Dept Of Computer Science & Engineering

Center For Data Science and Applied Machine Learning

Review 1

Team Details :

PES1UG19CS019 Abhishek Aditya BS, A Section

PES1UG19CS520 Supreeth G Kurpad, H Section

PES1UG19CS571 Vishal R, I Section

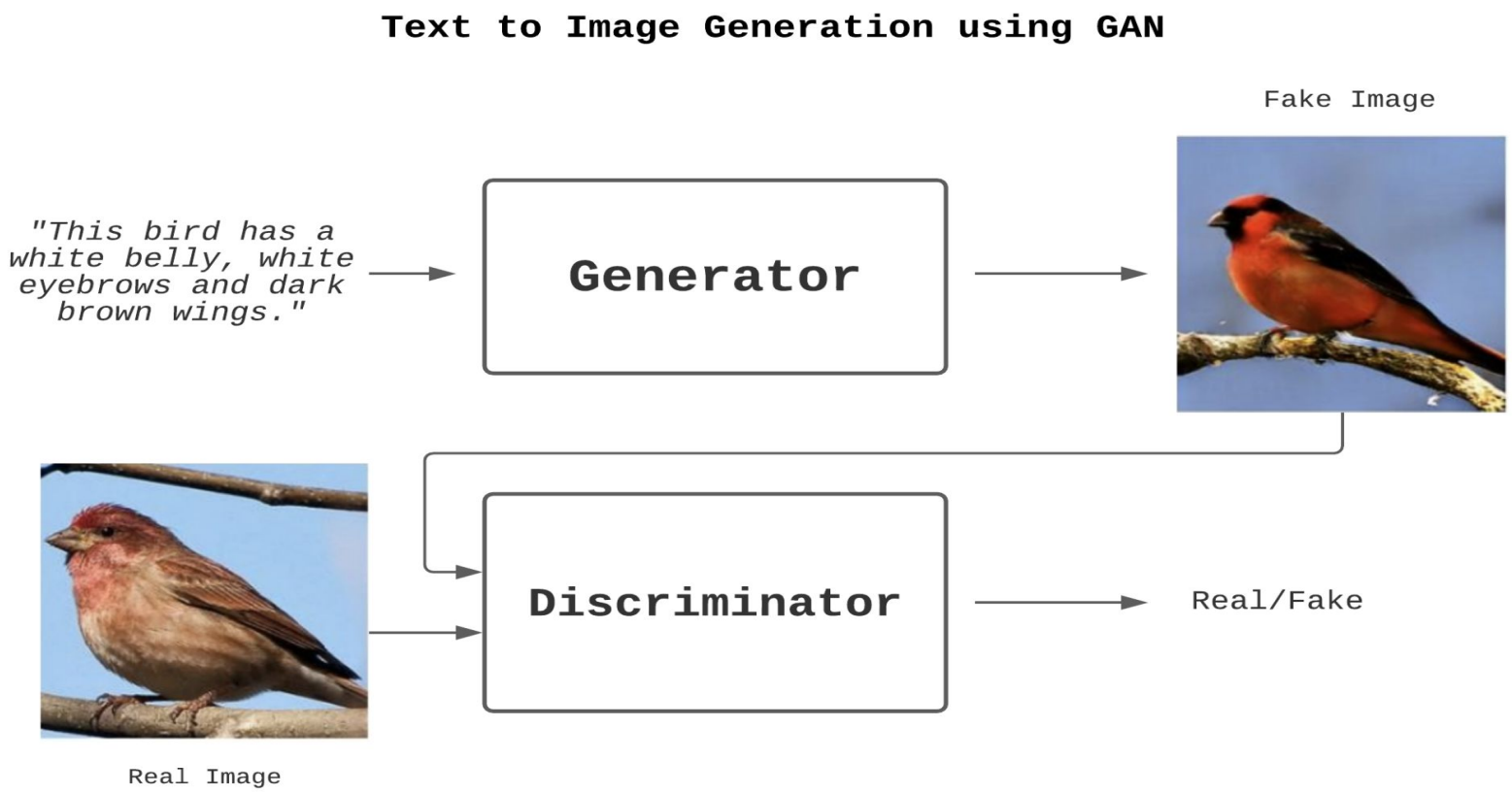
Project Guide :

Prof. P Rama Devi

Problem Statement

Generation of images from text captions using Generative Adversarial Networks (GAN's) .

Given a Text input the goal of the network is to generate a high-resolution image that matches text context.



Scope

1. Synthesizing high-quality images from text descriptions is a challenging problem in computer vision and has many practical applications.

2. Samples generated by existing text-to-image approaches can roughly reflect the meaning of the given descriptions, but they fail to contain necessary details and vivid object parts.

3. Can be used to generate images that doesn't exist in real life which can be useful in content creation.

Literature Review

[1]. **StackGAN** was used to generate images from text. It followed a 2 stage approach to generate high resolution images from text. First stage generated a low resolution image and used that to generate a high resolution image in the second stage.

[2]. **DF-GAN** proposed one stage GAN architecture to generate high resolution images from text and also proposed that 2-stage approach is hard to train and has its own challenges. DF-GAN uses an effective way to fuse the text embeddings and the image features together to produce a more realistic image.

[3]. **Residual Dense Blocks** were used in Image Super Resolution (SR) applications. One of the applications being converting a low resolution to high resolution image. RDB blocks effectively performs local residual learning and helps the model to retain low-level image details.

[4]. A **U-NET** based discriminator architecture was explored to modify the existing StackGAN discriminator model as proposed by the authors to improve the discriminator model. U-NET based discriminator allows discriminator to provide per pixel feedback and helps to distinguish between fake and real images.

Literature Review (continued)

[5] **Generative Adversarial Text to Image Synthesis** was the first approach that made use of GAN to generate images that matched a certain text description

[6] **Semantic Spatial Aware GAN** tries to achieve better images but effectively fusing textual embeddings in the batchnorm layers of the model. This GAN uses the same discriminator as proposed in DF-GAN due to its simplicity and effectiveness.

Proposed Work

1. Implement a Generative Adversarial Network **(IMAGAN)** that could generate high resolution images from a text input.

2. To try out other approaches like **VQ-VAE / VQ-GAN** to achieve high resolution images.

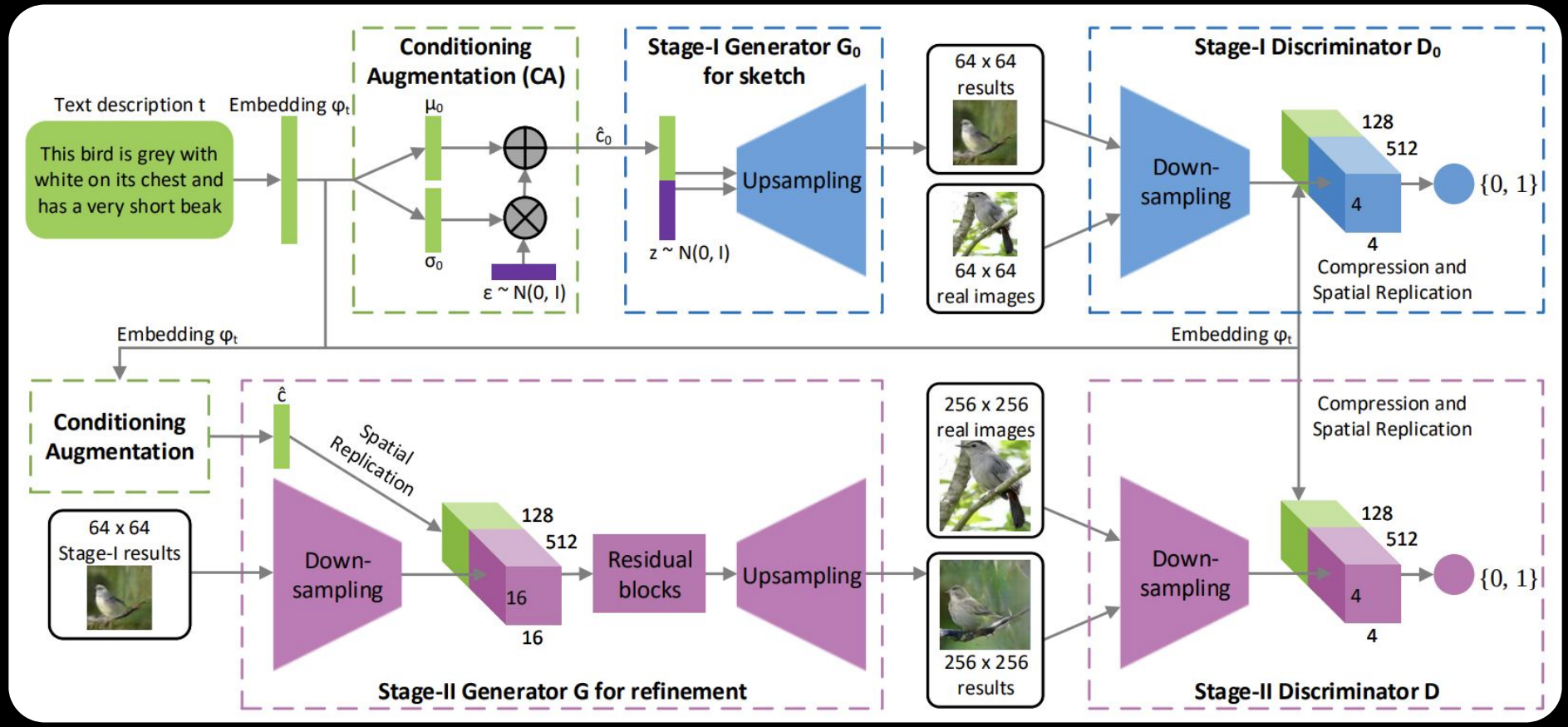
3. To perform hyperparameter tuning on the model by changing multiple hyperparameters to achieve better model metrics like increased accuracy, lower loss rate, higher Inception scores and and lower FID scores.

4. To beat the existing **SOTA** models by trying out different model architectures and publish a research paper on the same.

Work Carried Out

1. We implemented **StackGAN** according to the paper [1] and trained both the stages for 600 epochs on **CUB dataset**, but the outcome was not satisfying since 2 stage approach was harder to train and does not scale very well.
2. We tried to modify the **StackGAN** architecture by modifying the generator in both the stages with the intention of generating better quality images, however due to the limitation of StackGAN the result obtained was not as expected.
3. We also modified the Discriminator architecture of **StackGAN** by using a **U-NET** [4] based approach. However, we could not find any improvements in our model.

StackGAN Architecture



Images
Generated by
StackGAN Stage1

Stage 1 output

64x64

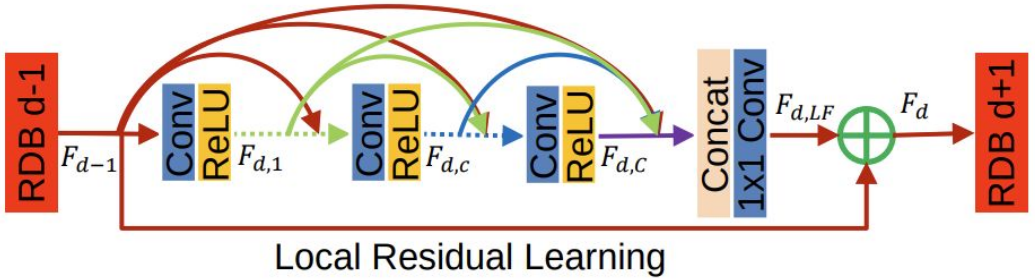
StackGAN Stage 2

256x256

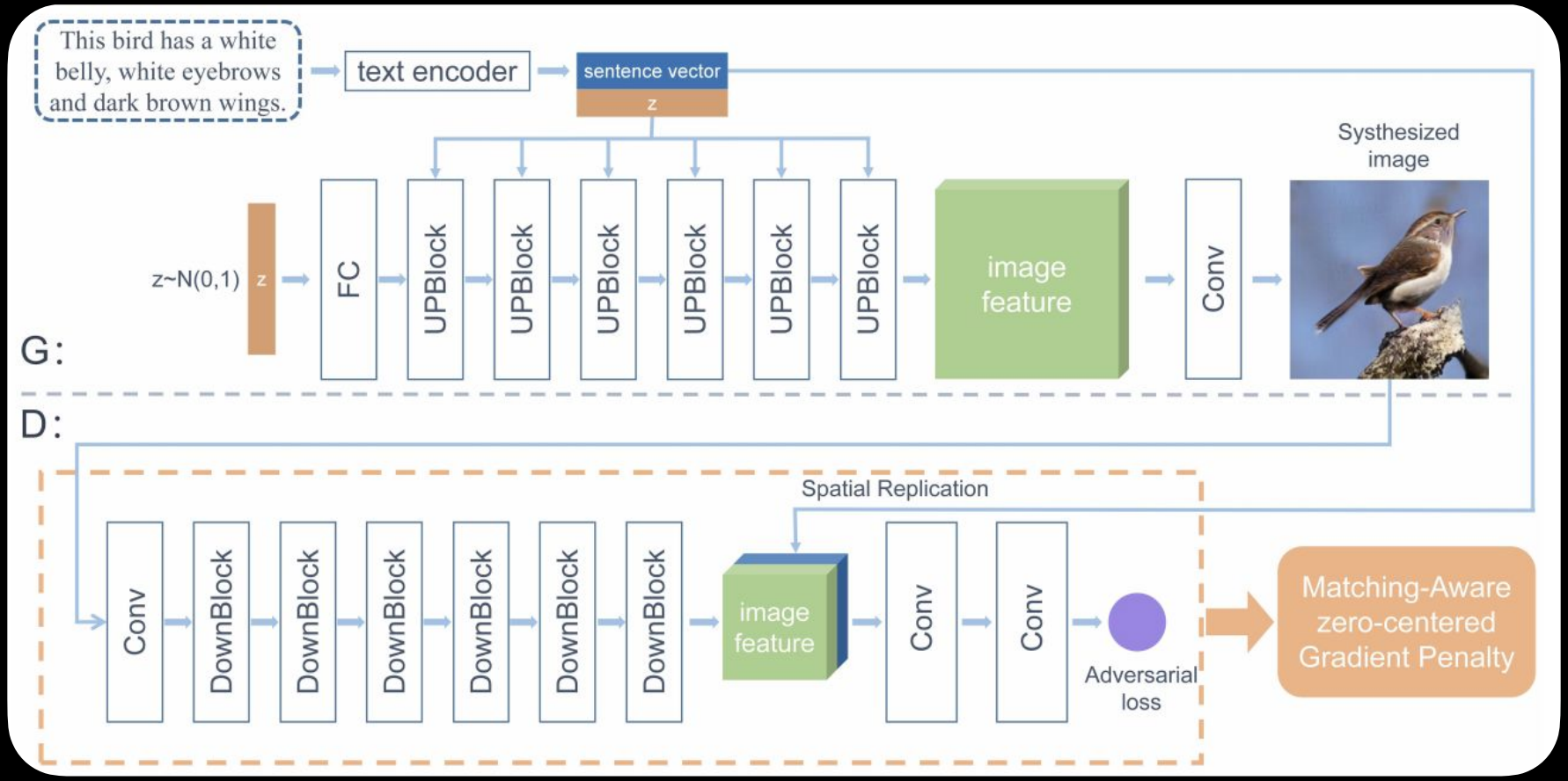
Output of StackGAN Stage 2 gets limited by the quality of image generated by stage 1

Work Carried Out (continued)

1. We implemented **Deep Fusion GAN (DF-GAN)** [2]. This GAN architecture fixed the limitation of StackGAN and was able to generate high resolution images.
2. We modified the Generator architecture of DF-GAN by using **Residual Dense Blocks (RDB)** [3].
3. Due to the limitation of hardware resources, we used **PyTorch Mixed Precision Training** to effectively utilize the GPU and reduce the training time by 300% on V100 GPU.



DF-GAN Architecture



Images
generated by
IMAGAN

Work that needs to be done

1. Training the new architecture was not completely successful as we obtained NaNs during the training process hence we would like to introspect the implementation. We believe it could be from the Mixed Precision Training.
2. We would also like to integrate the **VQ-VAE/VQ-GAN** framework in our current implementation to improve the quality of image generated.
3. We would like to test out the quality of images generated by changing multiple hyper-parameters like batch size, no. of epochs, learning rate, normalization and dropout values.
4. Implement User Interface (UI) for others to test out the model.

References

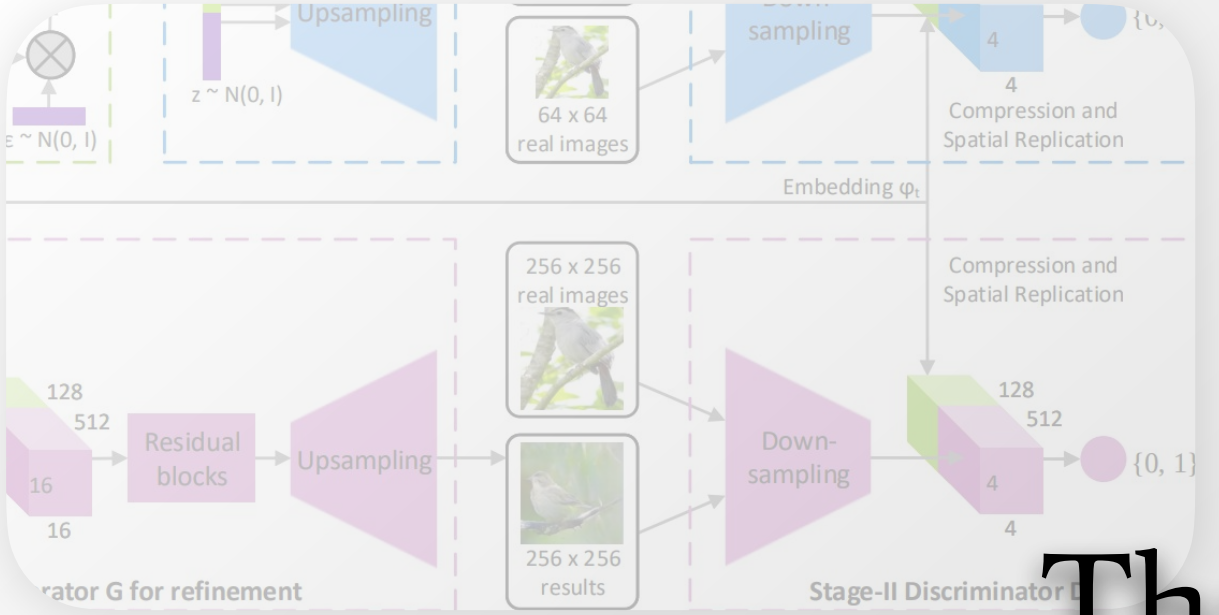
[1] Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 5907-5915. 2017.

[2] Tao, Ming, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. "Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis." *arXiv preprint arXiv:2008.05865* (2020).

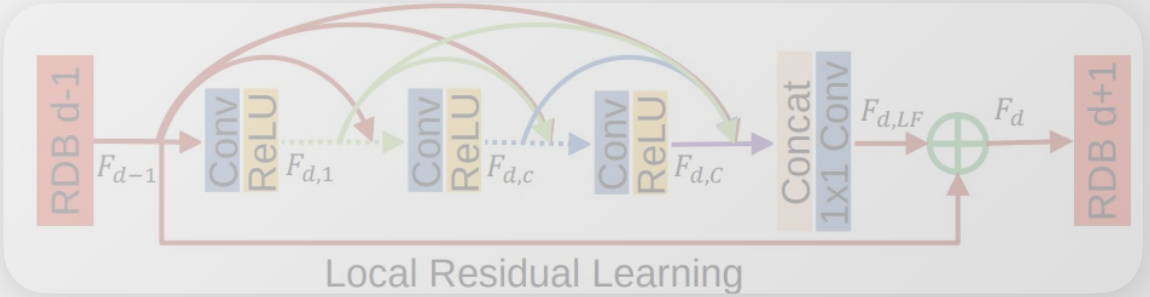
[3] Zhang, Yulun, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. "Residual dense network for image super-resolution." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472-2481. 2018.

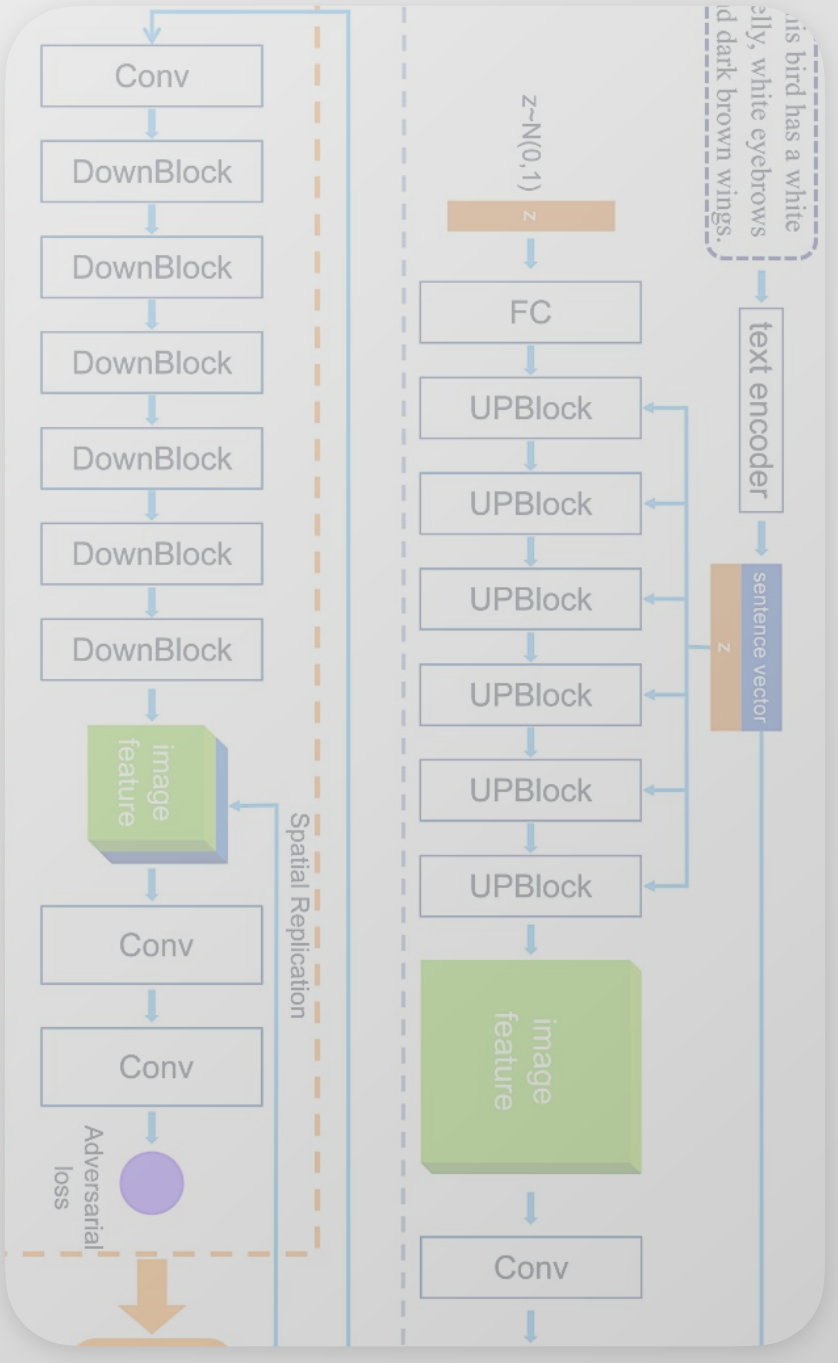
[4] Schonfeld, Edgar, Bernt Schiele, and Anna Khoreva. "A u-net based discriminator for generative adversarial networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8207-8216. 2020.

[5] Reed, Scott, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. "Generative adversarial text to image synthesis." In *International Conference on Machine Learning*, pp. 1060-1069. PMLR, 2016.


Upsampling
z ~ N(0, I)
64 x 64
real images
Down-
sampling
4
4
Compression and
Spatial Replication
Embedding φt
256 x 256
real images
Compression and
Spatial Replication
128
512
16
16
Residual
blocks
Upsampling
Down-
sampling
128
512
4
4
{0, 1}
256 x 256
results
rator G for refinement
Stage-II Discriminator D

Thank You


RDB d-1
F_{d-1}
Conv
ReLU
$F_{d,1}$
Conv
ReLU
$F_{d,c}$
Conv
ReLU
$F_{d,C}$
Concat
1x1 Conv
$F_{d,LF}$
F_d
RDB d+1
Local Residual Learning


Conv
DownBlock
DownBlock
DownBlock
DownBlock
DownBlock
DownBlock
image feature
Conv
Conv
Adversarial loss
Spatial Replication
z~N(0,1)
z
FC
UPBlock
UPBlock
UPBlock
UPBlock
UPBlock
UPBlock
image feature
Conv
text encoder
sentence vector
z