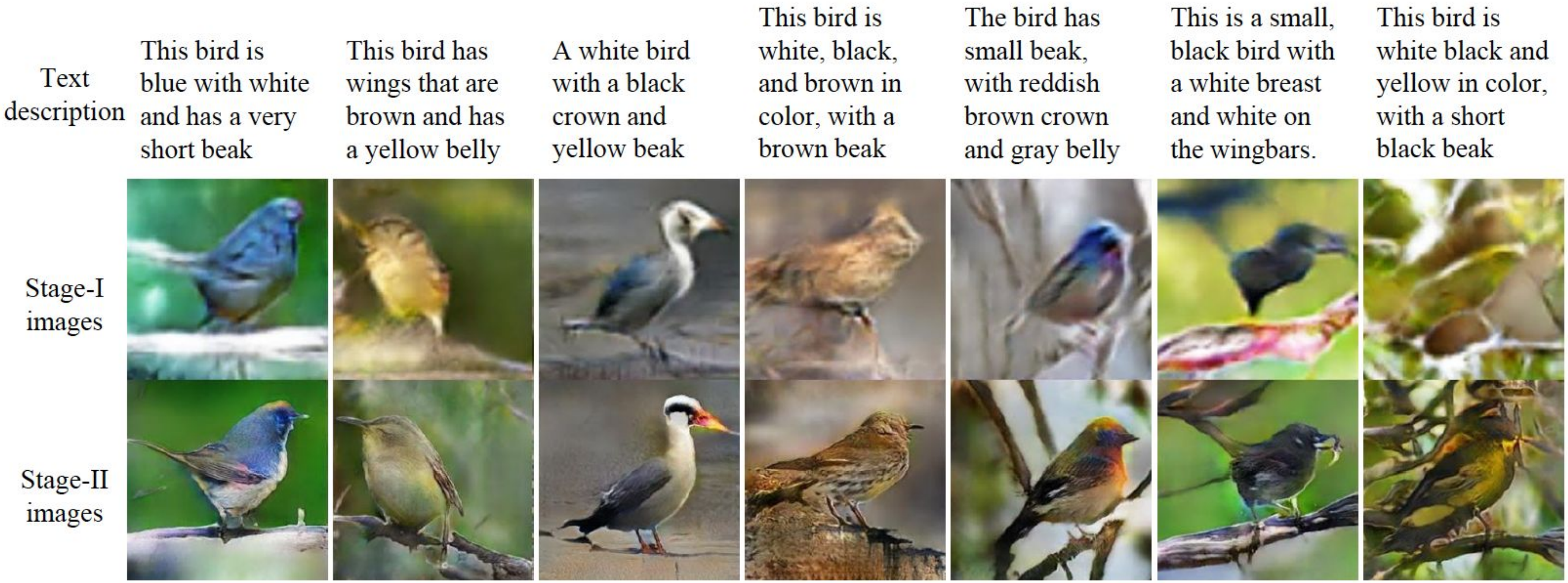
StackGAN

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

The Problem:



This bird is blue with white and has a very short beak

This bird has wings that are brown and has a yellow belly

A white bird with a black crown and yellow beak

This bird is white, black, and brown in color, with a brown beak

The bird has small beak, with reddish brown crown and gray belly

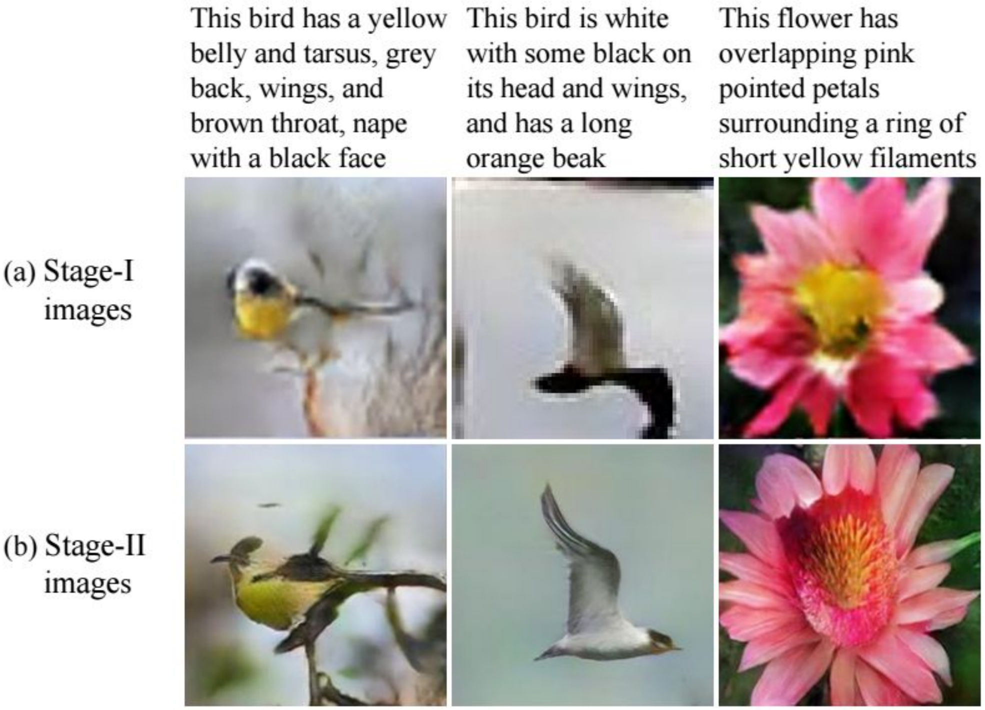
This is a small, black bird with a white breast and white on the wingbars.

This bird is white black and yellow in color, with a short black beak

2-Stage Network

- Stage 1.
  - Generates 64x64 images
  - Structural information
  - Low detail
- Stage 2.
  - Requires Stage 1. output
  - Upsamples to 256x256
  - Higher detail, photorealistic

Both stages take in the same conditioned textual input



This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face

This bird is white with some black on its head and wings, and has a long orange beak

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments

Generalized 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 Functions

Scores from The Discriminator:

$$s_r \leftarrow D(x, h) \ \{\text{real image, right text}\}$$
$$s_w \leftarrow D(x, \hat{h}) \ \{\text{real image, wrong text}\}$$
$$s_f \leftarrow D(\hat{x}, h) \ \{\text{fake image, right text}\}$$

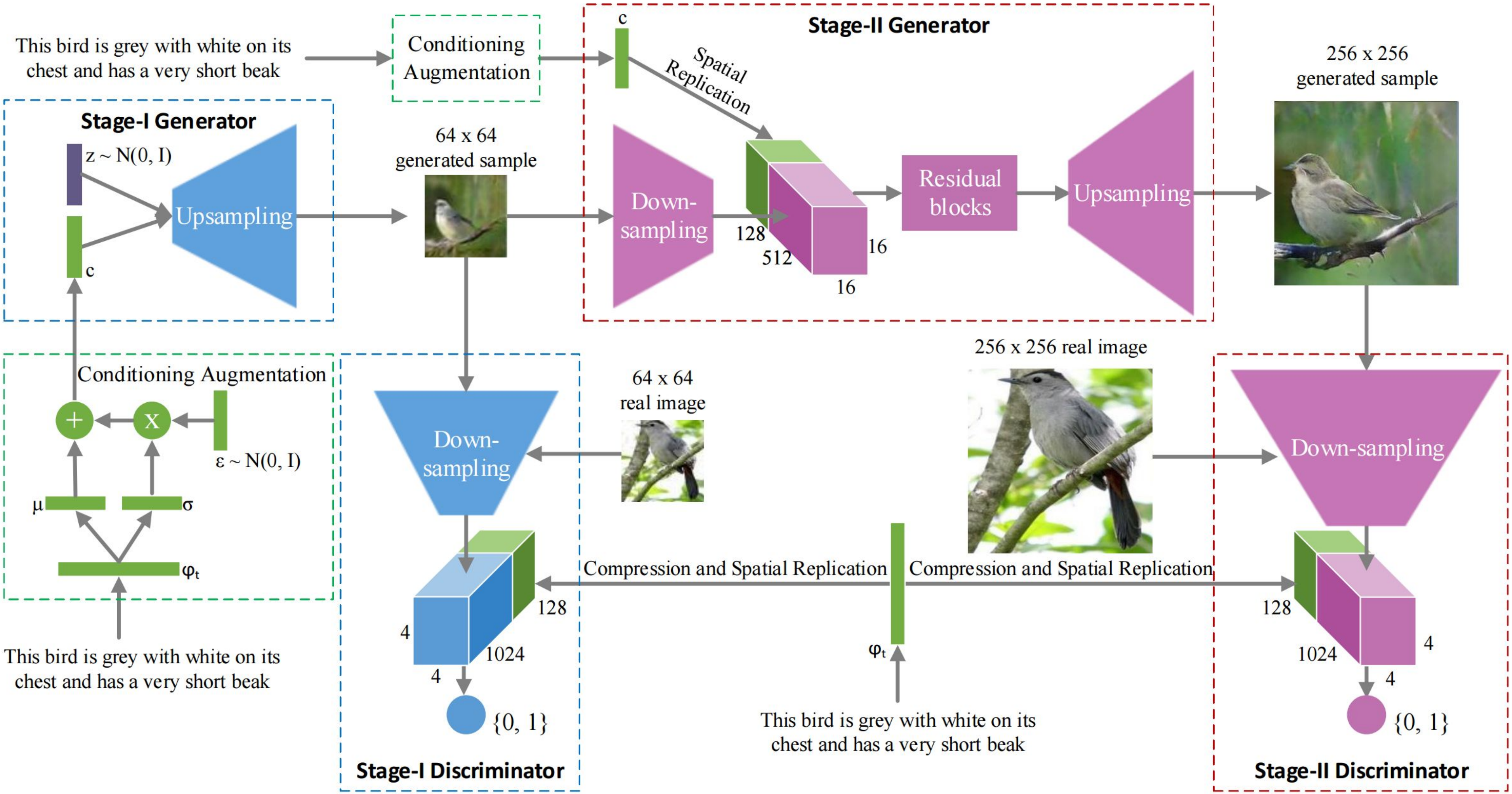
Then alternate:

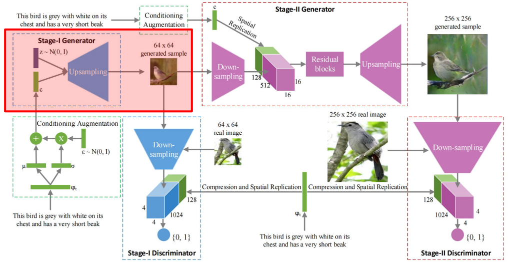
Maximizing

$$\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$$

and minimizing

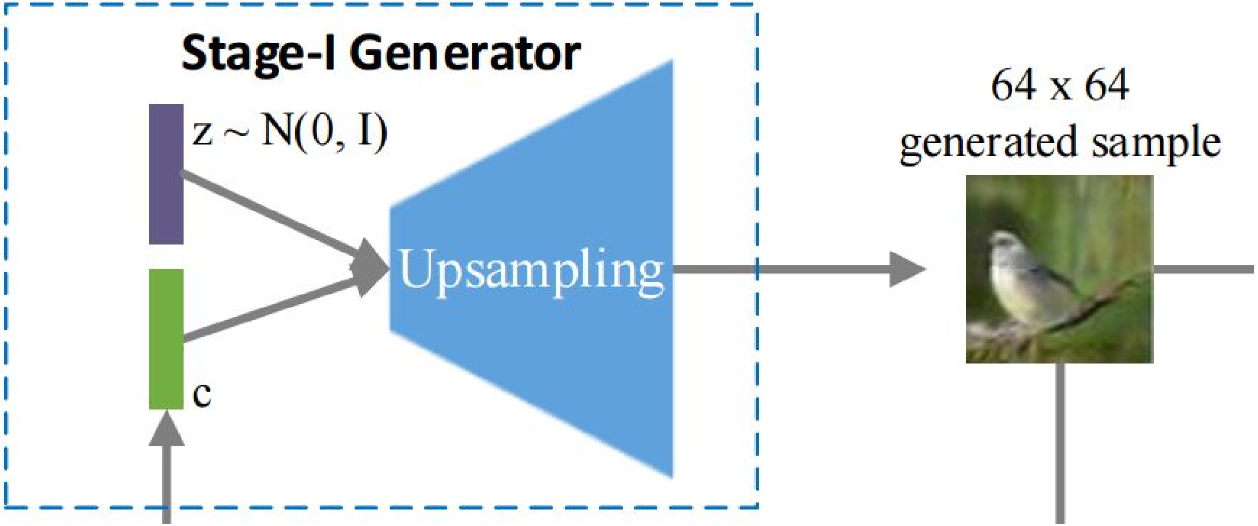
$$\mathcal{L}_G \leftarrow \log(1 - s_f) \quad + \quad \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi_t), \Sigma_0(\varphi_t)) \,||\, \mathcal{N}(0, I))$$


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

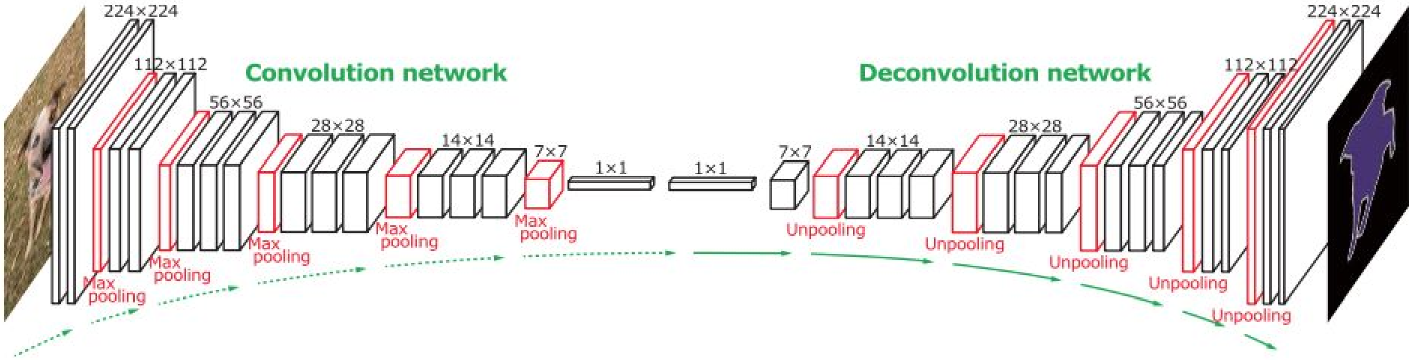

This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
c
Stage-II Generator
Spatial Replication
256 x 256 generated sample
Stage-I Generator
z ~ N(0, I)
Upsampling
c
64 x 64 generated sample
Down-sampling
128
512
16
16
Residual blocks
Upsampling
256 x 256 real image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

Stage-I Generator

- c - vector representing input sentence
- z - noise sampled from a unit gaussian distribution



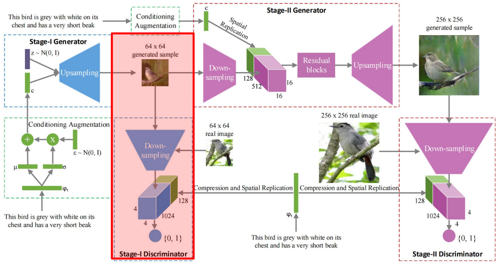
Actually Creating Images

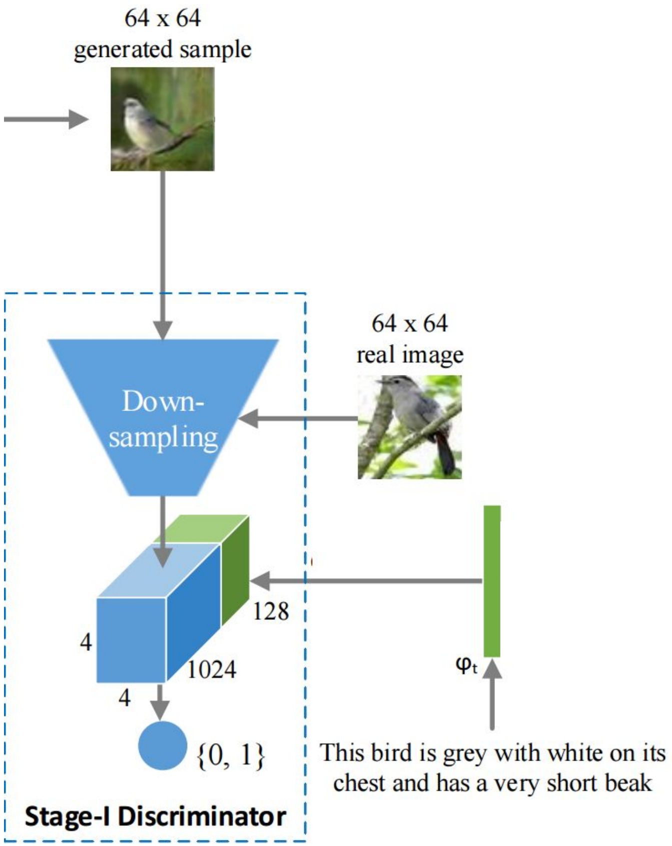


Nice Deconvolution Animation

But really they're upsampling the activation maps using nearest neighbors-- then applying deconvolution

Stage-I Discriminator


This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
Stage-II Generator
c
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
φ_t
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
φ_t
This bird is grey with white on its chest and has a very short beak
128
1024
4
4
{0, 1}
Stage-II Discriminator

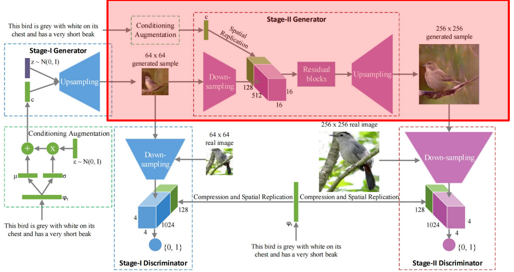


Down-Sampling

- Images
  - Stride-2 convolutions, Batch Norm., Leaky ReLU
  - 64 x 64 x 3 → 4 x 4 x 1024
- Text
  - Fully-connected layer: φ_t → 128
  - Spatially replicate to 4 x 4 x 128
- Depth Concatenate
  - Total of 4 x 4 x 1152

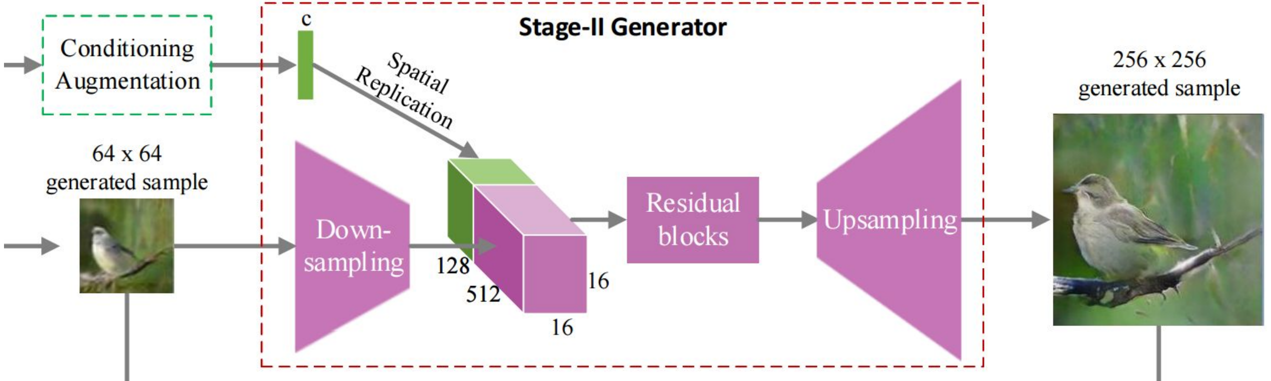
Score

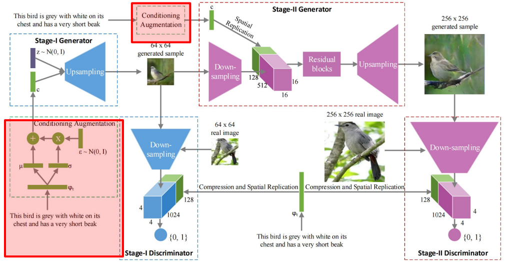
- 1x1 convolution, followed by 4x4 convolution
  - Produces scalar value between 0 and 1


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

Stage-II Generator

- Takes in…
  - 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




This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
Stage-II Generator
c
Spatial Replication
256 x 256 generated sample
Stage-I Generator
z ~ N(0, I)
c
Upsampling
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}
φt
This bird is grey with white on its chest and has a very short beak
128
1024
4
4
{0, 1}
Stage-I Discriminator
Stage-II Discriminator

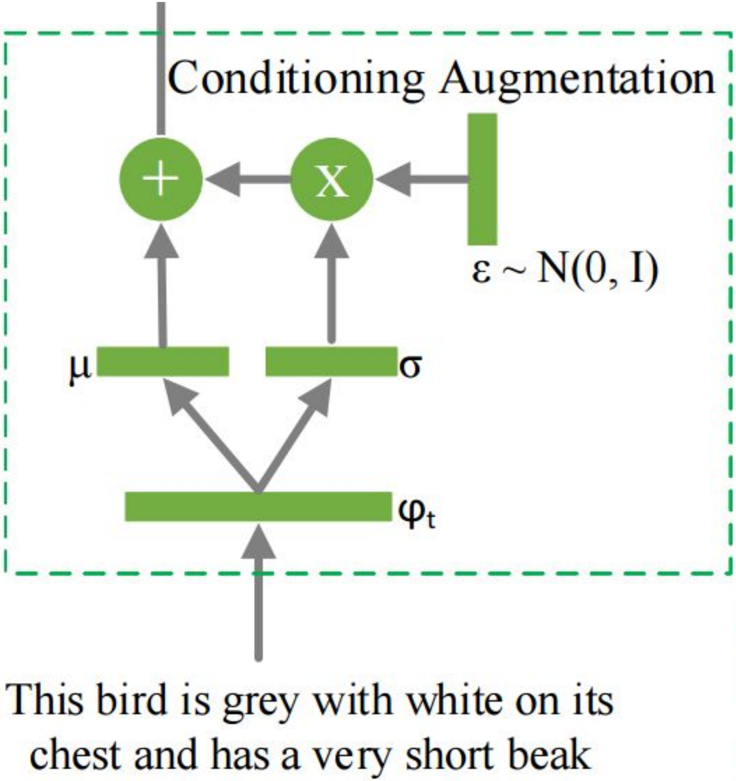
Conditioning Augmentation

Text Encoding

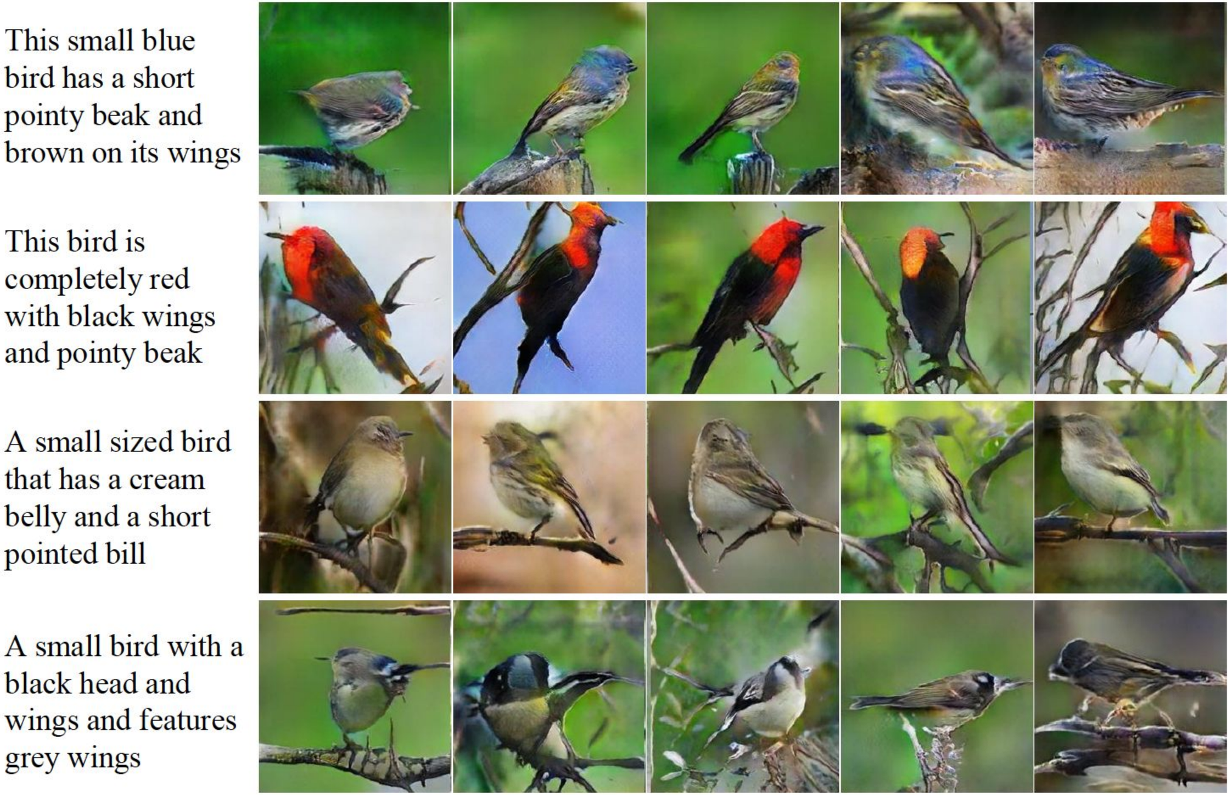
- Uses a “hybrid character-level convolutional recurrent neural network”
- Same as Reed et al. “GAN Text to Image Synthesis” paper

Augmentation

- Randomly sample “latent variables” from the independent Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}(\varphi_t), \boldsymbol{\Sigma}(\varphi_t))$


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 $N(\boldsymbol{\mu}(\varphi_t), \boldsymbol{\Sigma}(\varphi_t))$ actually change between images.

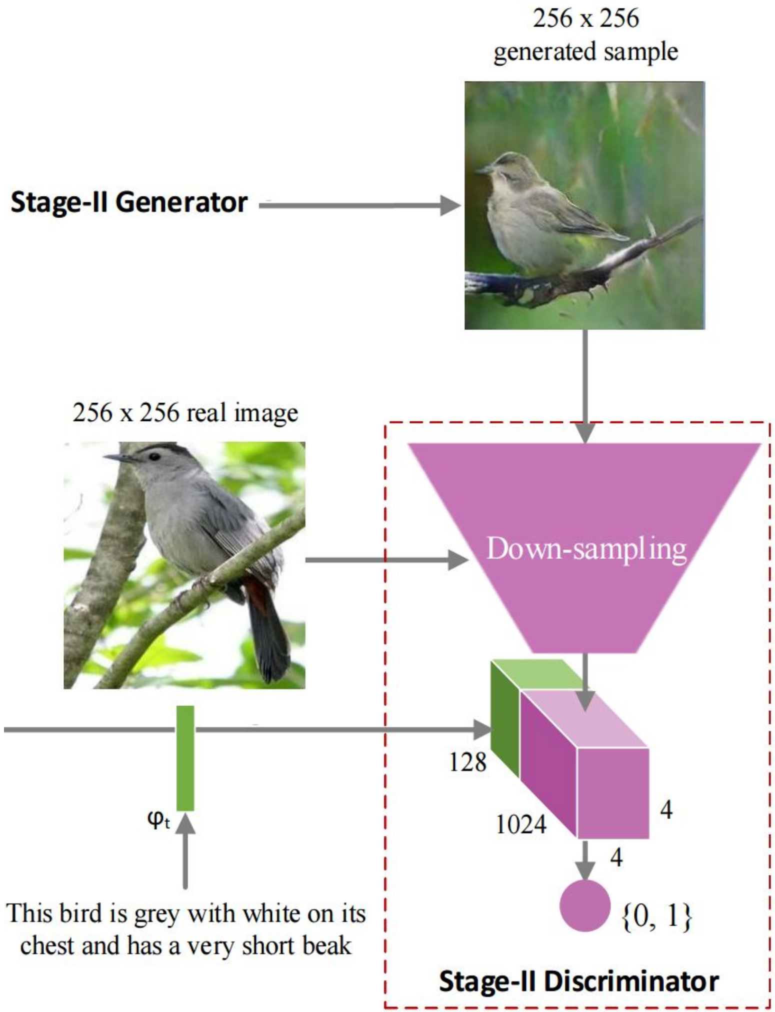
Stage-II Discriminator

Down-sampling

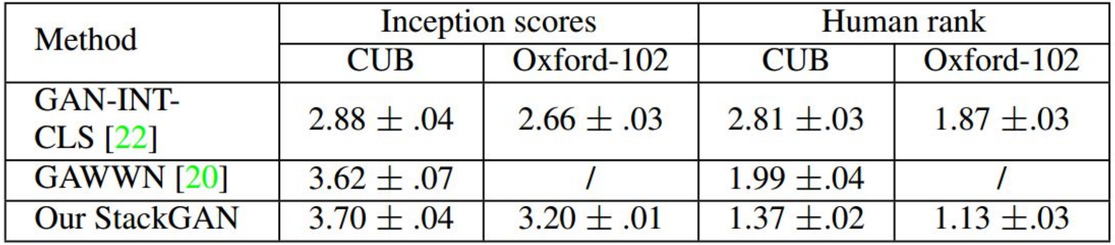
- Same as Stage-I, but more layers

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*.

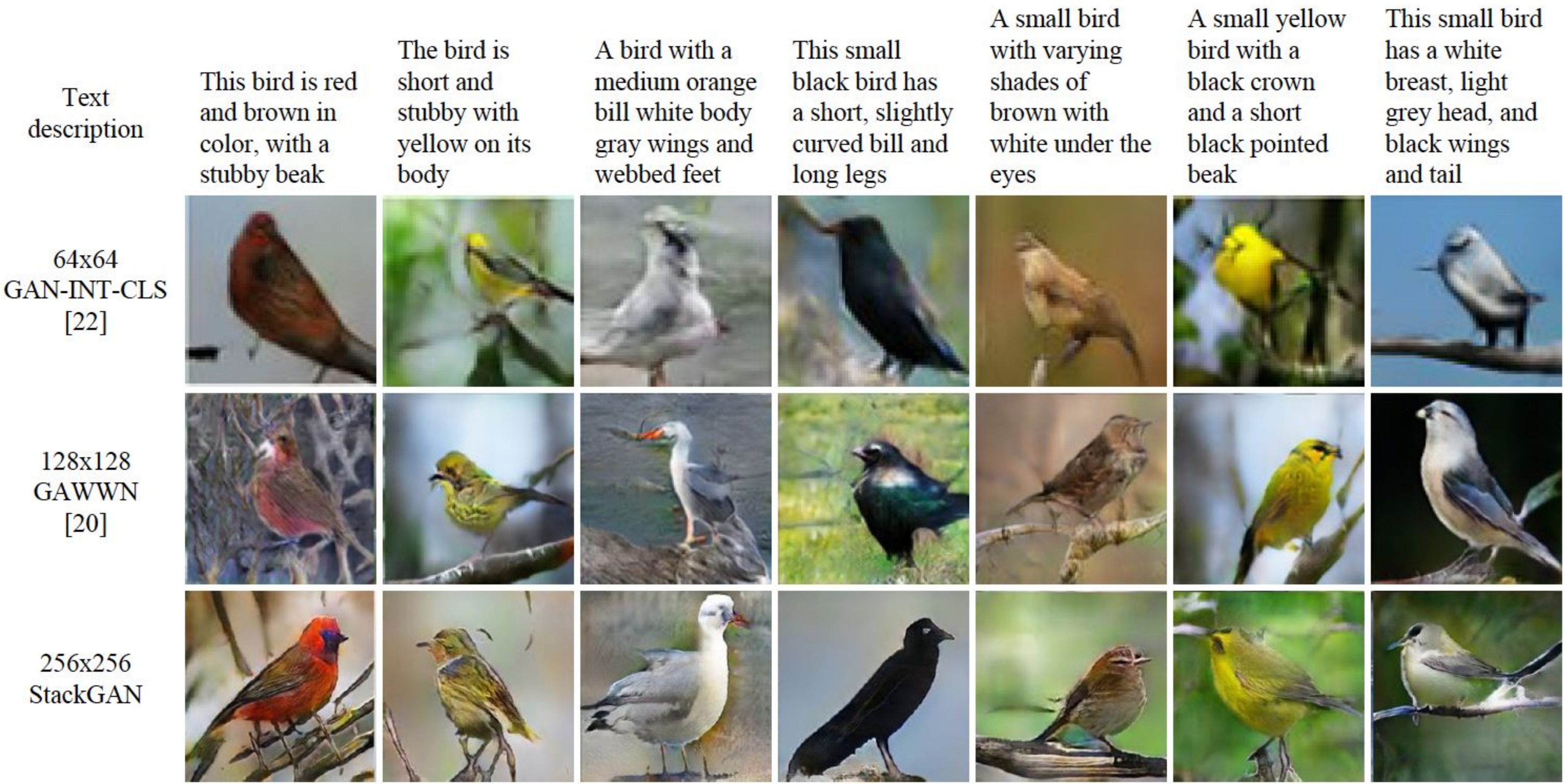


Evaluation



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	/
Our 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


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

Text description	This flower has petals that are white and has pink shading	This flower has a lot of small purple petals in a dome-like configuration	This flower has long thin yellow petals and a lot of yellow anthers in the center	This flower is pink, white, and yellow in color, and has petals that are striped	This flower is white and yellow in color, with petals that are wavy and smooth	This flower has upturned petals which are thin and orange with rounded edges	This flower has petals that are dark pink with white edges and pink stamen