StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

1st Abhishek Anand
Engineering Physics Department
Indian Insititute of Technology Bombay

2nd Kapil Bhagat
Engineering Physics Department
Indian Insititute of Technology Bombay

3rd Shadab Anjum
Engineering Physics Department
Indian Insititute of Technology Bombay

Abstract—**StackGAN is a machine learning technique for generating photo-realistic images from textual descriptions. The approach is based on a novel architecture of stacked Generative Adversarial Networks (GANs) that enable the synthesis of high-quality images with impressive visual fidelity. The StackGAN technique involves two stages of generation. In the first stage, a low-resolution image is generated from the input text using a GAN. Then, in the second stage, the low-resolution image is further refined to generate a high-resolution image that matches the input text description.**

The StackGAN method is capable of generating complex and diverse images that capture the essence of the input text. This has many potential applications, including image and video synthesis, virtual and augmented reality, and even artistic expression. In addition, StackGAN can be used for a variety of tasks, such as synthesizing images for training datasets or generating realistic images for computer graphics and design.

StackGAN has demonstrated state-of-the-art performance in terms of image quality, diversity, and realism, outperforming existing methods. The technique has been tested on a range of datasets, including birds, flowers, and natural scenes, and has consistently produced visually stunning results. StackGAN represents an exciting advancement in the field of image synthesis and has the potential to revolutionize the way we generate and manipulate visual media.

I. Introduction

Image synthesis is a challenging task in machine learning that involves generating new images based on input data or descriptions. With the rise of deep learning techniques, there has been significant progress in this field, particularly in the area of Generative Adversarial Networks (GANs). GANs are a type of neural network that can learn to generate realistic images by training two models: a generator and a discriminator.

Recently, a novel approach called StackGAN has been proposed for text-to-image synthesis, which uses a stacked GAN architecture to generate photo-realistic images from textual descriptions. This technique involves two stages of generation, with the first stage producing a low-resolution image from the text input, and the second stage using the low-resolution image to generate a high-resolution image that matches the input text.

The StackGAN technique has been shown to generate high-quality and diverse images that capture the essence of the input text, outperforming existing methods. It has many potential applications in areas such as computer vision, graphics, and art, and could revolutionize the way we generate and manipulate visual media.

In this project, we will explore the StackGAN technique for text-to-image synthesis and evaluate its performance on various datasets. We will also investigate the potential applications of this technique in areas such as virtual and augmented reality, and computer graphics. By the end of this project, we aim to gain a deeper understanding of the StackGAN approach and its implications for the field of image synthesis.

II. Background and Previous Work

The task of image synthesis has been studied extensively in the field of computer vision and machine learning. Early approaches for image synthesis focused on rule-based methods or statistical models, which had limited success in generating realistic images. However, with the advent of deep learning techniques, particularly Generative Adversarial Networks (GANs), there has been significant progress in this field.

GANs are a type of neural network that consists of two models: a generator and a discriminator. The generator learns to generate new images, while the discriminator learns to distinguish between real and generated images. The models are trained iteratively, with the generator attempting to fool the discriminator and the discriminator trying to correctly identify real images. This process results in a generator that can learn to produce high-quality, realistic images.

In recent years, GANs have been applied to the task of text-to-image synthesis, where the goal is to generate images from textual descriptions. However, previous methods have struggled to produce photo-realistic images that accurately capture the essence of the input text.

The StackGAN approach was proposed to address these limitations. The technique involves a two-stage process, where a low-resolution image is first generated from the text input, and then a high-resolution image is generated by conditioning on the low-resolution image and the text. The approach was shown to produce high-quality, diverse, and photo-realistic images that outperformed previous methods.

In addition to StackGAN, other methods have been proposed for text-to-image synthesis, including GAN-INT-CLS, AttnGAN, MirrorGAN, and CLIP-guided synthesis. These methods have introduced various improvements to the basic GAN architecture, such as attention mechanisms, adversarial losses, and fine-tuning with pre-trained models.

Overall, the field of text-to-image synthesis has seen significant progress in recent years, with StackGAN representing a promising approach for generating photo-realistic images from textual descriptions.

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Datasets and evaluation metrics

The Caltech-UCSD Birds-200-2011 dataset is a widely-used dataset in the field of computer vision and machine learning for studying fine-grained image classification and text-to-image synthesis. The dataset contains over 11,000 images of 200 bird species, with each species having approximately 50 images. The images were captured under varying conditions, such as different backgrounds, poses, and lighting, making the dataset challenging and realistic.

The dataset includes textual descriptions for each bird species, providing a rich source of information for text-to-image synthesis tasks. The textual descriptions include information about the bird's physical characteristics, habitat, and behavior.

IV. Implementation details

To generate high-resolution images with photo-realistic details, we propose a simple yet effective Stacked Generative Adversarial Networks. It decomposes the text-to-image generative process into two stages:

- **Stage I:** It sketches the primitive shape and basic colors of the object conditioned on the given text description, and draws the background layout from a random noise vector, yielding a low-resolution 64X64 image.
- **Stage II:** It corrects defects in the low-resolution image from Stage-I and completes details of the object by reading the text description again, producing a high resolution upsampled 256X256 photo-realistic image.

Generative Adversarial Networks (GAN) composed of two models that are alternatively trained to compete with each other:

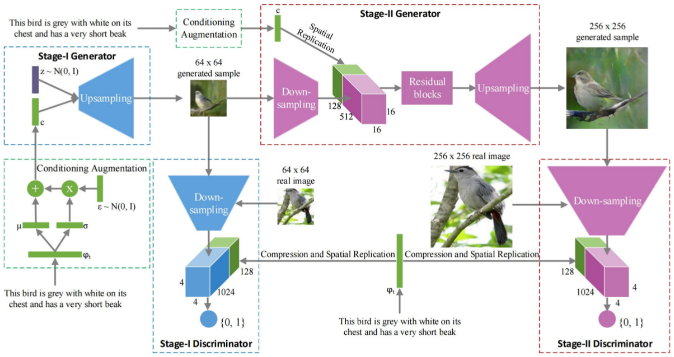
- **The Generator G** optimized to generate images that are difficult for the discriminator D to differentiate from real images.
- **The Discriminator D** optimized to distinguish real images from the synthetic images generated by G.

Loss Function: Scores from the Discriminator:

$$s_r \leftarrow D(x, h)\{realimage, righttext\}$$

$$s_w \leftarrow D(x, \hat{h})\{realimage, wrongtext\}$$

$$s_f \leftarrow D(\hat{x}, h)\{fakeimage, righttext\}$$


This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
c
Stage-II Generator
Spatial Replication
256 x 256 generated sample
Stage-I Generator
z ~ N(0, I)
Upsampling
c
64 x 64 generated sample
Down-sampling
128
512
16
16
Residual blocks
Upsampling
Conditioning Augmentation
ε ~ N(0, I)
μ
σ
φt
This bird is grey with white on its chest and has a very short beak
Down-sampling
64 x 64 real image
256 x 256 real image
Down-sampling
Compression and Spatial Replication
Compression and Spatial Replication
128
4
1024
4
{0, 1}
Stage-I Discriminator
φt
This bird is grey with white on its chest and has a very short beak
128
1024
4
4
{0, 1}
Stage-II Discriminator

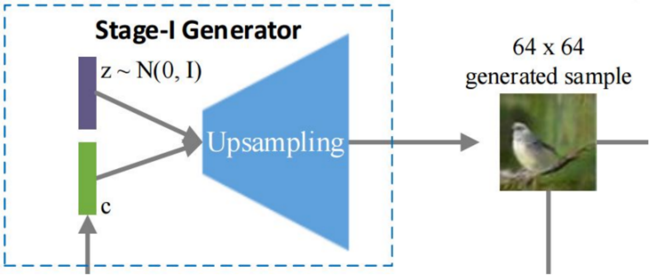
Then alternate: Maximizing

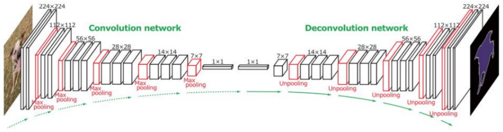
$$L_D \longrightarrow \log\left(S_r\right) + \left(\log\left(1 - S_\omega\right) + \log\left(1 - S_f\right)\right)/2$$

and Minimizing

$$L_G \longleftarrow \log\left(1 - S_f\right) + \lambda D_{KL}\left(N\left(\mu_0\left(\phi_t\right), \Sigma_0\left(\phi_t\right)\right) || N(0, I)\right)$$

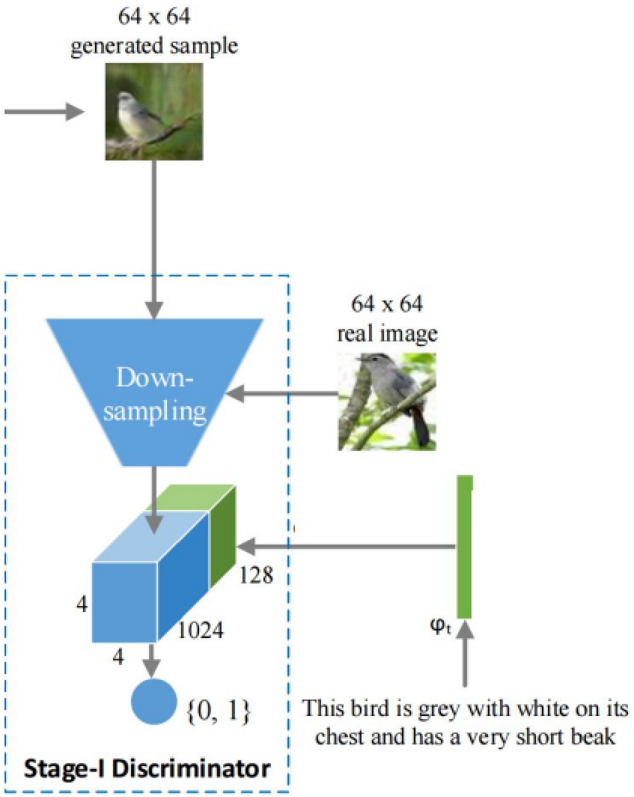
Stage I generator: Instead of directly generating a high-resolution image conditioned on the text description, we simplify the task to first generate a low-resolution image with our Stage-I GAN, which focuses on drawing only rough shape and correct colors for the object.





Stage I Discriminator: For the discriminator D_0 the text embedding ϕ_t is first compressed to Nd dimensions using a fully-connected layer and then spatially replicated to form a $M_d \times M_d \times N_d$ tensor. Meanwhile, the image is fed through a series of down-sampling blocks until it has $M_d \times M_d$ spatial dimension. Then, the image filter map is concatenated along the channel dimension with the text tensor. The resulting tensor is further fed to a 1X1 convolutional layer to jointly learn features across the image and the text. Finally, a fully connected layer with one node is used to produce the decision score.

Down-Sampling:

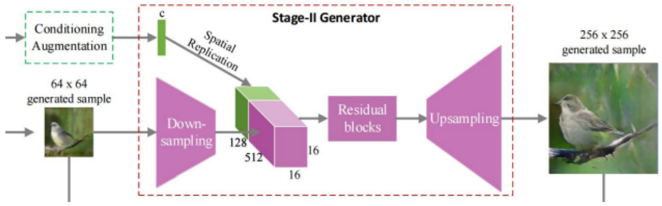


- **Images**
  - Stride-2 convolutions, Batch Norm., Leaky ReLU
  - $64 \times 64 \times 3 \rightarrow 4 \times 4 \times 1024$
- **Text**
  - Fully-connected layer: $\phi \rightarrow 128$
  - Spatially replicate to $4 \times 4 \times 128$
- **Depth Concatenate**
  - Total of $4 \times 4 \times 1152$

Score:

- **1x1 convolution, followed by 4x4 convolution**
  - Produces scalar value between 0 and 1

Stage II Generator: When generating low-resolution images with Stage-I GAN some details might be missing, and the shape of the object may not be accurate. This is because Stage-I GAN only focuses on creating basic shapes and colors. To overcome this, we use a Stage-II GAN, which takes the low-resolution images and the text embedding as inputs.

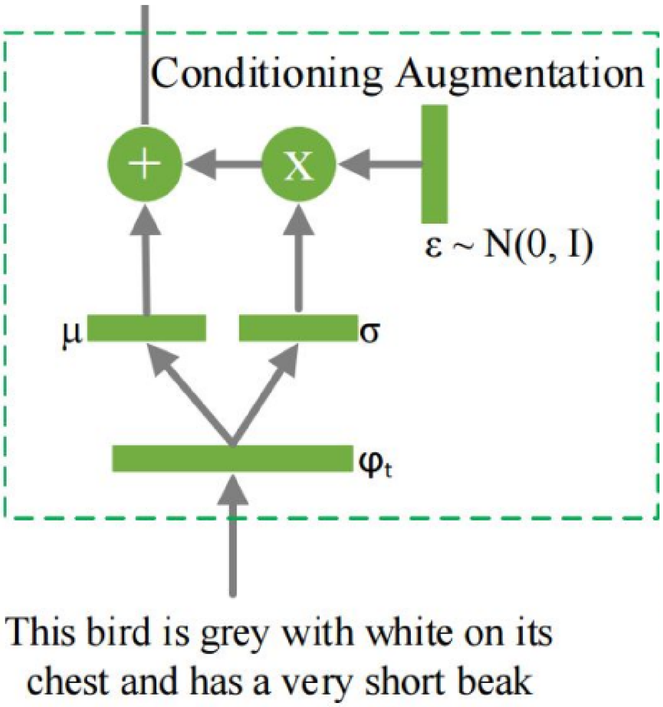


The Stage-II GAN refines the Stage-I results to create high-resolution images with more details and accuracy. It corrects any defects in the Stage-I results and completes any ignored text information to generate more realistic images. This means that the Stage-II GAN builds upon the Stage-I results to create more accurate and detailed images.

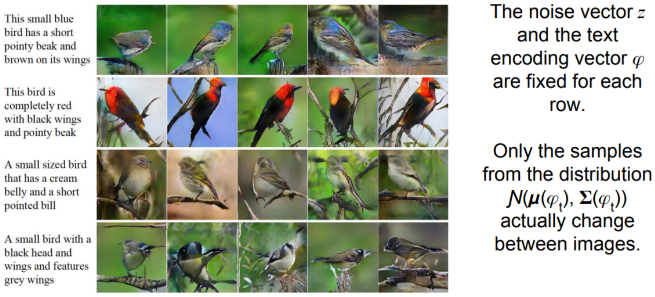
- **Takes:**
  - Stage-I's image
  - 'Conditioned augmentation' representing input text
- **Downsampling via CNN, Batch Norm, Leaky Relu**
- **Residual Blocks, similar to ResNet**
  - To jointly encode image and text features

Conditioning Augmentation

- Text Encoding
- Uses a "hybrid character-level convolutional recurrent neural network"
- Same as Reed et al. "GAN Text to Image Synthesis" paper
- **Augmentation**
- Randomly sample "latent variables" from the independent Gaussian distribution

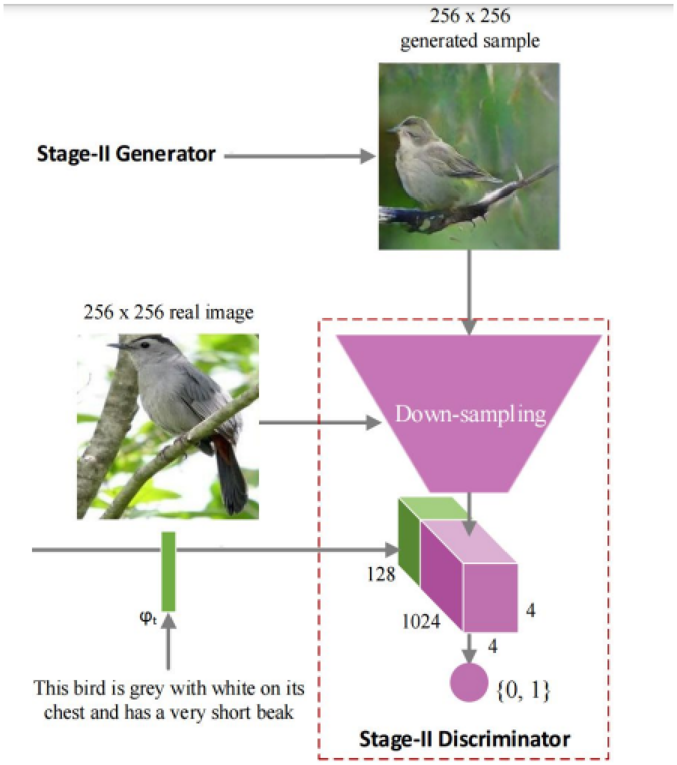

Conditioning Augmentation
+
X
ε ~ N(0, I)
μ
σ
φt
This bird is grey with white on its chest and has a very short beak

Variations due purely to Conditioning Augmentation

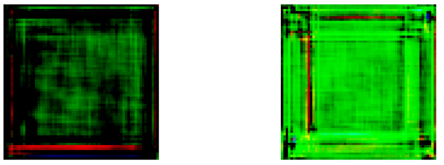

This small blue bird has a short pointy beak and brown on its wings
This bird is completely red with black wings and pointy beak
A small sized bird that has a cream belly and a short pointed bill
A small bird with a black head and wings and features grey wings
The noise vector z and the text encoding vector φ are fixed for each row.
Only the samples from the distribution $\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t))$ actually change between images.

Stage II Discriminator:

- **Down-sampling**
  - Same as Stage-I, but more layers
- **Loss functions**
  - Same as before, but now G is "encourage[d] to extract previously ignored information" in order to trick a more perceptive and detail-oriented D.



Our Results and Evaluation: Actually we have generated some Images in Stage I, but only for 6 Epochs, because of our device memory limit exceeded. Unfortunately, We can't generate images from Stage II. But If we run our code on some high computational device for nearly 600 Epochs, then we can get a very high Resolution Image. Anyway, we are attaching our generated images, after 6 Epochs, it is not worthy but it validate our code. But I am also attaching the fully generated images of birds by both Stage I and Stage II.

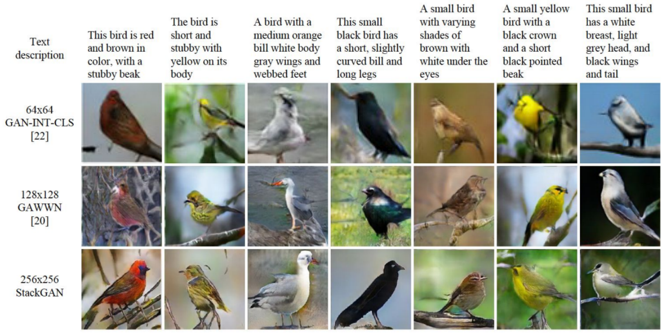


(a) After Epoch 1 (b) After Epoch 6

Method	Inception scores		Human rank	
	CUB	Oxford-102	CUB	Oxford-102
GAN-INT-CLS [22]	2.88 ± .04	2.66 ± .03	2.81 ±.03	1.87 ±.03
GAWWN [20]	3.62 ± .07	/	1.99 ±.04	/
Original StackGAN	3.70 ± .04	3.20 ± .01	1.37 ±.02	1.13 ±.03

- State of the art Inception score, 28.47 % and 20.30 % improvement
- People seem to like the results, too

Fully Generated Image:
Here is the final generated images of the birds after fully run of the code:


Text description
This bird is red and brown in color, with a stubby beak
The bird is short and stubby with yellow on its body
A bird with a medium orange bill white body gray wings and webbed feet
This small black bird has a short, slightly curved bill and long legs
A small bird with varying shades of brown with white under the eyes
A small yellow bird with a black crown and a short black pointed beak
This small bird has a white breast, light grey head, and black wings and tail
64x64 GAN-INT-CLS [22]
128x128 GAWWN [20]
256x256 StackGAN

V. Conclusion

The approach changes the text-to-image synthesis into a two-stage process that involves sketch-refinement. In the first stage, Stage-I GAN creates a low-resolution image based on the basic color and shape constraints from the text description. The second stage, Stage-II GAN, corrects the defects in the Stage-I results and adds more details to produce higher resolution images with better image quality. The results of extensive quantitative and qualitative evaluations demonstrate the effectiveness of the proposed method. Compared to existing text-to-image generative models, our approach generates higher resolution images with more photo-realistic details and diversity.

References

https://web.archive.org/web/20220101171059/
https://arxiv.org/pdf/1605.05396.pdf
https://github.com/hanzhanggit/StackGAN
https://www.youtube.com/watch?v=AALBGpLbj6Q
https://www.youtube.com/watch?v=Gib$_k iXgnvA$
http://www.vision.caltech.edu/visipedia/CUB-200.html