Authors:
Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiaoyuan Jing, Fei Wu, Bingkun Bao

DF-GANs

Text to image synthesis

Supervisor:
Dr. CKM

TA:
Mr. Prudhviraj Jeripothula

Presenter:
Aditya Agrawal
CS19B1003

Table of Content

Problem Statement

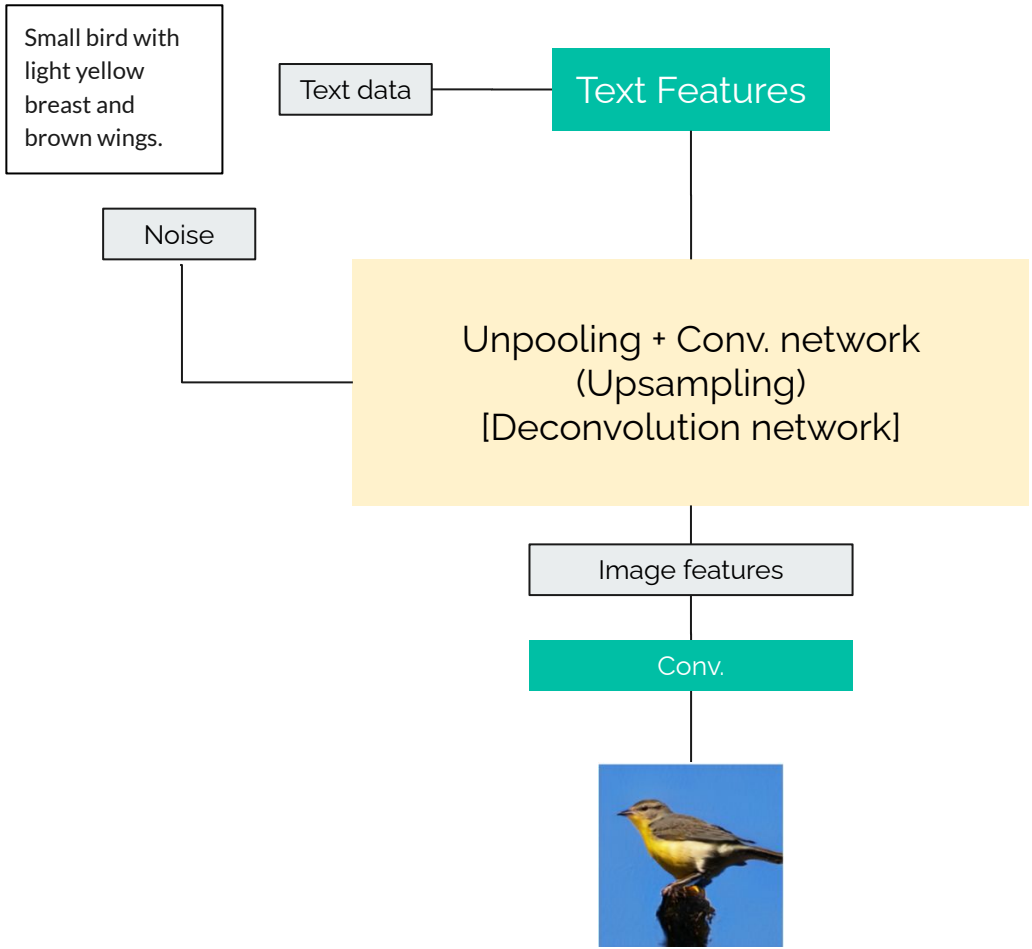
Generate high quality images from text descriptions which depicts the features.

Examples:

1. Small bird with light yellow breast brown wings. →

2. Small bird with red crown and light brown feathers. →

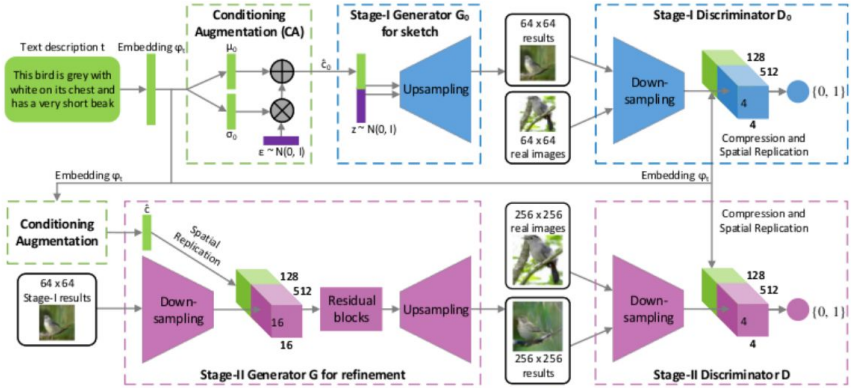
Initial Approach (no GANs for now)



Existing Methods

- Stacked-GAN
- Attn-GAN
- SD-GAN
- Obj-GAN
- DM-GAN

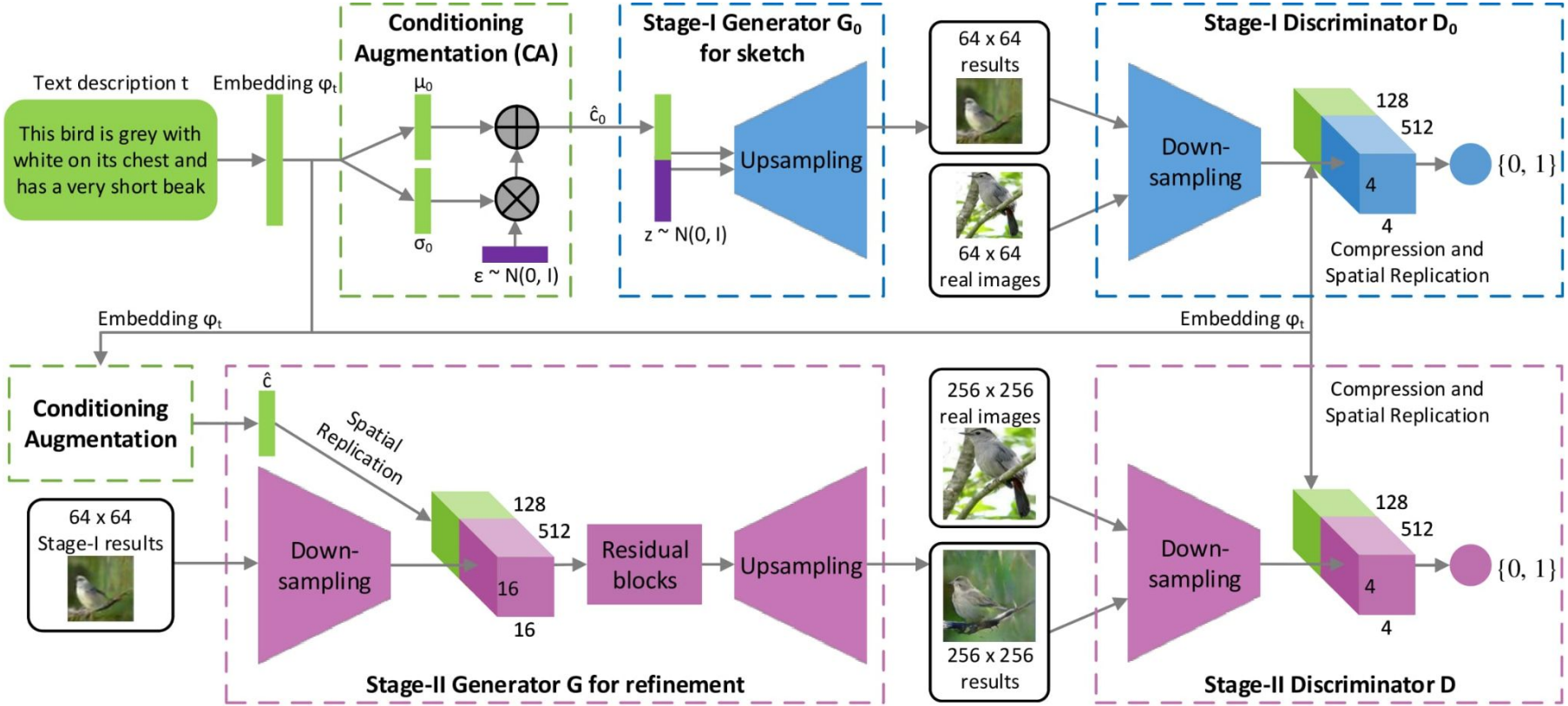
Stack-GAN



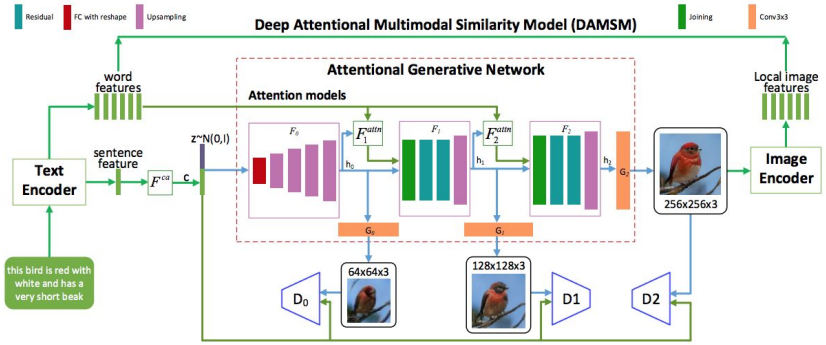
- Stacks multiple Gen. and Dis.

- Generated images looks unrealistic and refinements on coarse shapes.
- Text is provided to Gen. by concatenating text vector to input noises and intermediate features.

Stack GAN


Text description t
This bird is grey with white on its chest and has a very short beak
Embedding φ_t
Conditioning Augmentation (CA)
μ_0
σ_0
$\varepsilon \sim N(0, I)$
$\hat{c}_0$
Stage-I Generator G_0 for sketch
$z \sim N(0, I)$
Upsampling
64 x 64 results
64 x 64 real images
Stage-I Discriminator D_0
Down-sampling
128
512
4
4
{0, 1}
Compression and Spatial Replication
Embedding φ_t
Embedding φ_t
Conditioning Augmentation
$\hat{c}$
Spatial Replication
64 x 64 Stage-I results
Down-sampling
128
512
16
16
Residual blocks
Upsampling
Stage-II Generator G for refinement
256 x 256 real images
256 x 256 results
Compression and Spatial Replication
Down-sampling
128
512
4
4
{0, 1}
Stage-II Discriminator D

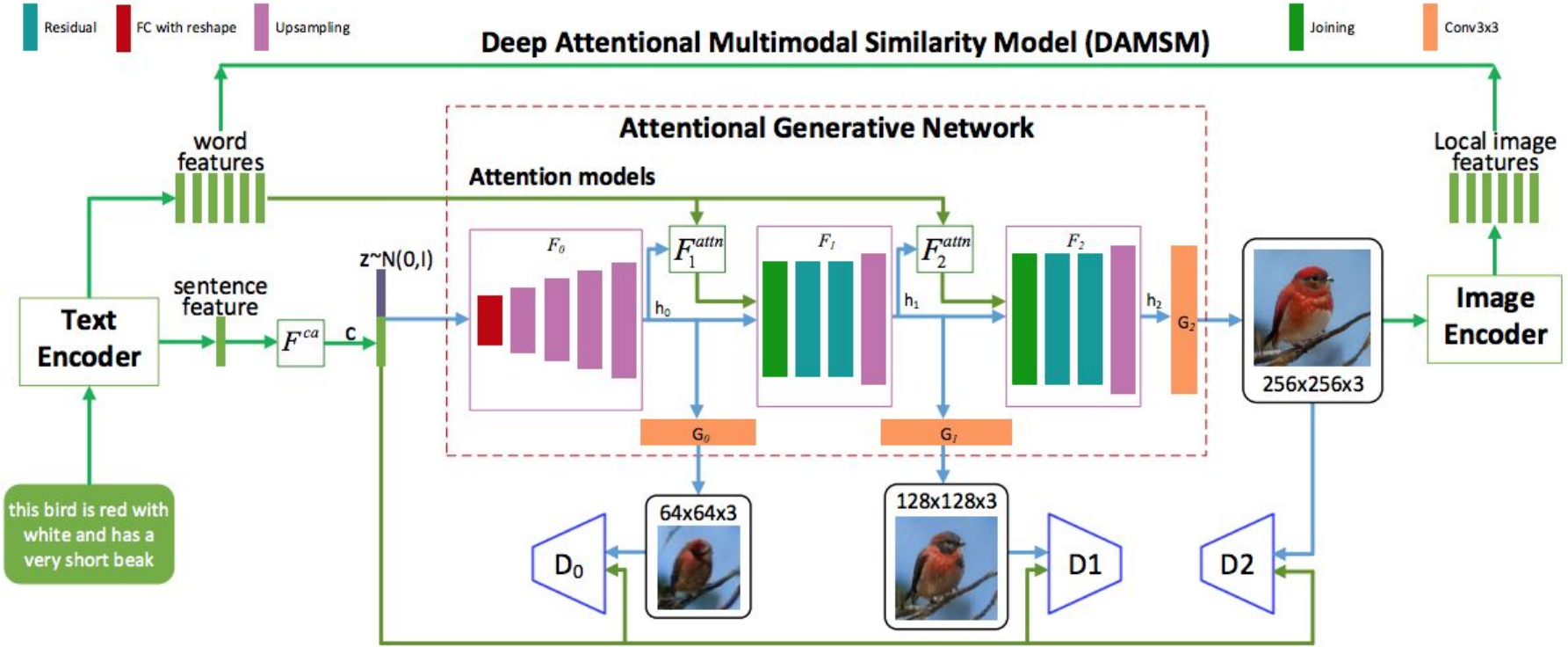
Attn-GAN



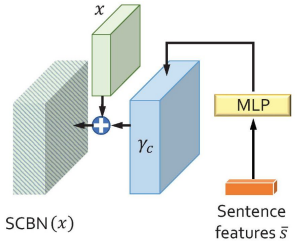
- Introduces cross-modal attention mechanism for more realistic images.
- Finds mapping between textual information and each pixel.

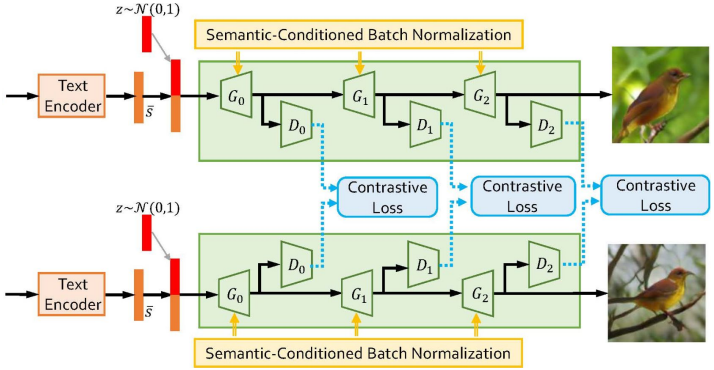
- Stacked architecture.
- Focuses on text information mostly (compared to DF-GAN).
- Bigger image, bigger network.

Attn-GAN


Residual
FC with reshape
Upsampling
Joining
Conv3x3
Deep Attentional Multimodal Similarity Model (DAMSM)
Attentional Generative Network
Attention models
word features
Local image features
sentence feature
z~N(0,I)
Text Encoder
Image Encoder
c
256x256x3
64x64x3
128x128x3
D0
D1
D2
this bird is red with white and has a very short beak

SD-GAN

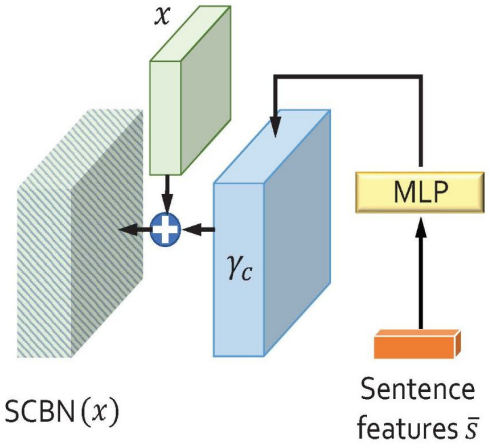


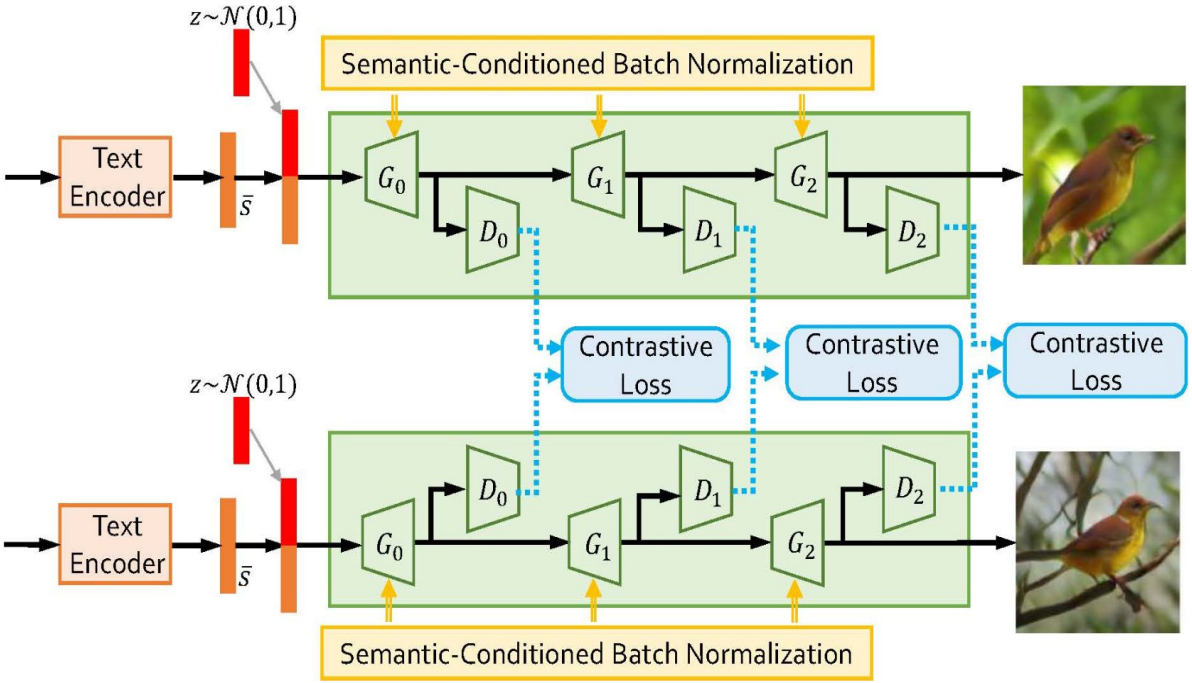


- Uses the Siamese structure to distill the semantic commons from texts for image generation consistency.
- Takes **contrastive loss** at multiple stages and try minimizing it.
- Conditional Batch Normalization for text-image fusion.

- Stacked architecture.
- Batch norm. doesn't use affine transformations effectively.

SD-GAN


x
γ_c
MLP
SCBN (x)
Sentence features $\bar{s}$


$z \sim \mathcal{N}(0,1)$
Semantic-Conditioned Batch Normalization
Text Encoder
$\bar{s}$
G_0
G_1
G_2
D_0
D_1
D_2
Contrastive Loss
Contrastive Loss
Contrastive Loss
$z \sim \mathcal{N}(0,1)$
Text Encoder
$\bar{s}$
G_0
G_1
G_2
D_0
D_1
D_2
Semantic-Conditioned Batch Normalization

Motivation

Applications: Media production, learning methods.

Text to image synthesis is a challenging task. Almost all the current available works employs a *stacked architecture* as the backbone which is inefficient and costly. The utilized text fusion methods, cross-modal att., batch norm, concatenating, are not much effective and introduces extra networks.

This model introduces a simpler architecture while producing better results.

Problems to solve

1

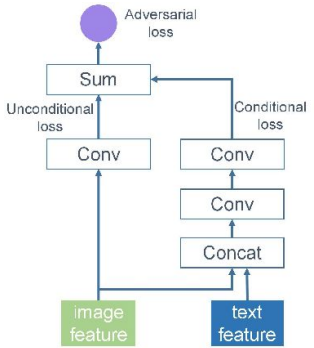
Use of stacked architecture
- Costly
- Dependence on initial stages
- Difference in scale of image cause instability in loss

2

Text-image fusion methods
- **Concatenation** is inefficient/naive
- **Cross-modal atten.** is based on spatial attention. (Image Size ↑ - Computation ↑)
- **Conditional Batch Norm.** as in SD-GAN

3

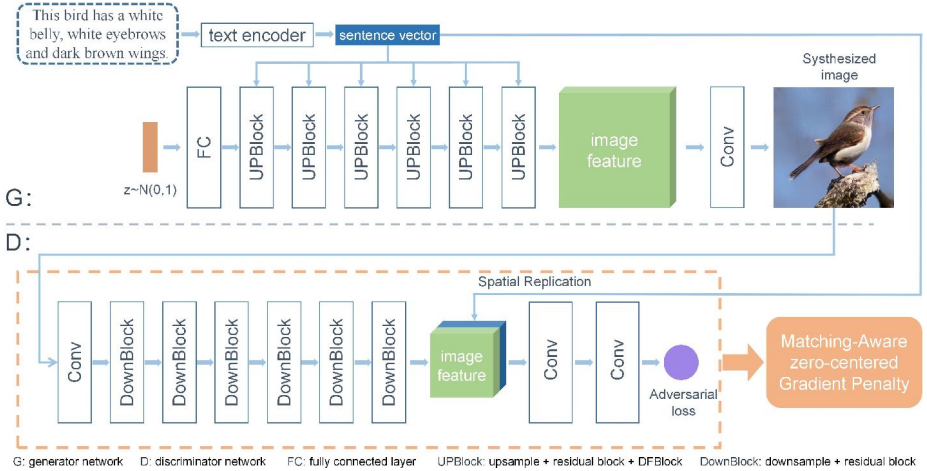
Two ways discriminator
- Use of extra network
- Increases training complexity
- Less efficient



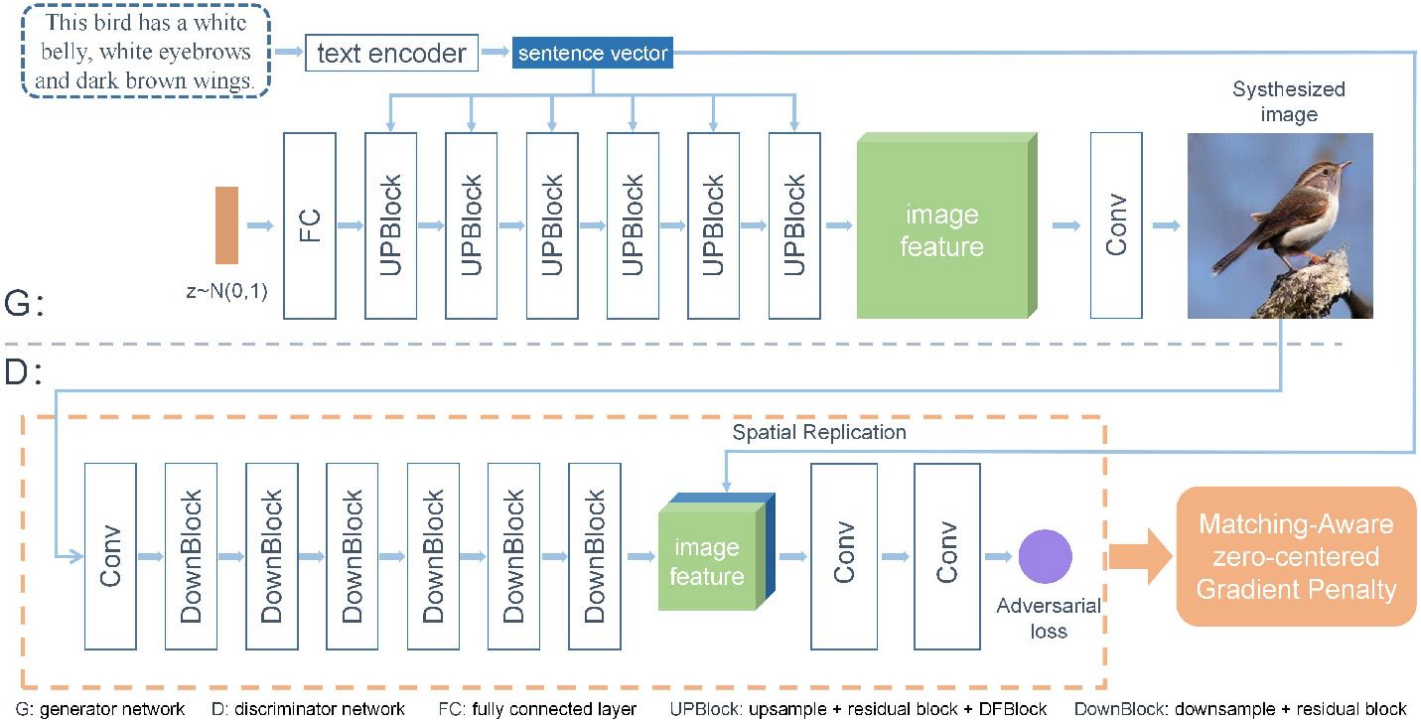
DF-GAN Model

The proposed model has the following features:

1. One stage text-to-image backbone.
2. A novel fusion module called deep text-image fusion block.
3. A novel target-aware discriminator:
   a. matching-aware gradient penalty
   b. 1-way discriminator output



DF-GAN


This bird has a white belly, white eyebrows and dark brown wings.
text encoder
sentence vector
Systhesized image
G:
z~N(0,1)
FC
UPBlock
UPBlock
UPBlock
UPBlock
UPBlock
UPBlock
image feature
Conv
D:
Spatial Replication
Conv
DownBlock
DownBlock
DownBlock
DownBlock
DownBlock
DownBlock
image feature
Conv
Conv
Adversarial loss
Matching-Aware zero-centered Gradient Penalty
G: generator network
D: discriminator network
FC: fully connected layer
UPBlock: upsample + residual block + DFBlock
DownBlock: downsample + residual block

One stage text-to-image backbone

Use of single Generator - Discriminator network.

Uses **hinge loss** to stabilize the training.

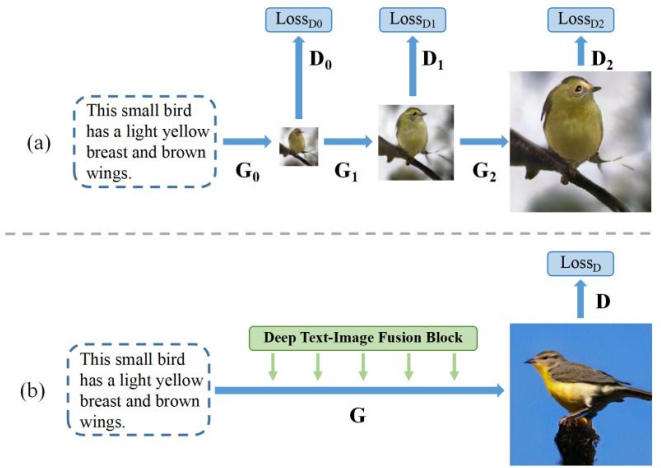
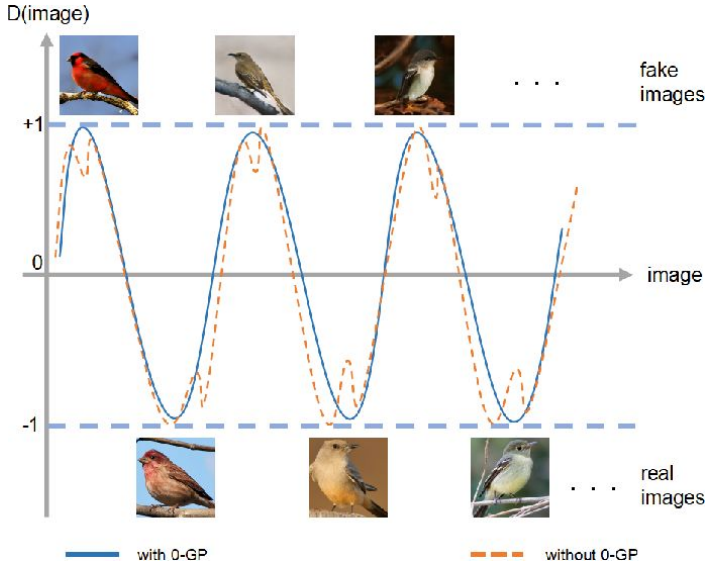


Figure 1: (a) Existing text-to-image models stack multiple generators and discriminators to generate high-resolution images. (b) Our proposed DF-GAN generates high-quality images directly and fuses the text and image features deeply by our deep text-image fusion blocks.

Target Aware Discriminator

Matching Aware Gradient Penalty

- Pushes real data points towards minimum of loss curve.
- Smoothens the surface for and around the real data points for smoother convergence.

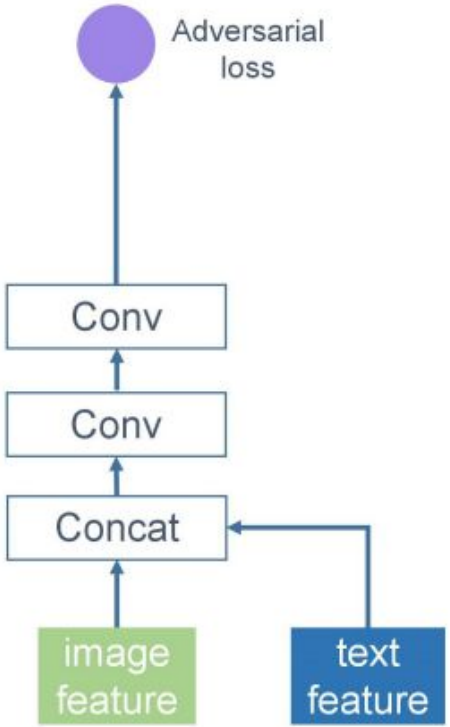


$$
\begin{aligned}
L_D = & - \mathbb{E}_{x \sim \mathbb{P}_r}[min(0, -1 + D(x, e))] \\
& - (1/2)\mathbb{E}_{G(z) \sim \mathbb{P}_g}[min(0, -1 - D(G(z), e))] \\
& - (1/2)\mathbb{E}_{x \sim \mathbb{P}_{mis}}[min(0, -1 - D(x, e))] \\
& + k\mathbb{E}_{x \sim \mathbb{P}_r}[(\|\nabla_x D(x, e)\| + \|\nabla_e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

Target Aware Discriminator

One-way output

- Reduces computational cost
- Unconditional loss deviates the Adversarial loss from the desired position



Target Aware Discriminator

One-way output

- Reduces computational cost
- Unconditional loss deviates the Adversarial loss from the desired position

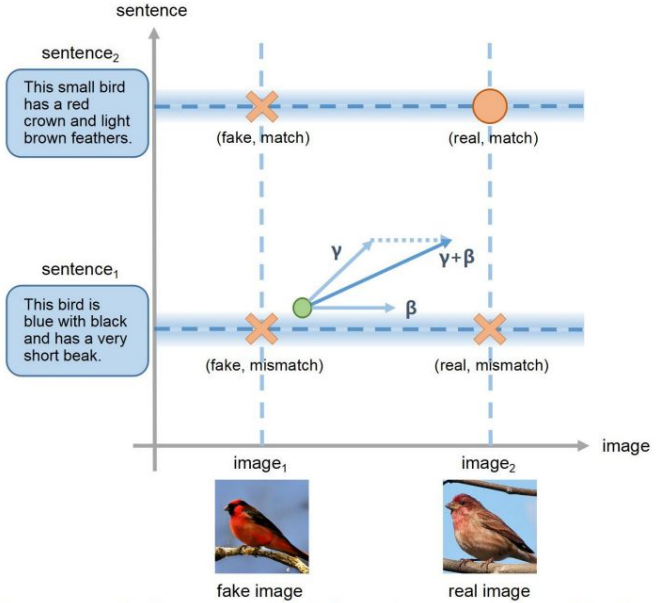
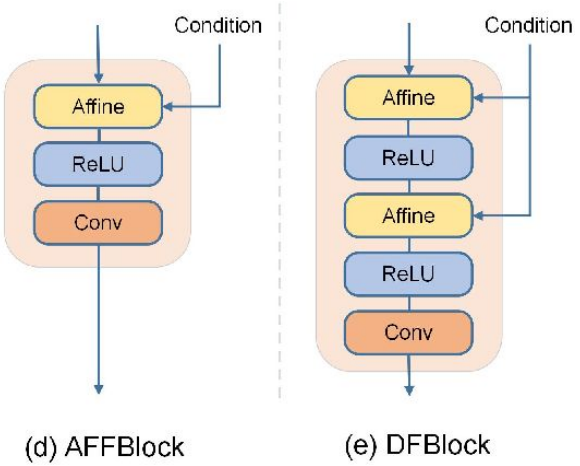


Figure 3: A diagram for Matching-Aware Gradient Penalty (MA-GP). The data point (real, match) marked by a circle should be applied MA-GP.

Text-image fusion

DF-Blocks

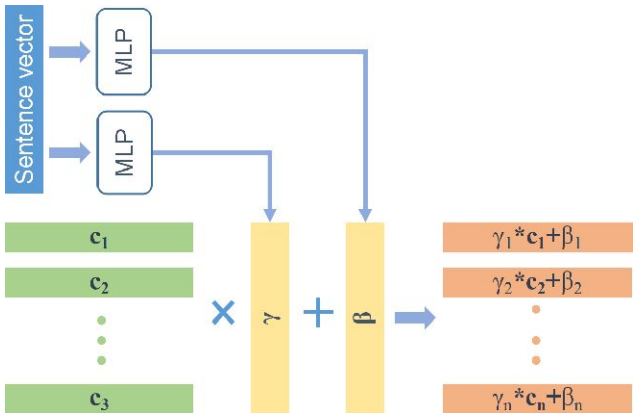
- Normalization of feature map is skipped.
- Affine Transformations are used.
- Affine+ReLU blocks are stacked together to form DFBlock.



Text-image fusion

DF-Blocks

- Normalization of feature map is skipped.
- Affine Transformations are used.
- Affine+ReLU blocks are stacked together to form DFBlock.



Experiments and Results

Data sets used:

- COCO: (80k training, 40k testing, 5 ld)
- CUB bird: (11788 img, 200 sp., 10 ld)

ld: Language Descriptions

Epochs:

- COCO : 121
- CUB-200 : 601

Optimizer: Adam ($\beta1=0.0$ and $\beta2=0.9$)

Learning Rate:

- Generator: 1e-4
- Discriminator: 4e-4

Evaluation Metric:

- COCO:
  - Frechet Inception Distance
  - Inception Score
- CUB birds:
  - Inception Score

Quantitative Results

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

Method	CUB-IS↑	CUB-FID↓	COCO-FID↓
AttnGAN [33]	4.36	23.98	35.49
MirrorGAN [23]	4.56	18.34	34.71
SD-GAN [34]	4.67	-	-
DM-GAN [40]	4.75	16.09	32.64
DF-GAN (Ours)	**5.10**	**14.81**	**21.42**

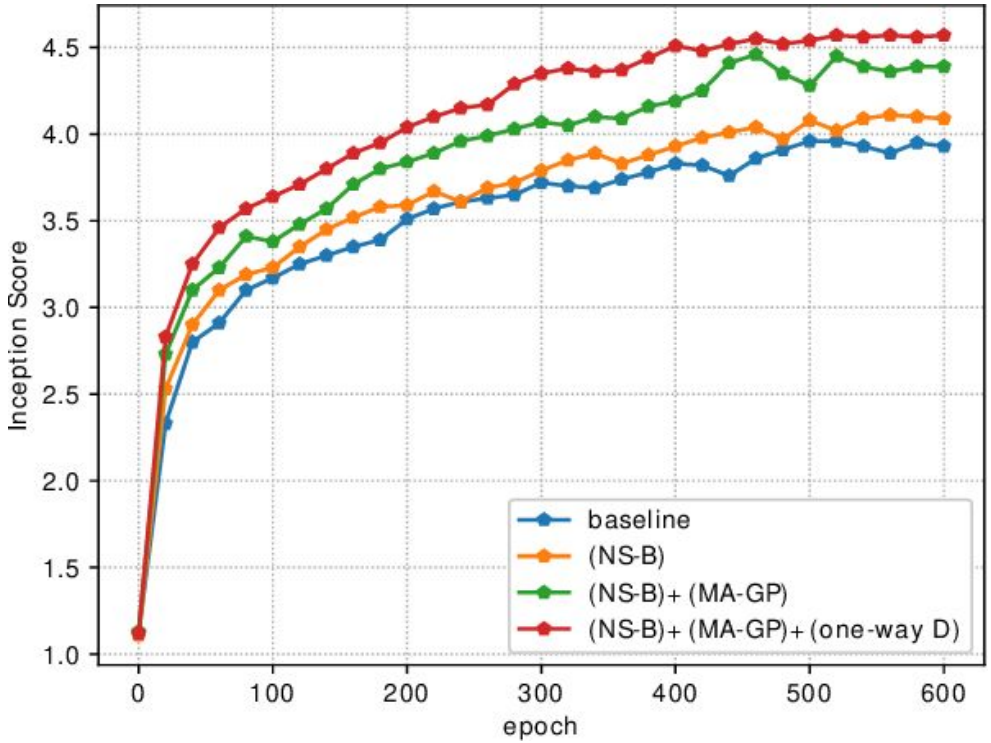
Qualitative Results

Figure 6: Examples of images synthesized by AttnGAN [33], DM-GAN [40], and our proposed DF-GAN conditioned on text descriptions from the test set of COCO and CUB datasets.

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/ MA-GP	4.46	32.52	3.55
OS-B w/ MA-GP w/ OW-O	**4.57**	**23.16**	**4.61**



Fusion Blocks

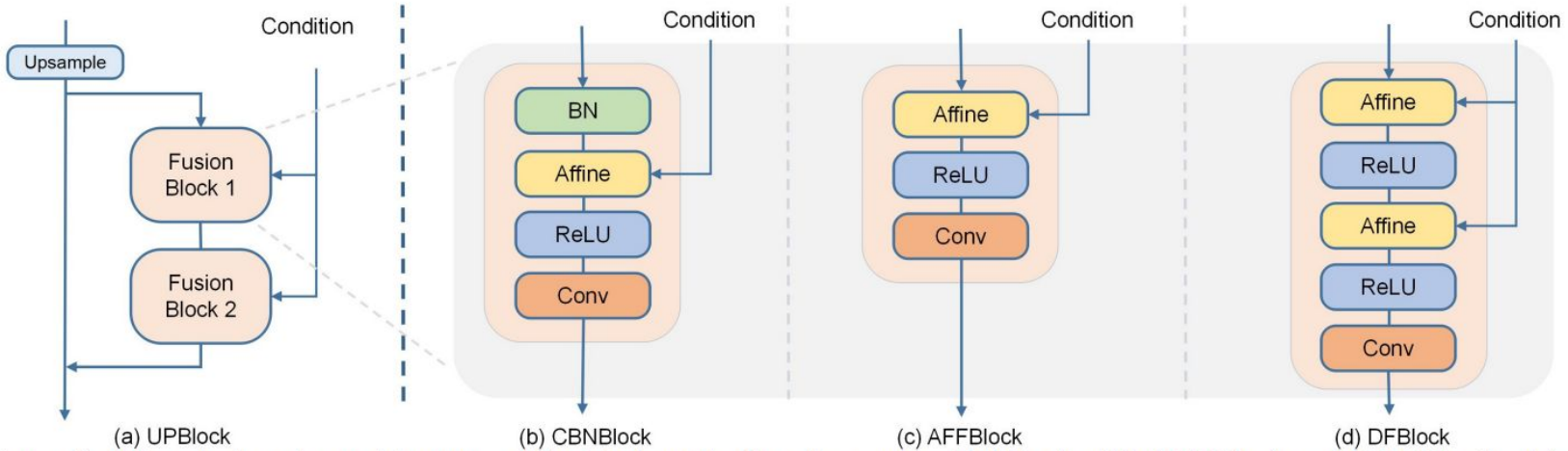
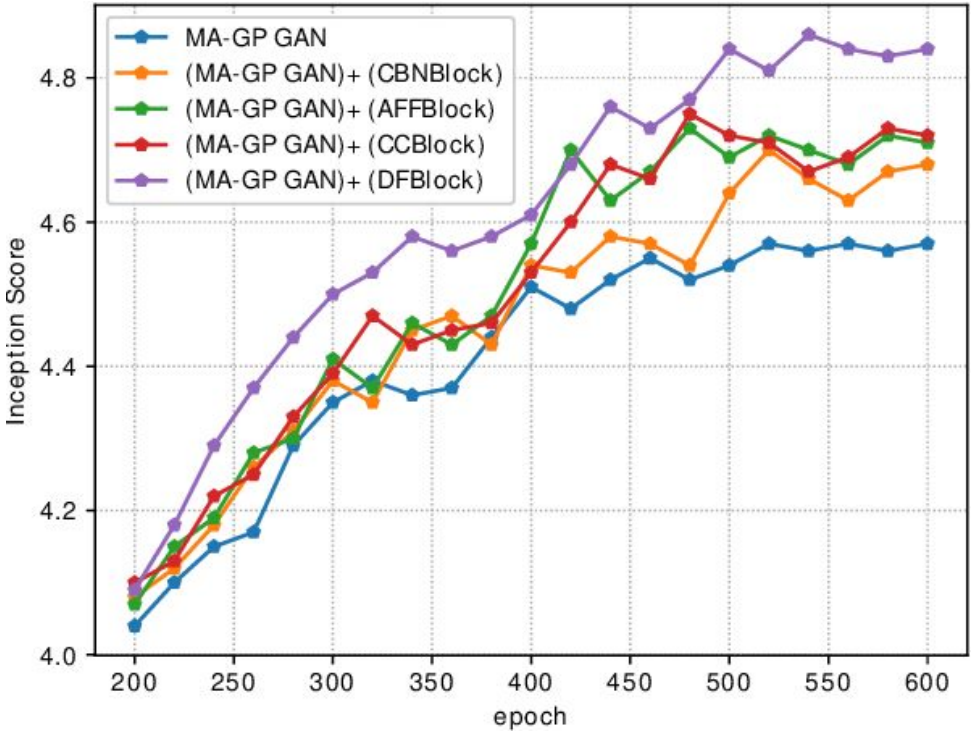


Figure 5: We redesign the architecture of the Fusion Block and compare DFBlock with AFFBlock and CBNBlock. (a) A typical UPBlock in the generator network. The UPBlock upsamples the image features and fuses text and image features by two Fusion Blocks. (b) The CBNBlock is a Fusion Block which employs the Conditional Batch Normalization to fuse text and image features. (c) AFFBlock is a simplified version of CBNBlock which removes the Batch Normalization layer. (d) The DFBlock is an enhanced version of AFFBlock, it deepens the text-image fusion process by stacking multiple Affine Transformations.

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**



Conclusion

DF-GAN model with following components:

- 1 stage backbone - efficient, more stable training, more realistic images..
- Novel target-aware discriminator - better convergence for generator, improved quality, and text semantic images
    - MAGP
    - 1 way output
- DF-Block - better text image fusion
- Results shows this model rank better on the CUB and COCO datasets.
- Rates better compared to other related work in IS and FID score.

Thank you.