DF-GAN:

A simple and effective baseline for Text-to-Image Synthesis

Phạm Bùi Nhật Huy 20521410
Nguyễn Thị Như Vân 20520855

01

Introduction

Introduction

The last few years have witnessed the great success of

Generative Adversarial Networks (GANs) for a variety of applications.

Among them, text-to-image synthesis is one of the most important applications of GANs.

Introduction

Two major challenges for text-to-image synthesis are:

- ***Authenticity*** of the generated image
- The ***semantic consistency*** between the given text and the generated image

Introduction

Due to the instability of the GAN model, most recent models adopt the stacked architecture as the backbone to generate high-resolution images.

- Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. ***Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks***. In Proceedings of the IEEE international conference on computer vision, pages 5907–5915, 2017

- Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. ***Stackgan++: Realistic image synthesis with stacked generative adversarial networks***. IEEE TPAMI, 41(8):1947–1962, 2018.

Introduction

They employ cross-modal attention to fuse text and image features and then introduce ***DAMSM network, cycle consistency, or Siamese network*** to ensure the textimage semantic consistency by extra networks.

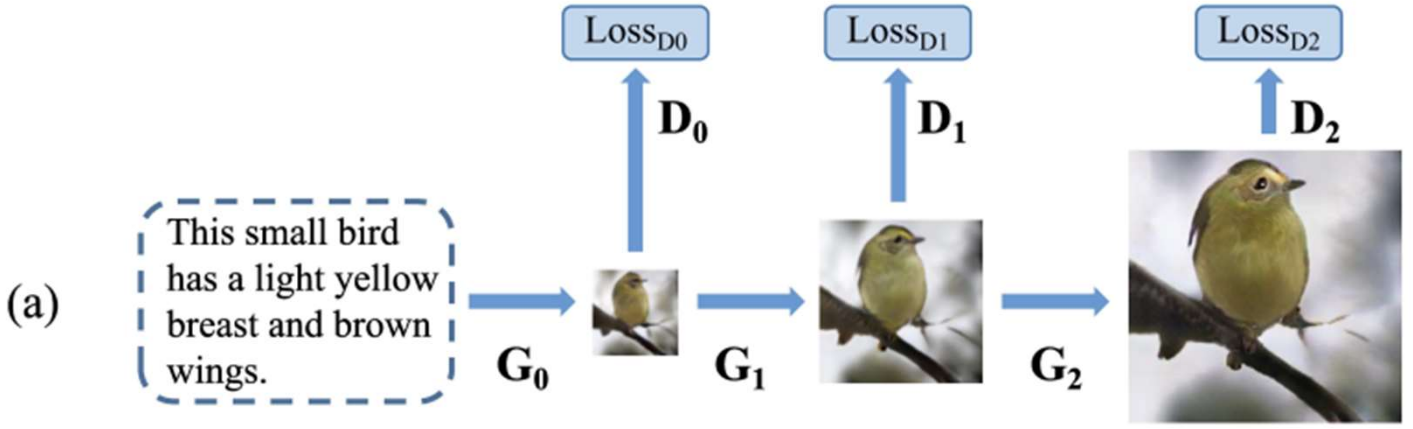
- DAMSM: Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. ***Attngan: Finegrained text to image generation with attentional generative adversarial networks*****.** In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1316–1324, 2018

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

- Siamese network: Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. ***Semantics disentangling for text-to-image generation***. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2327–2336, 2019.

Introduction

Although impressive results have been presented by previous works there still remain three problems.

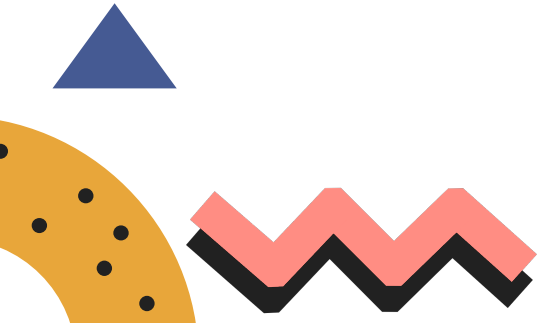


Introduction

For the first issue, ***replace the stacked backbone with a one-stage backbone***. It is composed of hinge loss and residual networks which stabilizes the GAN training process to synthesize high-resolution images directly.

For the second issue, we ***design a Target-Aware Discriminator*** composed of Matching-Aware Gradient Penalty (MA-GP) and One-Way Output to enhance the text-image semantic consistency.

For the third issue, we propose a ***Deep text-image Fusion Block (DFBlock)*** to fuse the text information into image features more effectively



02

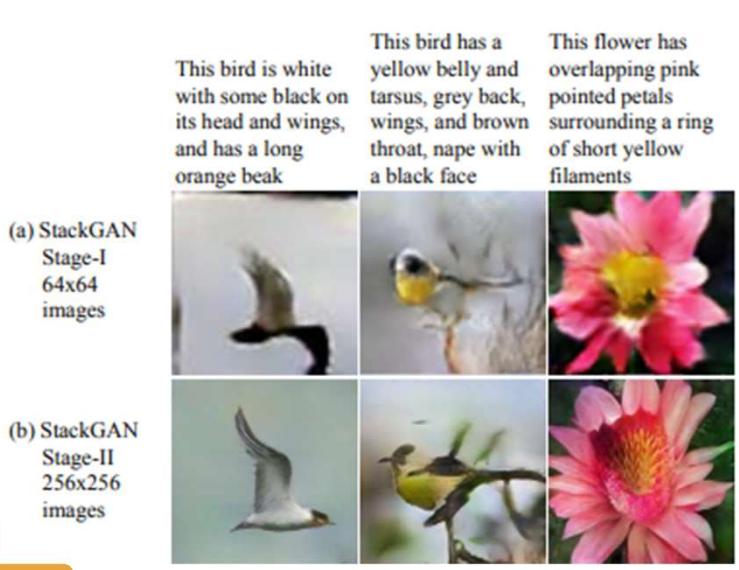
Related works

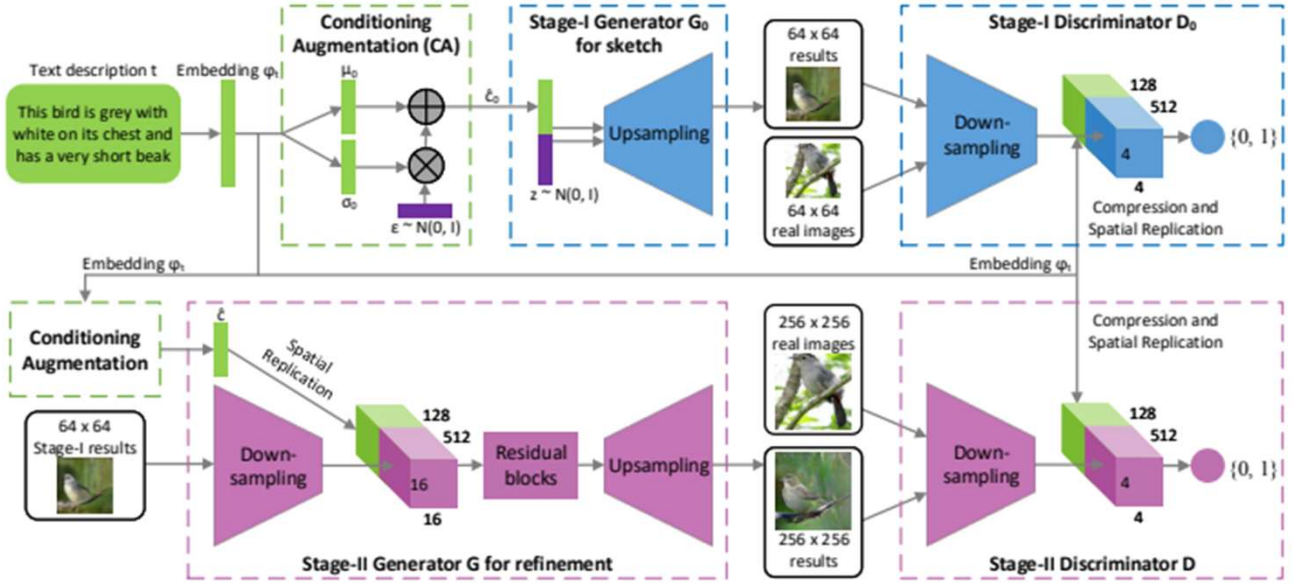
Related works

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

Related works

- ***Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks***. In Proceedings of the IEEE international conference on computer vision, pages 5907–5915, 2017





Related works

- ***Attngan: Finegrained text to image generation with attentional generative adversarial networks***. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1316–1324, 2018

Related works

- ***Mirrorgan: Learning text-to-image generation by redescription***. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1505–1514, 2019.

Related works

- ***SD-GAN: Semantics disentangling for text-to-image generation***. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2327–2336, 2019.

Related works

- ***DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis***. In Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition, pages 5802–5810, 2019

03

The proposed

DF-GAN

The proposed DF-GAN

3.1

Model overview

3.2

One-Stage Text-to-Image backbone

3.3

Target-Aware Discriminator (Matching-Aware Gradient-Penalty, One-Way Output)

3.4

Efficient text-image fusion

3.1

Model overview

Model overview

The proposed DF-GAN is composed of a generator, a discriminator, and a pre-trained text encoder

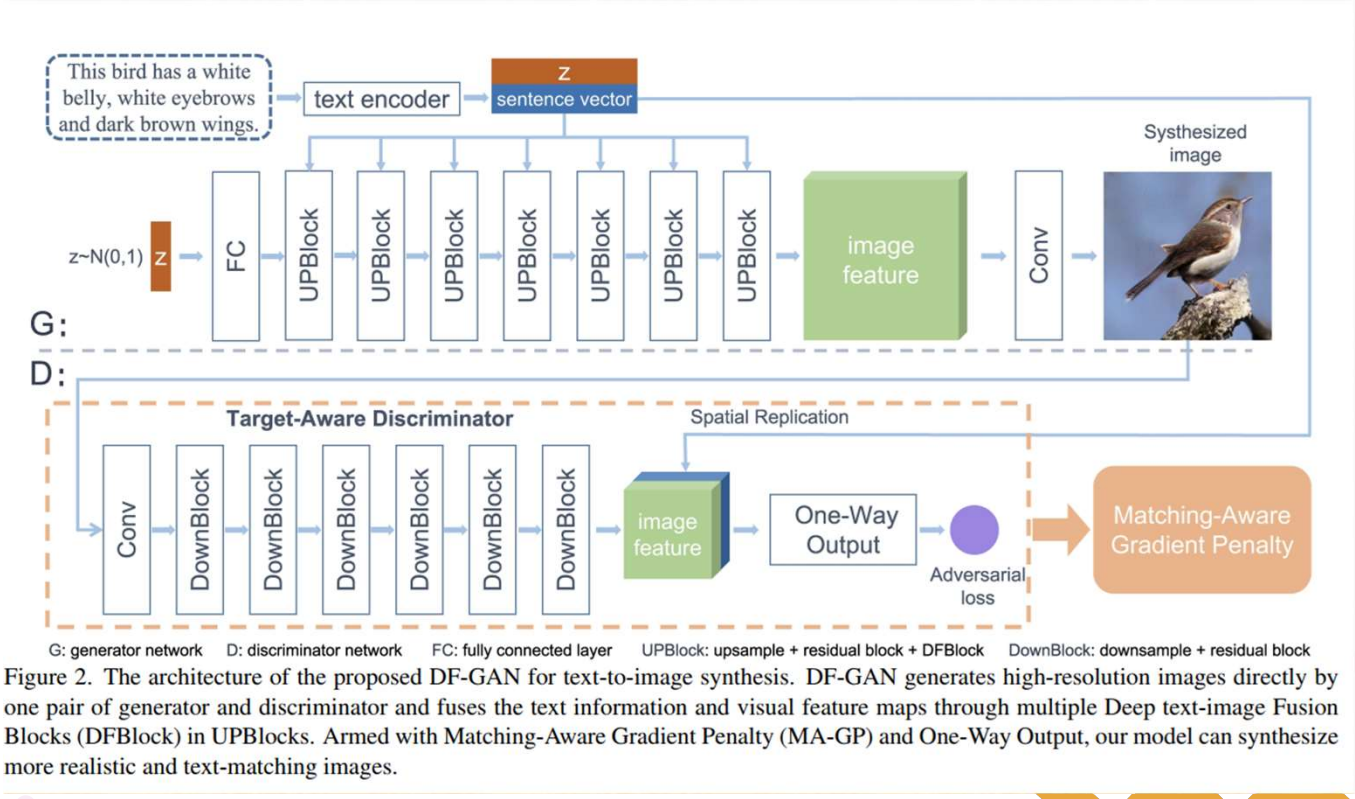


Figure 2. The architecture of the proposed DF-GAN for text-to-image synthesis. DF-GAN generates high-resolution images directly by one pair of generator and discriminator and fuses the text information and visual feature maps through multiple Deep text-image Fusion Blocks (DFBlock) in UPBlocks. Armed with Matching-Aware Gradient Penalty (MA-GP) and One-Way Output, our model can synthesize more realistic and text-matching images.

Model overview

The text encoder is a bi-directional Long Short-Term Memory (LSTM) that extracts semantic vectors from the text description. We directly use the pre-trained model provided by ***AttnGAN*.**

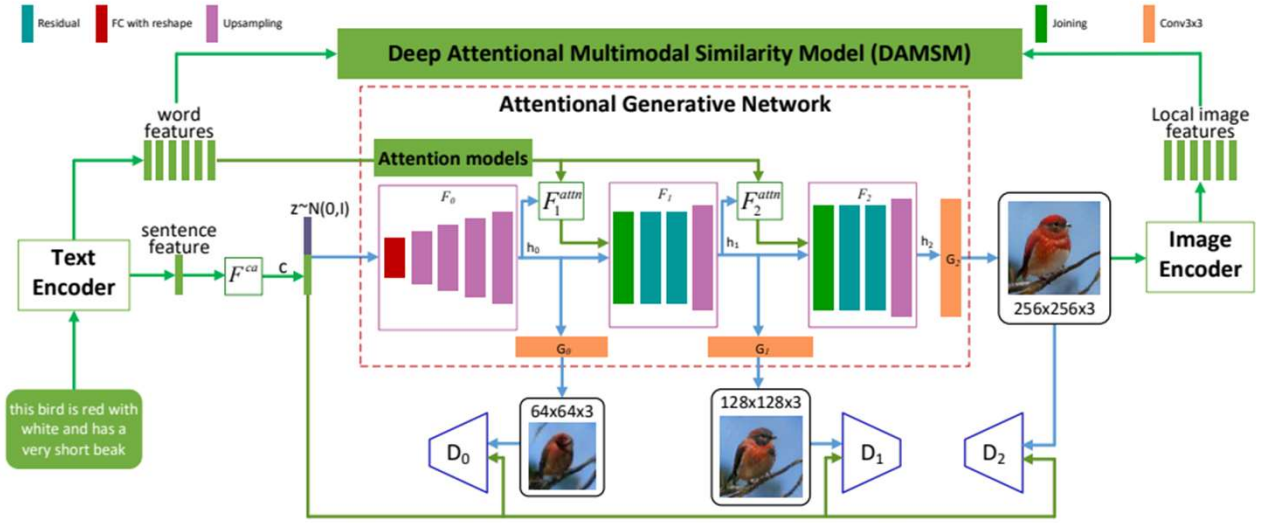


Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

3.2

One-Stage Text-to-Image backbone

One-Stage Text-to-Image backbone

Propose a one-stage text-to-image backbone that can synthesize high-resolution images directly by a single pair of generator and discriminator.

Employ the ***hinge loss*** to stabilize the adversarial training process.

As the single generator in our one-stage framework needs to synthesize high-resolution images from noise vectors directly, it ***must contain more layers*** than previous generators in stacked architecture.

To train these layers effectively, we introduce ***residual networks*** to stabilize the training of deeper networks.

One-Stage Text-to-Image backbone

The formulation of our ***one-stage method with hinge loss*** is as follows:

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

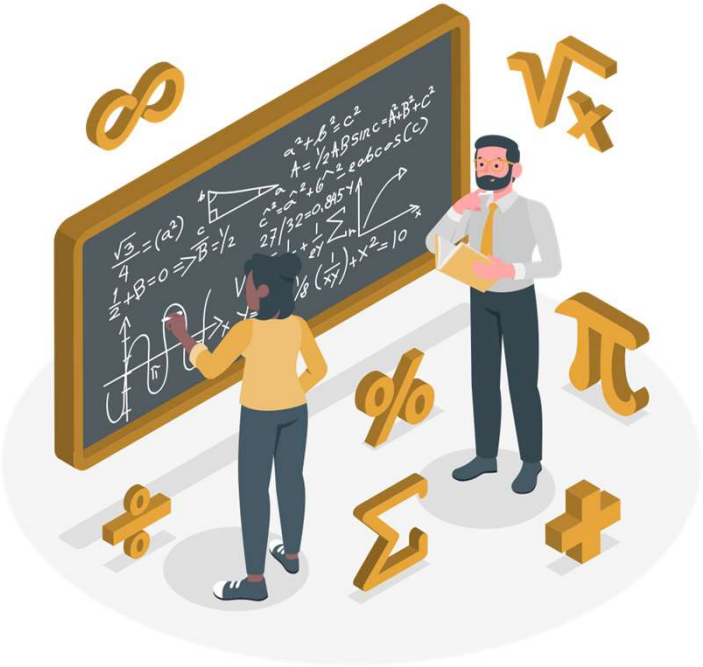
3.3

Target-Aware Discriminator (Matching-Aware Gradient-Penalty, One-Way Output)

Target-Aware Discriminator

In this section, we detailed the proposed ***Target-Aware Discriminator***, which is composed of Matching-Aware Gradient Penalty (MA-GP) and One-Way Output.

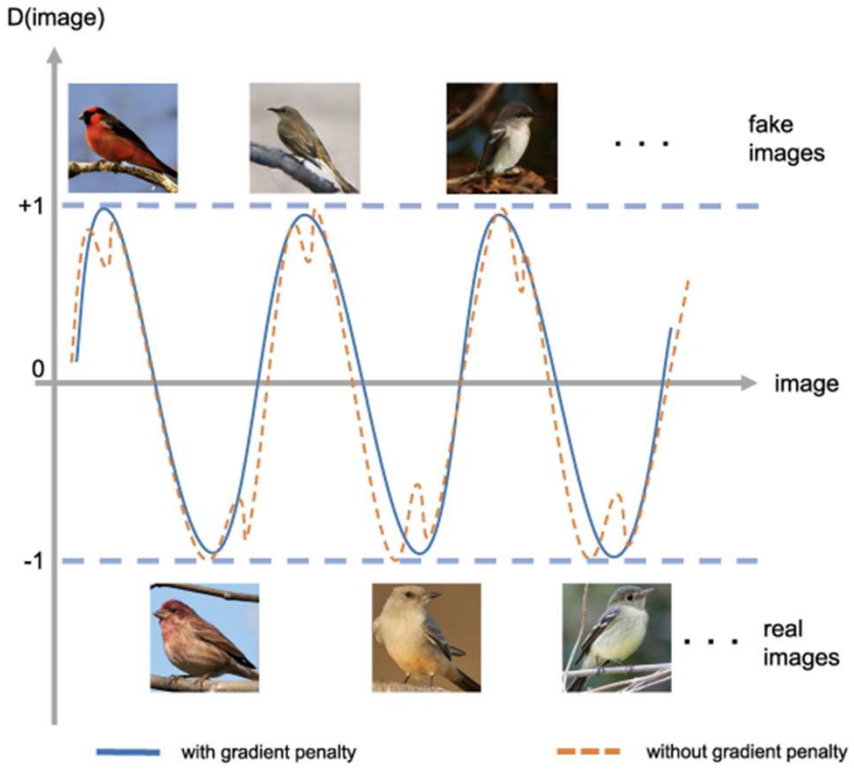
The Target-Aware Discriminator promotes the generator to synthesize more realistic and text-image semantic-consistent images



Matching-Aware Gradient Penalty

The Matching-Aware zero-centered Gradient Penalty (MA-GP) is newly designed strategy to enhance text-image semantic consistency.

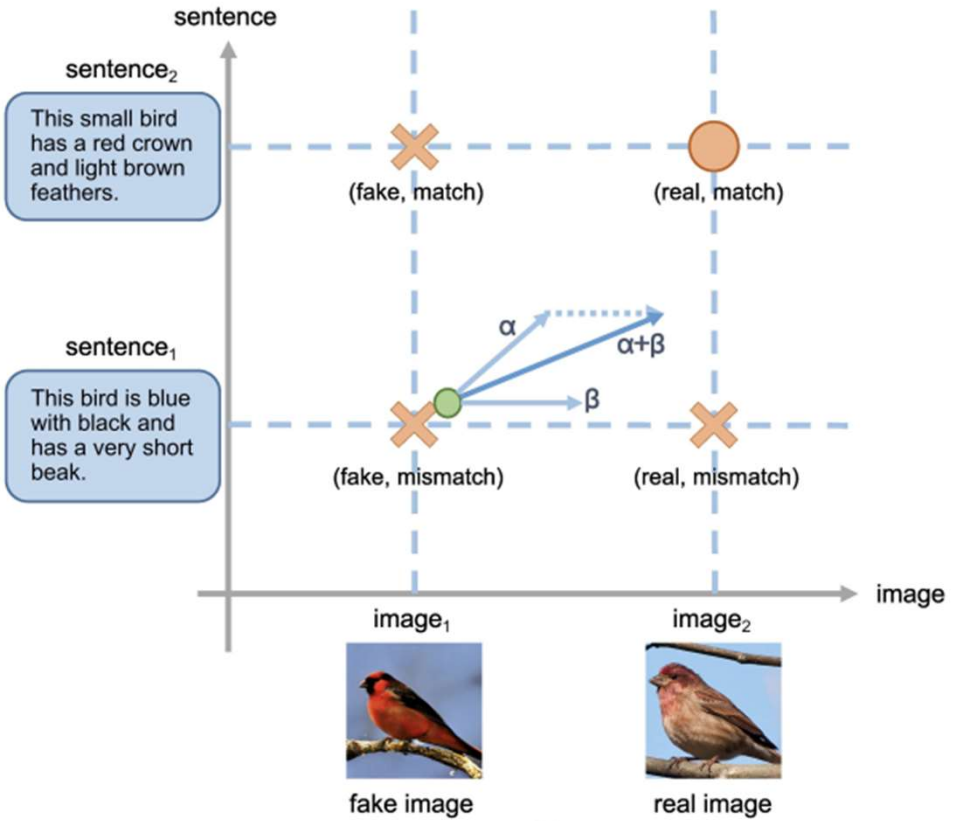
- Target data → low discriminator loss
- Synthesis images → high discriminator loss
- Hinge loss limits the discriminator loss between -1 and 1
- The GP on real data will reduce the gradient of the real data point and its vicinity
- The surface of the loss function around the real data point is then smoothed which is helpful for the synthetic data point to converge to the real data point



Matching-Aware Gradient Penalty

In text-to-image generation, the discriminator observes 4 kinds of inputs:

- synthetics images with matching text ***(fake, match)***
- synthetics images with mismatching text ***(fake, mismatch)***
- real images with matching text ***(real, match)***
- real images with mismatch text ***(real, mismatch)***



Matching-Aware Gradient Penalty

For text-visual semantic consistency, we tend to apply gradient penalty on the text-matching real data, the target of text-to-image synthesis.

Therefore, in MA-GP, ***the gradient penalty should be applied on real images with matching text.***

The whole formulation of our model with MA-GP is as follows:

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

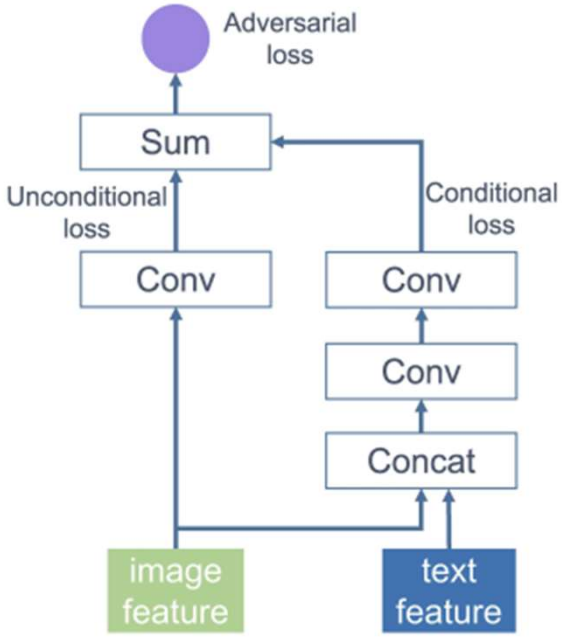
Matching-Aware Gradient Penalty

By using the MA-GP loss as a regularization on the discriminator
→ our model can better converge to the text-matching real data
→ synthesizing more text-matching images.

The discriminator is jointly trained in our network → prevents the generator from synthesizing adversarial features of the fixed extra network.

MA-GP does not incorporate any extra networks for text-image consistency and the gradients are already computed by back propagation process
→ the only computation introduced by our proposed MA-GP is the gradient summation, which is more computational friendly than extra networks.

One-Way Output



(a) Two-Way Output

- One determines whether the image is real or fake
- The other concatenates the image feature and sentence vector to evaluate text-image semantic consistency

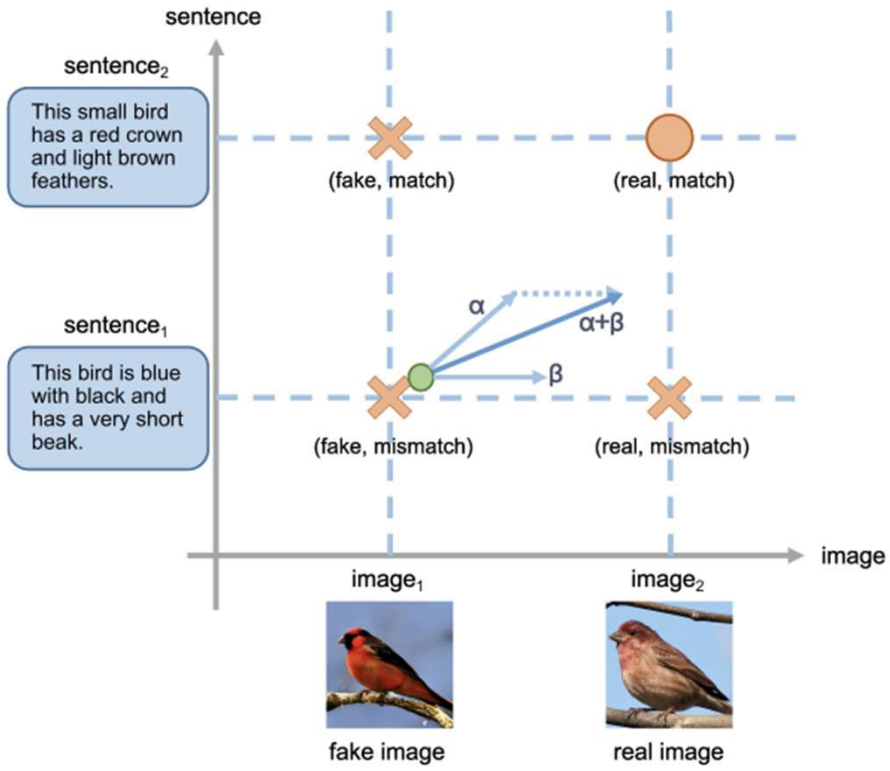
→ The unconditional loss and conditional loss are computed

One-Way Output

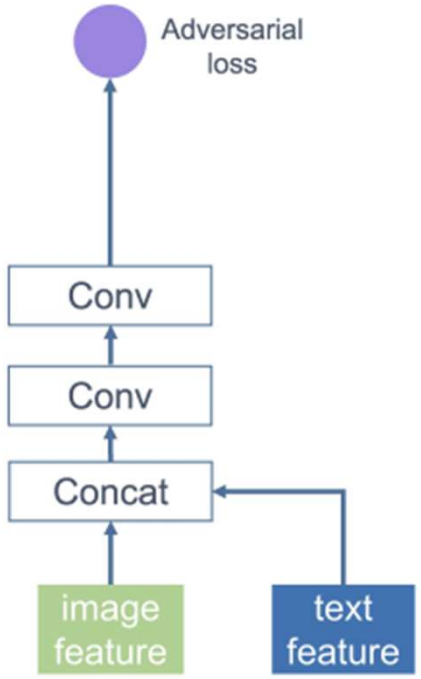
Two-Way Output weakens the effectiveness of MA-GP and slows down the convergence of the generator.

- α pointing to the real and matching inputs after back propagation
- the unconditional loss gives a gradient β only pointing to the real images

The target of the generator: synthesize real and text-matching images, the final gradient with deviation cannot well achieve text-image semantic consistency and slows down the convergence process of the generator → ***One-Way Output***



One-Way Output



(b) One-Way Output

The discriminator concatenates the image feature and sentence vector, then outputs only one adversarial loss through two convolution layers.

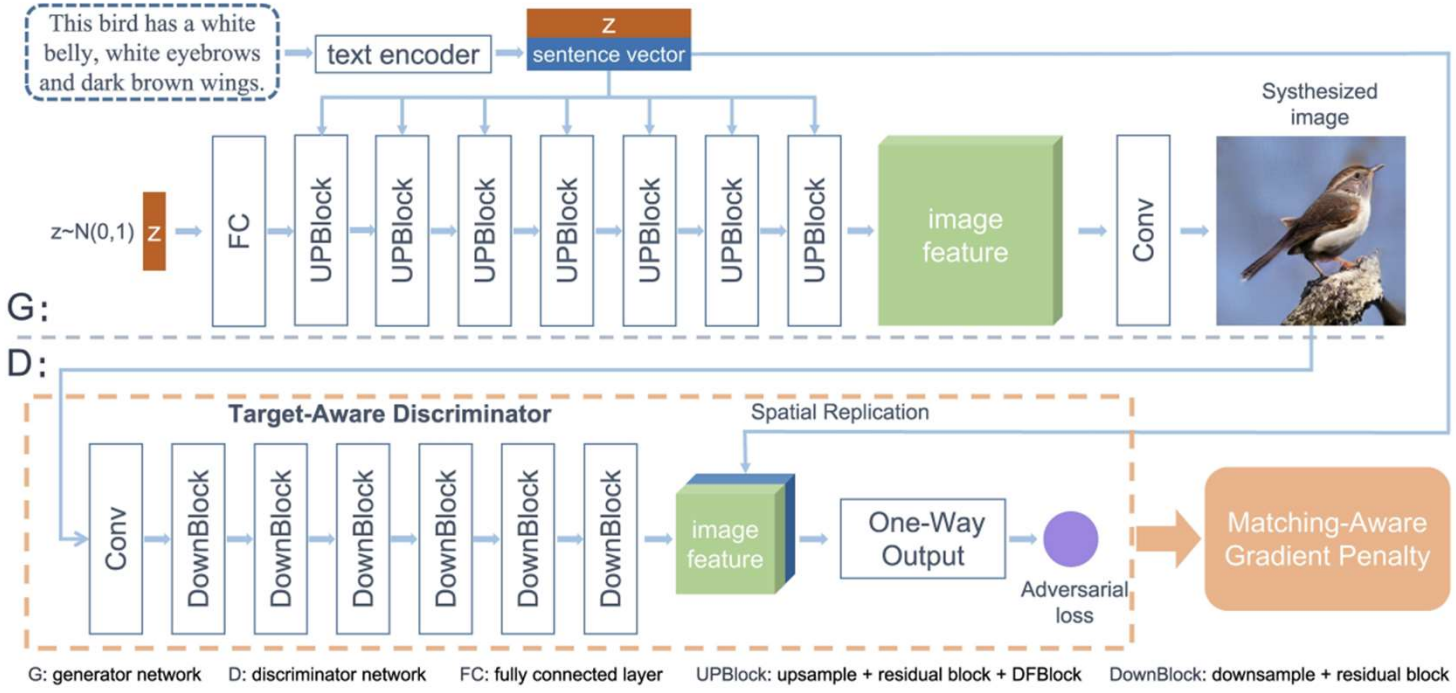
Through the One-Way Output, we are able to make the single gradient γ pointed to the target data points (real and match) directly, which optimize and accelerate the convergence of the generator.

3.4

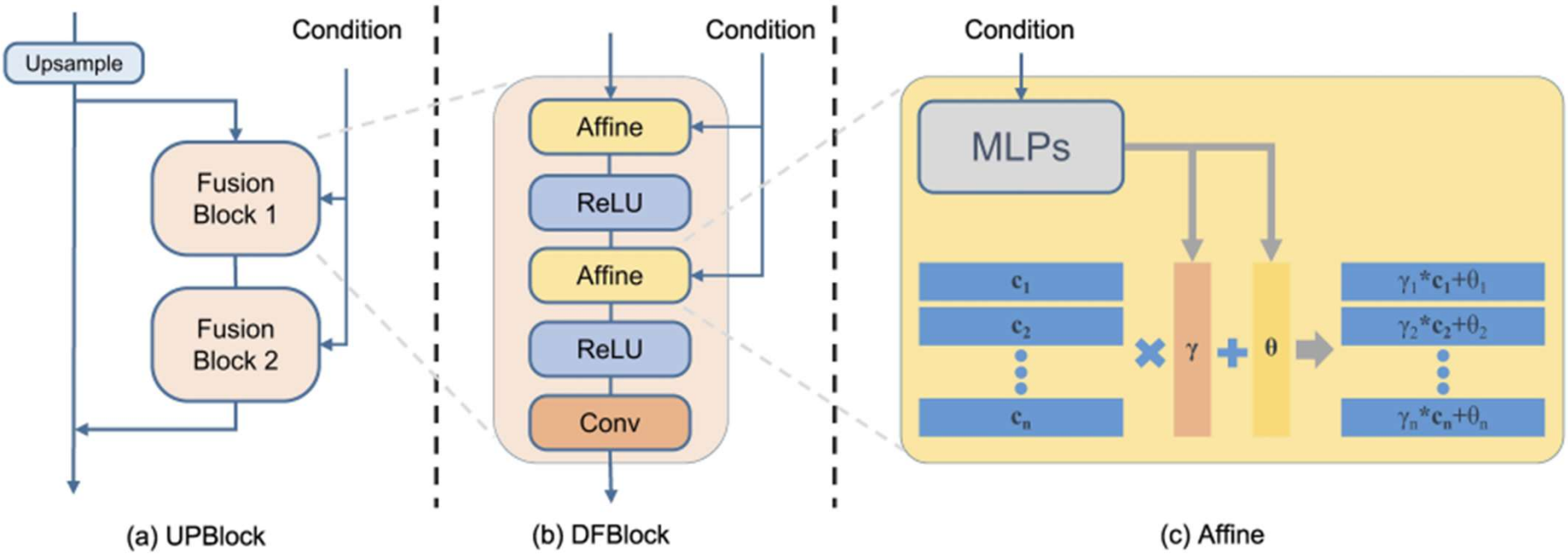
Efficient text-image fusion

Efficient Text-Image Fusion

The generator of our DF-GAN consists of 7 UPBlocks.



Efficient Text-Image Fusion



Efficient Text-Image Fusion

For Affine transformation, adopt two MLPs (Multilayer Perceptron) to predict the language-conditioned channel-wise scaling parameters γ and shifting parameters θ from sentence vector e, respectively:

$$\boldsymbol{\gamma} = MLP_1(\boldsymbol{e}), \qquad \boldsymbol{\theta} = MLP_2(\boldsymbol{e}). \qquad (3)$$

For a given input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, first conduct the channel-wise scaling operation on X with the scaling parameter γ, then apply the channel-wise shifting operation with the shifting parameter θ. Such a process can be expressed as follows:

$$AFF(\boldsymbol{x_i}|\boldsymbol{e}) = \gamma_i \cdot \boldsymbol{x_i} + \theta_i, \qquad (4)$$

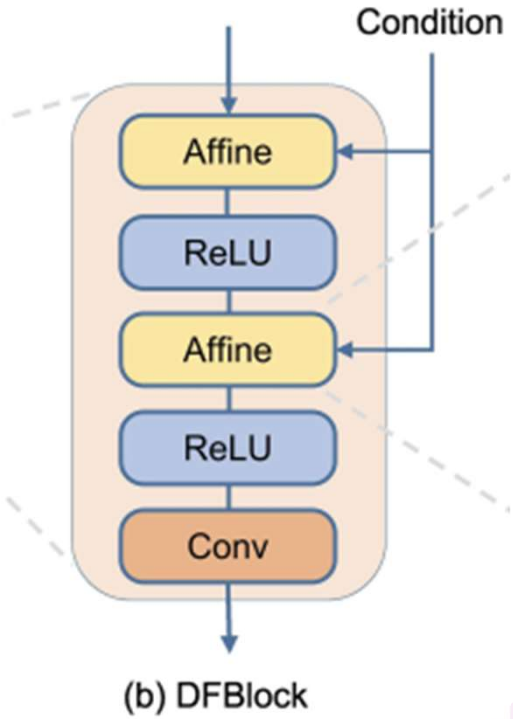
Efficient Text-Image Fusion

The Affine layer expands the conditional representation space of the generator. **However, the Affine Transformation is a linear transformation for each channel.**

→ ***Limit the effectiveness of text-image fusion process.***

Solution: add a ReLU layer between 2 Affine layer

-> Brings the nonlinearity into the fusion process



Efficient Text-Image Fusion

DFBlock is partly inspired by ***Conditional Batch Normalization (CBN)*** and ***Adaptive Instance Normalization (AdaIN)*** which contain the Affine transformation.

- Harm De Vries, Florian Strub, Jer´ emie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. ***Modulating early visual processing by language.*** In Advances in Neural Information Processing Systems, pages 6594–6604, 2017
- Xun Huang and Serge Belongie. ***Arbitrary style transfer in real-time with adaptive instance normalization***. In Proceedings of the IEEE International Conference on Computer Vision, pages 1501–1510, 2017
- Sergey Ioffe and Christian Szegedy. ***Batch normalization: Accelerating deep network training by reducing internal covariate shift***. In International Conference on Machine Learning, 2015.

Efficient Text-Image Fusion

However, ***both CBN and AdaIN employ normalization layers***, which transform the feature maps into the normal distribution.

It generates an ***opposite effect*** to the Affine Transformation which is expected to ***increase the distance between different samples***

→ **remove the normalization process**

Efficient Text-Image Fusion

With the deepening of the fusion process, the DFBlock brings ***two main benefits*** for text-to-image generation:

- It makes the generator more fully exploit the text information when fusing text and image features
- Deepening the fusion process enlargers the representation space of the fusion module, which is beneficial to generate semantic consistent images from different text descriptions

Compared with previous text-to-image GANs, the proposed DFBlock makes our model no longer consider the limitation from image scales when fusing the text and image features because existing text-to-image GANs generally employ the cross-modal attention mechanism which suffers a rapid growth of computation cost along with the increase of image size.

04

Experiments

Experiments

Dataset

- The CUB bird dataset contains 11,788 images belonging to 200 bird species. Each bird image has ten language descriptions *(published by two universities: California Institute of Technology and University of California)*
- The COCO dataset contains 80k images for training and 40k images for testing. Each image in this dataset has five language descriptions.

➢ *Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick.* ***Microsoft coco: Common objects in context.*** In European conference on computer vision, *pages 740–755. Springer, 2014*

Experiments

Training details

Optimize network using Adam:

- $\beta_1 = 0.0$
- $\beta_2 = 0.9$
- The learning rate is set to 0.0001 for the generator and 0.0004 for the discriminator according to Two Timescale Update Rule (TTUR)

Experiments

Evaluation details

To evaluate the performance of our network:

- **Inception Score (IS):** IS computes the Kullback-Leibler (KL) divergence between a conditional distribution and marginal distribution. Higher IS means higher quality of the generated images, and each image clearly belongs to a specific class.
- **Frechet Inception Distance (FID):** computes the Frechet distance between the distribution of the synthetic images and real-world images in the feature space of a pre-trained Inception v3 network. Contrary to IS, more realistic images have a lower FID.
- **Number of parameters (NoP):** to compare the model size with current methods.

Experiments

Quantitative evaluation

Compare the proposed method with several state-of-the-art methods, including **StackGAN, StackGAN++, AttnGAN, MirrorGAN, SD-GAN** and **DM-GAN**, which have achieved the remarkable success of text-to-image synthesis by using stacked structures.

Also compared with more recent models (recent models always use extra knowledge or supervisions). For example, **CPGAN** uses the extra pretrained YOLO-V3, **XMC-GAN** uses the extra pretrained VGG-19 and Bert, **DAEGAN** uses extra NLTK POS tagging and manually designs rules for different datasets, and **TIME** uses extra 2-D positional encoding.

Experiments

Quantitative evaluation

Table 1. The results of IS, FID and NoP compared with the state-of-the-art methods on the test set of CUB and COCO.

Model	CUB		COCO	
	IS ↑	FID ↓	FID ↓	NoP ↓
StackGAN [56]	3.70	-	-	-
StackGAN++ [57]	3.84	-	-	
AttnGAN [50]	4.36	23.98	35.49	230M
MirrorGAN [33]	4.56	18.34	34.71	-
SD-GAN [51]	4.67	-	-	-
DM-GAN [60]	4.75	16.09	32.64	46M
CPGAN [22]	-	-	55.80	318M
XMC-GAN [55]	-	-	9.30	166M
DAE-GAN [39]	4.42	15.19	28.12	98M
TIME [26]	4.91	14.30	31.14	120M
DF-GAN (Ours)	5.10	14.81	19.32	19M

Experiments

Quanlitative evaluation

Experiments

Ablation study

Table 2. The performance of different components of our model on the test set of CUB.

Architecture	IS ↑	FID ↓	SC ↑
Baseline	3.96	51.34	-
OS-B	4.11	43.45	1.46
OS-B w/ DAMSM	4.28	36.72	1.79
OS-B w/ MA-GP	4.46	32.52	3.55
OS-B w/ MA-GP w/ OW-O	**4.57**	**23.16**	**4.61**

Table 3. The performance of MA-GP GAN with different modules on the test set of CUB.

Architecture	IS↑	FID ↓
MA-GP GAN w/ Concat	4.57	23.16
MA-GP GAN w/ CBN	4.81	18.56
MA-GP GAN w/ AdaIN	4.85	17.52
MA-GP GAN w/ AFFBLK	4.87	17.43
MA-GP GAN w/ DFBLK (DF-GAN)	**5.10**	**14.81**

Experiments

Re-train: details

Dataset: CUB Bird with 11788 images and 10 language descriptions for each image

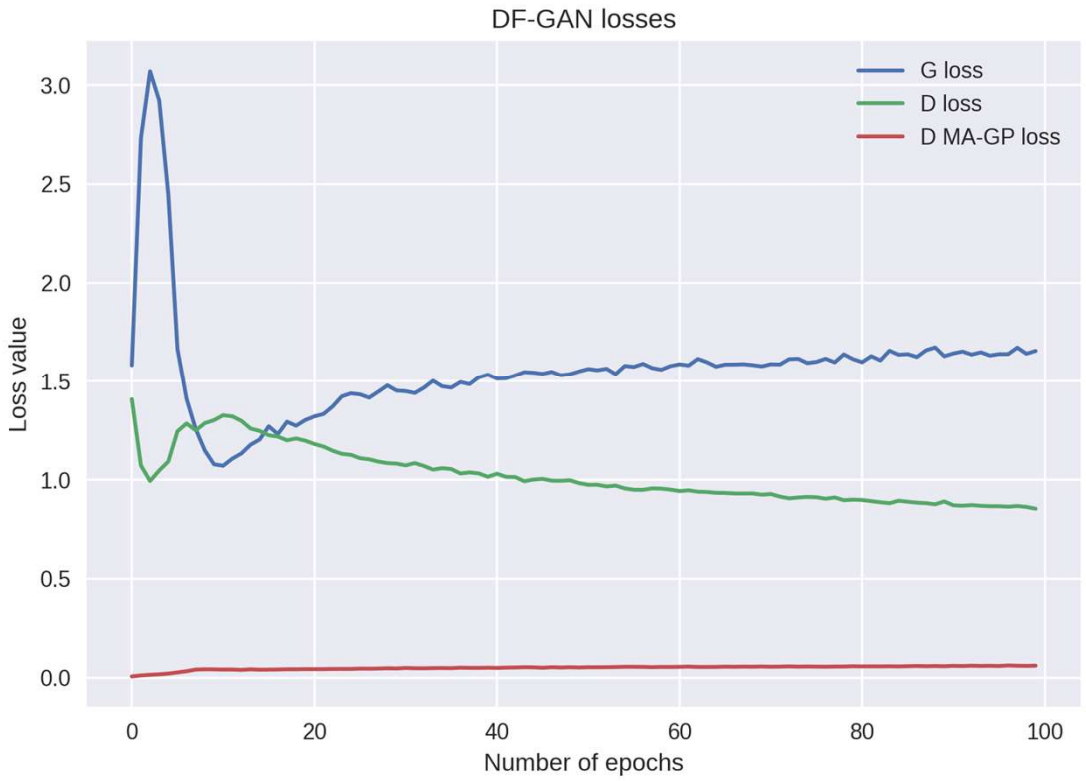
- Training: 8855 images and 88550 descriptions
- Testing: 2933 images and 29330 descriptions

Set up:

- GPU: T4
- Batch size: 32
- Time: ~15mins/epoch
- Number of epochs: 310 (paper: 600 epochs)

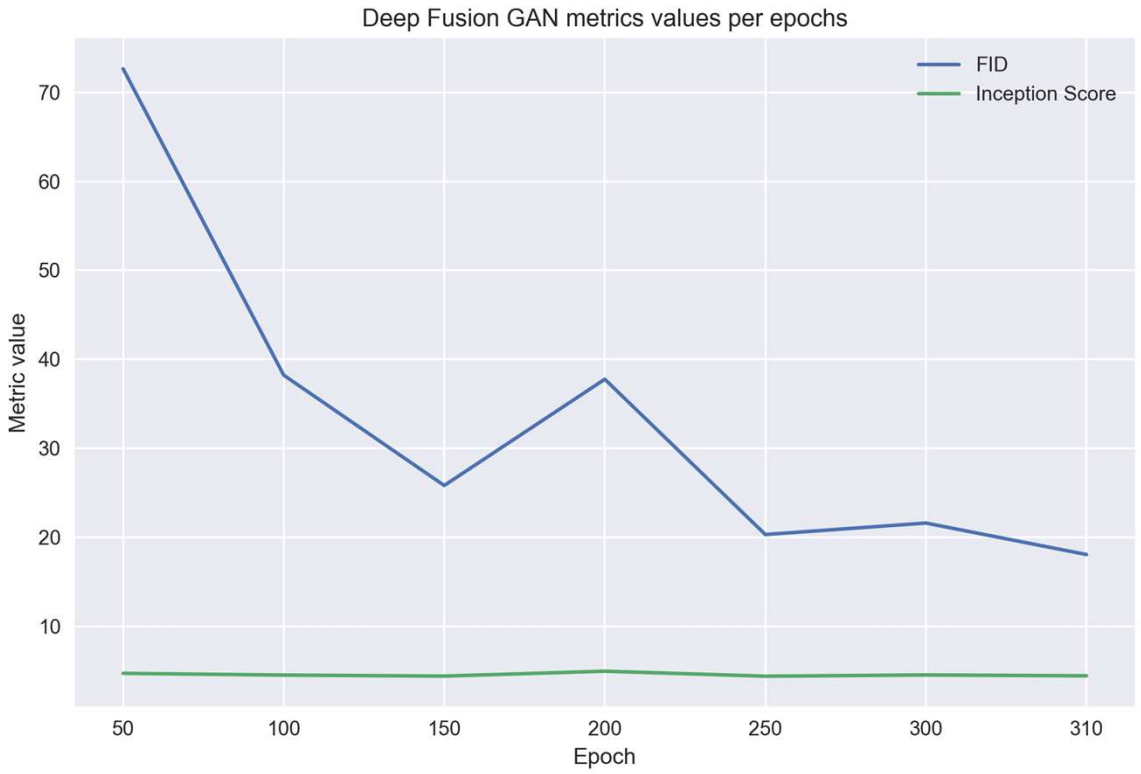
Experiments

Re-train: losses for first 100 epoch



Experiments

Re-train: metrics per epoch



Experiments

Re-train: final metrics

	Ours	Paper's
IS	4.43	5.10
FID	18.10	21.42

Experiments

Visualizing samples

“this bird has a white belly and breast, with a blue crown and nape”

Experiments

Visualizing samples

“A bird with a brown and black wings and a black stripe by its eyes”

Experiments

Visualizing samples

A small red bird has grey wings

A small red bird has grey ***long*** wings

05

Conclusion

Conclusion

Advantages

- One-stage text-to-image backbone that can synthesize high-resolution images directly without entanglements between different generators.
- Target-Aware Discriminator composed of Matching-Aware Gradient Penalty (MAGP) and One-Way Output can further enhance the text-image semantic consistency without introducing extra networks.
- Deep text-image Fusion Block (DFBlock) fully fuses text and image features more effectively and deeply.
- Extensive experiment results demonstrate that our proposed DF-GAN significantly outperforms current state-of-the-art models on the CUB dataset and more challenging COCO dataset

Conclusion

Disadvantages and limitations

- First, model only introduces the sentence-level text information, which limits the ability of fine-grained visual feature synthesis.

- Second, introducing pre-trained large language models to provide additional knowledge may further improve the performance.

Thank for listening!