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

(ICCV 2017)

Varun Syal and Sparsh Gupta
Problem & Results

Given textual descriptions, synthesize photo-realistic images.
- High resolution (256x256)
- Detailed with vivid object parts!

28.47% improvement on CUB, 20.3% improvement on Oxford-102 in inception scores

Higher Resolution images
More results

The flower has petals that are yellow with shades of orange (Oxford-102)

A living room with hard wooden floor filled with furniture (MS COCO)
Motivation

Lots of practical applications:

- Photo Editing, Computer Aided Design
- Summarization tools for text
- Assistive communication tools
- Language learning tools - Educational

Current Text-to-image approaches fail to capture details, only rough reflections!

No previous attempt to synthesize high resolution (256x256) images from text.
Related Work
Autoregressive Models

- Given the training set - generate new samples by learning explicit distribution
- PixelRNN, PixelCNN
- Used for Image completion

Issues:
- Training is very slow
- Not suited for high resolution images
- Not text-to-image
Generative Adversarial Networks (GANs)

- Better performance and sharper images
- Learn implicit distribution from the given data

Issues:
- Training is unstable, hard to train for high resolution images.

Conditional GANs (Reed et al.)
- 64x64 images, lack details & vividness

Results from Reed et. al on CUB
Multiple GANs in Series

**S2-GAN (Wang et al.)**
Structure GAN + Style GAN
(Not from text)

**Stacked GAN (Huang et al.)**
Low Resolution (32x32) images
StackGANs: Overview

**Text Encoder**: Converts text description into a text embedding.

**Conditioning Augmentation (CA)**: Novel technique for smoothness in conditioning manifold

**Stage-I GAN**: Sketches primitive shape and colors of the objects.

**Stage-II GAN**: Corrects defects in Stage-I, adds compelling details → Sketch Refinement process
Background - What is a GAN?

**Generator G**: optimized to reproduce true data distribution $p_{data}$.

**Discriminator D** optimized to distinguish between real and synthetic images generated by G.
Background - What is a GAN?

Similar to a two Player Min-Max Game with Objective Function:

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]
\]

- x: Real Image
- z: noise vector
Background: Conditional GAN

Generator learns to generate a fake sample with a specific condition or characteristics (say Y) rather than from unknown noise distribution.

This condition is fed into both Generator & Discriminator.
Background - Embeddings

Embeddings are learned representation of Text as a vector of numbers.

- Dimensionality Reduction - It is a more efficient representation
- Contextual Similarity - It is a more expressive representation
Stacked Generative Adversarial Network (StackGAN)
StackGAN - Conditioning Augmentation

This bird is grey with white on its chest and has a very short beak.
StackGAN - Conditioning Augmentation

This bird is grey with white on its chest and has a very short beak

Conditioning Augmentation (CA)

\[ \hat{c}_0 \]

\[ \varepsilon \sim N(0, I) \]
StackGAN - Conditioning Augmentation

-- Latent Variable $\hat{c}$ sampled from an independent Gaussian distribution:

$$\hat{c} \in N(\mu(\varphi_t), \Sigma(\varphi_t))$$

-- This is implemented as:

$$\hat{c}_0 = \mu_0 + \sigma_0 \circ \epsilon$$

$$\epsilon \sim N(0, I)$$

-- Regularization to ensure smoothness over conditioning manifold:

$$D_{KL}(N(\mu(\varphi_t), \Sigma(\varphi_t)) || N(0, I))$$
StackGAN - Stage I

This bird is grey with white on its chest and has a very short beak.
StackGAN - Stage I

Primitive Shape and Colors

Stage-I Generator $G_0$

for sketch

Stage-I Discriminator $D_0$

64 x 64 results

64 x 64 real images

Upsampling

Downsampling

Embedding $\varphi_t$

{0, 1}

Compression and Spatial Replication
StackGAN - Stage I

Sketches the primitive shape and basic colors of the object

**Discriminator:**

\[
L_{D_0} = \mathbb{E}_{(I_0, t) \sim p_{data}}[\log D_0(I_0, \varphi_t)] + \mathbb{E}_{z \sim p_z, t \sim p_{data}}[\log(1 - D_0(G_0(z, \hat{c}_0), \varphi_t))] 
\]

**Generator:**

\[
L_{G_0} = \mathbb{E}_{z \sim p_z, t \sim p_{data}}[\log(1 - D_0(G_0(z, \hat{c}_0), \varphi_t))] + \lambda D_{KL} \left( N(\mu(\varphi_t), \Sigma(\varphi_t)) \mid \mid N(0, I) \right) 
\]
StackGAN - Stage II

This bird is grey with white on its chest and has a very short beak.
StackGAN - Stage II

High Resolution Image with Photo-realistic Details
StackGAN - Stage II

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 photo-realistic image.

Discriminator:
\[
L_D = \mathbb{E}_{(I,t) \sim p_{data}}[\log D(I, \varphi_t)] + \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}}[\log(1 - D(G(s_0, \hat{c}), \varphi_t))]
\]

Generator:
\[
L_G = \mathbb{E}_{s_0 \sim p_{G_0}, t \sim p_{data}}[\log(1 - D(G(s_0, \hat{c}), \varphi_t)) + \lambda D_{KL}(N(\mu(\varphi_t), \Sigma(\varphi_t)) || N(0, I))]
\]
StackGAN - Implementation Details

-- Iteratively train $D_0$ and $G_0$ of Stage I for 600 epochs by fixing Stage II parameters.

-- Then, iteratively train $D$ and $G$ of Stage II by fixing Stage I parameters for another 600 epochs.

-- Adam Optimizer with initial learning rate 0.0002 and batch size 64
Experiments - Dataset

- Caltech-UCSD-Birds (CUB)
  -- 200 bird species with 11788 images

(a) Californian gull

(b) Glaucous gull
Experiments - Dataset

- MS COCO
  - 80k images in training set, 40k images in validation set
Experiments - Dataset

- Oxford-102
  -- 102 flower categories, between 40 to 258 images per class
Experiments - Results on StackGAN

The small bird has a red head with feathers that fade from red to gray from head to tail.

Stage-I images

Stage-II images
Experiments - Results on StackGAN

This bird is black with green and has a very short beak

Stage-I images

Stage-II images
This flower is yellow in color, with petals that are vertically layered.
Experiments - Results on StackGAN

This flower has white petals with a yellow tip and a yellow pistil
Experiments - Results on StackGAN

A living room with hard wood floors filled with furniture

Stage-I images

Stage-II images
Experiments - Results on StackGAN

There are many pieces of broccoli and vegetables here

Stage-I images

Stage-II images
Experiments - What does StackGAN learn?

-- Visual features were extracted from Stage II Discriminator for each generated image and all training images.
-- For each generated image, its nearest neighbors in training set were found.
Experiments - Comparative Study

Comparison with Previous State-of-the Art Work:
-- GAN-INT-CLS (Reed et al, ICML 2016)
  Only Stage I; Training for mismatched (text, image pairs) and interpolation
-- GAWWN (Reed et al, NIPS 2016)
  Use of spatial constraints to generate different parts of images

Comparison with Baselines:
-- Stage-I GAN
-- StackGAN for different sizes
-- Effect of Conditional Augmentation
-- Effect of Text Input at both stages
Experiments - Evaluation Metrics

Inception Score
-- Randomly selected 30k samples
-- $I = \exp\left(E_x D_{KL}(p(y|x) \parallel p(y))\right)$
-- Good models should generate diverse but meaningful images

Human Ranking
-- 10 users
-- 50 text descriptions for each class of CUB and Oxford-102
-- 4k text descriptions from MS COCO Validation dataset
## Experiments - Comparison with Previous Work

<table>
<thead>
<tr>
<th>Metric</th>
<th>Dataset</th>
<th>GAN-INT-CLS Mean ± SD</th>
<th>GAWWN Mean ± SD</th>
<th>Our StackGAN Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception score</td>
<td>CUB</td>
<td>2.88 ± 0.04</td>
<td>3.62 ± 0.07</td>
<td>3.70 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>Oxford</td>
<td>2.66 ± 0.03</td>
<td>/</td>
<td>3.20 ± 0.01</td>
</tr>
<tr>
<td></td>
<td>COCO</td>
<td>7.88 ± 0.07</td>
<td>/</td>
<td>8.45 ± 0.03</td>
</tr>
<tr>
<td>Human rank</td>
<td>CUB</td>
<td>2.81 ± 0.03</td>
<td>1.99 ± 0.04</td>
<td>1.37 ± 0.02</td>
</tr>
<tr>
<td></td>
<td>Oxford</td>
<td>1.87 ± 0.03</td>
<td>/</td>
<td>1.13 ± 0.03</td>
</tr>
<tr>
<td></td>
<td>COCO</td>
<td>1.89 ± 0.04</td>
<td>/</td>
<td>1.11 ± 0.03</td>
</tr>
</tbody>
</table>

Higher is better

Lower is better
Experiments - Comparison with Previous Work

<table>
<thead>
<tr>
<th>Text description</th>
<th>64x64 GAN-INT-CLS</th>
<th>128x128 GAWWN</th>
<th>256x256 StackGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>This bird is red and brown in color, with a stubby beak</td>
<td>This bird is short and stubby with yellow on its body</td>
<td>A bird with a medium orange bill white body gray wings and webbed feet</td>
<td>A small bird with varying shades of brown with white under the eyes</td>
</tr>
<tr>
<td>A bird with a short, slightly curved bill and long legs</td>
<td>A small yellow bird with a black crown and a short black pointed beak</td>
<td>This small bird has a white breast, light grey head, and black wings and tail</td>
<td></td>
</tr>
</tbody>
</table>
## Experiments - Comparison with Previous Work

<table>
<thead>
<tr>
<th>Text Description</th>
<th>64x64 GAN-INT-CLS</th>
<th>256x256 StackGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>This flower has a lot of small purple petals in a dome-like configuration</td>
<td>![Image of flower 1]</td>
<td>![Image of flower 2]</td>
</tr>
<tr>
<td>This flower is pink, white, and yellow in color, and has petals that are striped</td>
<td>![Image of flower 3]</td>
<td>![Image of flower 4]</td>
</tr>
<tr>
<td>This flower has petals that are dark pink with white edges and pink stamen</td>
<td>![Image of flower 5]</td>
<td>![Image of flower 6]</td>
</tr>
<tr>
<td>This flower is white and yellow in color, with petals that are wavy and smooth</td>
<td>![Image of flower 7]</td>
<td>![Image of flower 8]</td>
</tr>
<tr>
<td>A picture of a very clean living room</td>
<td>![Image of living room 1]</td>
<td>![Image of living room 2]</td>
</tr>
<tr>
<td>A group of people on skis stand in the snow</td>
<td>![Image of skiers 1]</td>
<td>![Image of skiers 2]</td>
</tr>
<tr>
<td>Eggs fruit candy nuts and meat served on white dish</td>
<td>![Image of food 1]</td>
<td>![Image of food 2]</td>
</tr>
<tr>
<td>A street sign on a stoplight pole in the middle of a day</td>
<td>![Image of street sign 1]</td>
<td>![Image of street sign 2]</td>
</tr>
</tbody>
</table>
## Experiments - Comparison with Baselines

<table>
<thead>
<tr>
<th>Method</th>
<th>CA</th>
<th>Text twice</th>
<th>Inception score</th>
</tr>
</thead>
<tbody>
<tr>
<td>64×64 Stage-I GAN</td>
<td>no</td>
<td>/</td>
<td>2.66 ± .03</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>/</td>
<td>2.95 ± .02</td>
</tr>
<tr>
<td>256×256 Stage-I GAN</td>
<td>no</td>
<td>/</td>
<td>2.48 ± .00</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>/</td>
<td>3.02 ± .01</td>
</tr>
<tr>
<td>128×128 StackGAN</td>
<td>yes</td>
<td>no</td>
<td>3.13 ± .03</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>3.20 ± .03</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>yes</td>
<td>3.35 ± .02</td>
</tr>
<tr>
<td>256×256 StackGAN</td>
<td>yes</td>
<td>no</td>
<td>3.45 ± .02</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>3.31 ± .03</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>yes</td>
<td>3.70 ± .04</td>
</tr>
</tbody>
</table>

Higher is better
Experiments - Comparison with Baselines

A small bird with a black head and wings and features grey wings

This bird is completely red with black wings and pointy beak

256x256 Stage-I GAN without CA

256x256 Stage-I GAN with CA

256x256 StackGAN with CA, Text twice
Experiments - Sentence Embedding Interpolation

Gradual appearance changes from the first sentence’s meaning to that of the second can be observed.

The bird is completely red → The bird is completely yellow

This bird is completely red with black wings and pointy beak → this small blue bird has a short pointy beak and brown on its wings
Summary

- Two stacked GANs with Conditioning Augmentation to generate high resolution images with photo-realistic details.
- Decomposes the text-to-image generation synthesis to a novel sketch-refinement process.
- First model to generate 256 x 256 images with photo-realistic details.
- Does not simply memorize the training samples, but captures complex underlying language-image relations.
- Achieves significant improvements on benchmark datasets compared to state-of-the-art models.
- Improvement – StackGAN++