separately fuse text conditions to each channel, and then stack a convolution network for local modeling. The sole Affine-based method can't realize sufficient global cross-modal fusion between channels. Inspired by SE-Net (Hu, Shen, and Sun 2018), we propose Channel Attention Fusion Module (Details in Fig. 2(d)).

In Channel Attention Fusion Module, we first stack a 3×3 DW-Conv operation followed by a ReLU function to extract channel-wise features. Then, we employ average pooling to scale the size of feature map to $C{\times}1{\times}1$. After that, we employ two fully connected layers to get the initial weights for each channel. The followed sigmoid function scales the initial weights of each channel between (0,1). The obtained result is channel-wisely multiplied with original feature map to complete the fusion between channels. Besides, to verify the novelty of CAFM, we also conduct ablation study with channel attention module (Woo et al. 2018) and SE block (Hu, Shen, and Sun 2018) (see Table 2).

Moreover, we can regard SAFM as spatial mixer and CAFM as channel mixer. They all can be replaced by the components with same role to further improve model capability. From this perspective, we hope FuseGAN can become a global cross-modal fusion baseline to inspire future research.

Global Spatial-Channel Fusion Discriminator The architecture of discriminator is shown in Fig. 3(a). In discriminator, most previous works directly concatenate text and image feature maps, followed by a series of convolution operations, which can't achieve sufficient information fusion. The recently proposed RAT-GAN (Ye, Liu, and Tan 2022) takes the issue into account, introducing spatial attention in the discriminator. However, RAT-GAN only enhances the fusion in spatial and works on text feature maps. To enhance global fusion in discriminator, we propose Global Spatial-Channel Fusion Discriminator. Specifically, after several down-sampling blocks, we introduce SAFM and CAFM to facilitate cross-modal information fusion. Compared with previous works, our discriminator can significantly improve cross-modal fusion ability.

3.3 Attention-based Text Encoder

Converting text descriptions into embedding is an important prior work for generating images conditioned on natural language. The previous bidirectional LSTM text encoder has shown its effectiveness in many works, which can capture context relationships in text descriptions well. But it ignores the difference between different words, which leads to some semantically irrelevant words also having an impact on the model. To solve this, we propose attention-based text encoder, which aims to reflect the attention between different words (Details in Fig. 3(b)).

First, we get the initial word matrix $e \in \mathbb{R}^{n \times d}$ for text description, where n is the number of words and d is the length of word embedding, respectively. Then the word matrix e is transformed by fully connected layers. And the self-attention function occurs among the transformed results. The mathematical form is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V. \tag{3}$$

In fact, we employ multi head self-attention (Vaswani et al. 2017), which improves the representation space of model. Dividing Q, K, and V into h heads in parallel for dot product attention. The dot-product attention for each head is computed separately. The mathematical form is as follows:

$$\begin{aligned} \text{MHSA} &= \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{4}$$

where W^O is a learnable matrix, respectively. Then a feedforward network is used to capture relationships within words. Finally, stacking a bidirectional LSTM to get sentence vector. In this way, our proposed attention-based encoder can better focus on semantically relevant words. Our proposed attention-based text encoder can be a general component for other text to image synthesis works.

Method	CUB			COCO		
	FID↓	IS↑	R-precision↑	FID↓	R-precision↑	NoP↓
StackGAN++ (Zhang et al. 2018)	15.30	4.04 ± .06	-	81.59	-	103M
AttnGAN (Xu et al. 2018)	23.98	4.36 ± .03	0.246	35.49	0.183	169M
DM-GAN (Zhu et al. 2019)	16.09	4.75 ± .07	0.287	32.64	0.236	46M
MirrorGAN (Qiao et al. 2019)	18.34	4.54 ± .17	-	34.71	-	170M
ControlGAN (Li et al. 2019)	13.92	4.58 ± .09	0.308	33.58	0.248	200M
DAE-GAN (Ruan et al. 2021)	15.19	4.42 ± .04	0.321	28.12	0.257	49M
XMC-GAN (Zhang et al. 2021)	-	-	-	9.33	-	166M
Lafite (Zhou et al. 2022)	10.48	5.97 ± .- -	-	**8.12**	0.318	75M
DF-GAN (Tao et al. 2022)	14.81	5.10 ± .- -	0.306	19.32	0.278	19M
SSA-GAN (Liao et al. 2022)	15.61	5.17 ± .08	0.326	19.37	0.264	110M
RAT-GAN (Ye, Liu, and Tan 2022)	13.91	5.36 ± .20	0.368	14.60	0.298	186M
FuseGAN (ours)	**10.16**	**5.99 ± .12**	**0.386**	11.92	**0.326**	**15M**

Table 1: The results of IS, R-precision, FID and NoP compared with the state-of-the-art methods on the test set of CUB and COCO. ↓ means lower is better. ↑ means higher is better.

3.4 Loss Function

Previous single-stage works tend to employ single text-adaptive discriminator to provide weak feedback for generator. To generate more realistic and text-matching images, we introduce image contrastive loss and semantic contrastive loss.

Image Contrastive Loss To better guide the generator by comparing the difference between the generated and original image, we introduce image contrastive loss in addition. Specifically, we employ perceptual loss (Johnson, Alahi, and Fei-Fei 2016). The mathematical form is as follows:

$$\mathcal{L}_{image} = ||\mathrm{F}(x) - \mathrm{F}(\hat{x}))||_2^2, \tag{5}$$

where F is the pretrained VGG network (Simonyan and Zisserman 2015), x is the original image, and $\hat{x}$ is the generated image, respectively.

Semantic Contrastive Loss To promote the semantic consistency between the generated image and text description, we introduce the DAMSM loss (Xu et al. 2018). The mathematical form is as follows:

$$\mathcal{L}_{semantic} = \mathrm{DAMSM}(s, \hat{x}), \tag{6}$$

Where DAMSM is the pretrained aligned tool, s is the text description, and $\hat{x}$ is the generated image, respectively.

Overall Loss Similar to DF-GAN, we use hinge loss with MA-GP (Tao et al. 2020) as the discriminator loss. The specific mathematical form of discriminator loss is as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{D} = \; & \mathbb{E}_{x \sim P_{data}}[\max(0, 1 - \mathrm{D}(x, s))] \\ & + \frac{1}{2}\mathbb{E}_{x \sim P_{G}}[\max(0, 1 + \mathrm{D}(\hat{x}, s))] \\ & + \frac{1}{2}\mathbb{E}_{x \sim P_{data}}[\max(0, 1 + \mathrm{D}(x, \hat{s}))],\end{aligned} \tag{7}$$

where x is the real image, $\hat{x}$ is the generated image, s is the matched sentence, $\hat{s}$ is the mismatched sentence, and D is the discriminator, respectively. The training loss of the generator is as follows:

$$\begin{aligned}\mathcal{L}_{G} &= \lambda_1 \mathcal{L}_{semantic} + \lambda_2 \mathcal{L}_{image} + \mathcal{L}_{adv}^{G}, \\ \mathcal{L}_{adv}^{G} &= -\mathbb{E}_{x \sim P_{data}}[\mathrm{D}(\hat{x}, s)],\end{aligned} \tag{8}$$

where λ_1 and λ_2 are hyperparameters, respectively.

4 Experiments

In this section, we first introduce the datasets, training details, and evaluation details. Then, we evaluate FuseGAN qualitatively and quantitatively on two challenging datasets and conduct ablation study to get an insight of the effectiveness of every module we proposed.

Datasets We conduct our experiments on two challenging datasets CUB (Wah et al. 2011) and COCO (Lin et al. 2014). The CUB bird datasets (200 categories) contain 8855 training images and 2933 testing images. Each image has 10 text descriptions. The COCO datasets contain 80k images for training and 40k images for testing. Each image has five language descriptions.

Training details Our model is implemented in PyTorch. The Adam optimizer (Kingma and Ba 2015) with β_1 = 0.0 and β_2 = 0.9 is used in the training. The learning rate is set to 0.0001 for generator and 0.0004 for discriminator according to TTUR (Heusel et al. 2017). The hyper-parameters λ_1 and λ_2 are set to 0.02 and 0.004, respectively.

Evaluation Metric We choose Fréchet Inception Distance (FID) (Heusel et al. 2017), Inception Score (IS) (Salimans et al. 2016), and top-1 R-precision (Xu et al. 2018) to evaluate the performance of our work. For FID, it computes the Fréchet distance between the distribution of the generated images and real-world images in the feature space of a pre-trained Inception v3 network (Szegedy et al. 2016). For IS, it computes the Kullback-Leibler (KL) divergence between a conditional distribution and marginal distribution. Lower

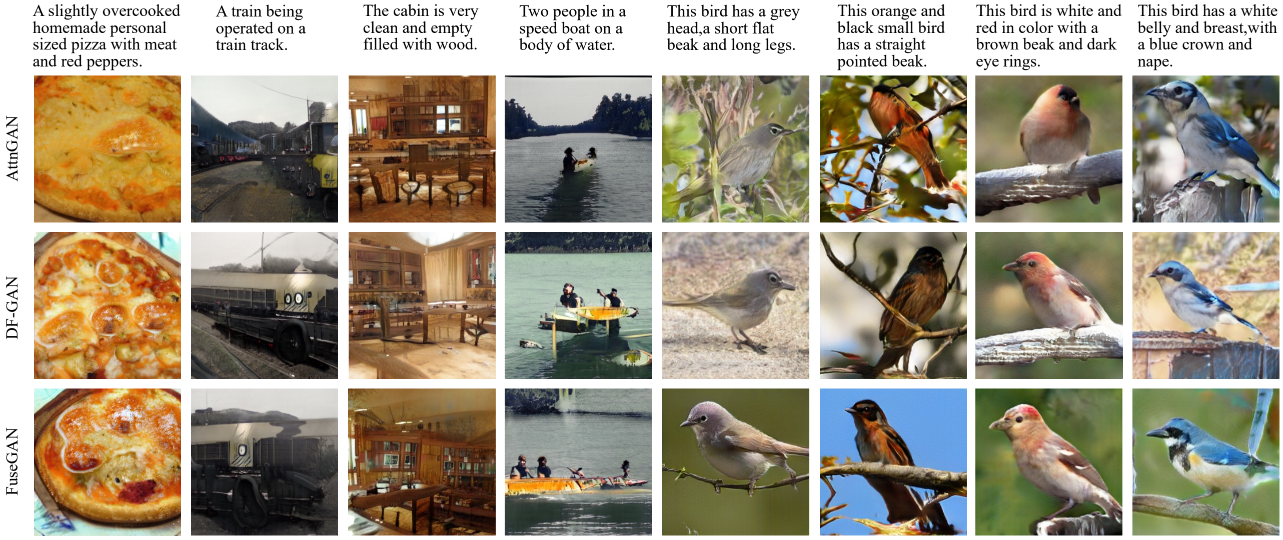
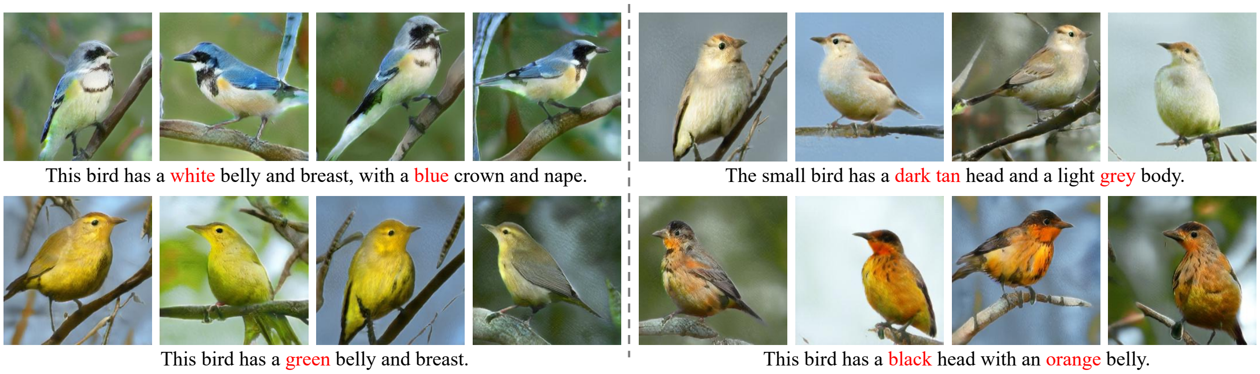

A slightly overcooked homemade personal sized pizza with meat and red peppers.
A train being operated on a train track.
The cabin is very clean and empty filled with wood.
Two people in a speed boat on a body of water.
This bird has a grey head,a short flat beak and long legs.
This orange and black small bird has a straight pointed beak.
This bird is white and red in color with a brown beak and dark eye rings.
This bird has a white belly and breast,with a blue crown and nape.
AttnGAN
DF-GAN
FuseGAN

Figure 4: Qualitative comparison between AttnGAN (Xu et al. 2018), DF-GAN (Tao et al. 2022), and our proposed FuseGAN conditioned on text descriptions from the test set of COCO datasets (1st - 4th columns) and CUB datasets (5th - 8th columns).



This bird has a white belly and breast, with a blue crown and nape.

The small bird has a dark tan head and a light grey body.

This bird has a green belly and breast.

This bird has a black head with an orange belly.

Figure 5: Diversity image generation examples of FuseGAN on CUB datasets. Our proposed FuseGAN can generate multiple text-matching and photo-realistic images on same text description. Each caption corresponds to four images.

FID and higher IS mean model achieves better performance. For R-precision, we use CLIP (Radford et al. 2021) to calculate the cosine similarity between original image and given description. To compute these, each model generates 30,000 images (256×256 resolution) from text descriptions randomly selected from test datasets.

Following previous works (Liao et al. 2022; Tao et al. 2022), we do not use IS on COCO datasets because it can't evaluate the image quality well. Moreover, we evaluate the number of parameters (NoP) to compare the model size with other methods.

4.1 Quantitative Evaluation

As shown in Table 1, we conduct the quantitative comparison between our proposed FuseGAN and previous stacked or single-stage methods. On two challenging datasets, we achieve state-of-the-art performance with fewer model parameters.

On CUB datasets, we reach a new state-of-the-art FID 10.16. Compared with the previous singe-stage baseline DF-GAN (Tao et al. 2022), our FuseGAN decreases the FID metric from 14.81 to 10.16 and improves the IS metric from 5.10 to 5.99, R-precision from 0.306 to 0.386. On COCO datasets, compared with current state-of-the-art model Lafite (Zhou et al. 2022), our proposed FuseGAN achieves comparable performance (FID 11.92 *vs.* 8.12) only with 20% model parameters. Compared with DF-GAN, our FuseGAN decreases the FID metric from 19.32 to 11.92 and improves R-precision from 0.278 to 0.326. And compared with SSA-GAN, RAT-GAN, and other stacked methods, FuseGAN demonstrates outstanding performance.

4.2 Qualitative Evaluation

As shown in Fig. 4, we conduct the qualitative comparison between our proposed FuseGAN, the single-stage method DF-GAN (Tao et al. 2022) and the stacked method AttnGAN (Xu et al. 2018). Compare with other works, our generated images are more photo-realistic and text-matching. For ex-

Ablation	Variant	FID ↓	R-precision ↑
Baseline	FuseGAN (ours)	10.16	0.386
Component	Previous Single-stage Baseline (Tao et al. 2022)	14.81	0.306
	+ Style-based Backbone Network	13.96	0.328
	+ SAFM & CAFM	11.88	0.362
	+ Contrastive Loss	10.90	0.368
	+ Attention-based Text Encoder	10.16	0.386
SAFM	SAFM w/o 7×7 Conv	13.64	0.346
	SAFM → Multi Head Self Attention (Vaswani et al. 2017)	10.04	0.366
	SAFM → Spatial Attention Module (Woo et al. 2018)	11.88	0.357
	SAFM → Convolutional Block Attention Module (Woo et al. 2018)	13.88	0.345
CAFM	CAFM w/o DW-Conv	12.48	0.375
	CAFM → SE Block (Hu, Shen, and Sun 2018)	12.97	0.363
	CAFM → Channel Attention Module (Woo et al. 2018)	12.56	0.371
	CAFM → Convolutional Block Attention Module (Woo et al. 2018)	13.43	0.381
Affine	Affine → Concat (Reed et al. 2016b)	18.96	0.287
	Affine → CBN (Brock, Donahue, and Simonyan 2019)	13.48	0.331
	Affine → AdaIN (Karras, Laine, and Aila 2019)	12.64	0.361
	Affine → AFF Block (Tao et al. 2022)	11.86	0.377

Table 2: Ablation Study of different components, SFAM, CAFM, and Affine on the test set of CUB. ↓ means lower is better. ↑ means higher is better.

ample, in the 6th column, given the text "This orange and black small bird has a straight pointed beak", the image generated by FuseGAN has all the mentioned attributes. However, the image generated by DF-GAN does not reflect "straight pointed beak" and the image generated by AttnGAN is a little blurry. In the 8th column, FuseGAN produces the desired image, but other methods don't produce clear images to meet all attributes. Besides, as shown in Fig. 5, FuseGAN can generate a variety of realistic images on same text condition.

4.3 Ablation Study

As shown in Table 2, to verify the superiority of each component in our proposed FuseGAN, we deploy our experiments on the CUB test set (Wah et al. 2011).

Baseline The baseline is our proposed FuseGAN, a novel single-stage backbone network to achieve global cross-modal text-image fusion with slight computational cost.

Effectiveness of Component We start from previous single-stage baseline (Tao et al. 2022) and add each component sequentially. The style-based backbone network, SAFM & CAFM, contrastive loss, and attention-based encoder are added in order. It shows that each component improves performance, which verifies the effectiveness of our proposed each innovative modules.

Effectiveness of SAFM We ablate our proposed Spatial Attention Fusion Module. Compared with spatial attention module (Woo et al. 2018), CBAM (Woo et al. 2018), our SAFM achieves better performance. Compared with multi head self-attention (Vaswani et al. 2017), although our performance is slightly worse, SAFM has lighter training burden. Furthermore, we also verify the effectiveness of the employed 7×7 convolution, which can bring larger local receptive field to model.

Effectiveness of CAFM We ablate our proposed Channel Attention Fusion Module. Compared with channel attention module (Woo et al. 2018), CBAM (Woo et al. 2018), SE Block (Hu, Shen, and Sun 2018), our CAFM achieves better performance. Importantly, compare with similar modules SE Block, our proposed CAFM is more suitable for the current network. We also try to replace CAFM or SAFM with CBAM but do not achieve better performance. Furthermore, we verify the effectiveness of DW-Conv, which can better extract channel-wise information.

Effectiveness of Affine Following previous works, we explore the effectiveness of Affine module. Compared with the previously used Concat (Reed et al. 2016b), CBN (Brock, Donahue, and Simonyan 2019), AdaIN (Karras, Laine, and Aila 2019), AFF Block (Tao et al. 2020), the Affine module is more suitable for our proposed FuseGAN.

5 Conclusion

In this paper, we propose FuseGAN, a global cross-modal fusion baseline for text to image synthesis. We build a new single-stage backbone network and propose a novel fusion block (FuseBlock). Furthermore, we propose an attention-based text encoder to reflect the attention of different words. We introduce image contrastive loss and semantic contrastive loss to generate desired images. Extensive experiments on two challenging datasets demonstrate that FuseGAN achieves state-of-the-art performance with fewer model parameters. In the future, we hope to explore more efficient fusion blocks to further improve model capability.

References

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale GAN training for high fidelity natural image synthesis. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *In Proceedings of the Advances in neural information processing systems (NeurIPS)*, 27.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10696–10706.

Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2022. Visual attention network. *arXiv preprint arXiv:2202.09741*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *In Proceedings of the Advances in neural information processing systems (NeurIPS)*, 30.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 7132–7141.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, 694–711.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 4401–4410.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. 2019. Controllable text-to-image generation. *In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Liao, W.; Hu, K.; Yang, M. Y.; and Rosenhahn, B. 2022. Text to image generation with semantic-spatial aware GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18187–18196.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision (ECCV)*, 740–755.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1505–1514.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016a. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 49–58.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016b. Generative adversarial text to image synthesis. In *Proceedings of the International conference on machine learning (ICML)*, 1060–1069.

Ruan, S.; Zhang, Y.; Zhang, K.; Fan, Y.; Tang, F.; Liu, Q.; and Chen, E. 2021. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 13960–13969.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *In Proceedings of the Advances in neural information processing systems (NeurIPS)*, 29.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673–2681.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2818–2826.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16515–16525.

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

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 34: 24261–24272.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *In Proceedings of the Advances in neural information processing systems (NeurIPS)*, 30.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22–31.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2256–2265.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 1316–1324.

Ye, S.; Liu, F.; and Tan, M. 2022. Recurrent Affine Transformation for Text-to-image Synthesis. *arXiv preprint arXiv:2204.10482*.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 833–842.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 5907–5915.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1947–1962.

Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; and Sun, T. 2022. Towards Language-Free Training for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17907–17917.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5802–5810.