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

Sanghyuck Na

April, 17, 2020

Dongguk University

Artificial Intelligence Laboratory

shna@Dongguk.edu

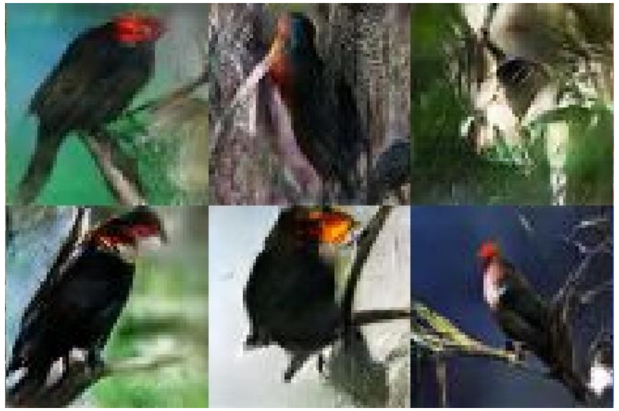
# 0 Contents

1. Introduction
2. GAN-INT-CLS
3. SRGAN
4. StackGAN
5. Result
6. Reference

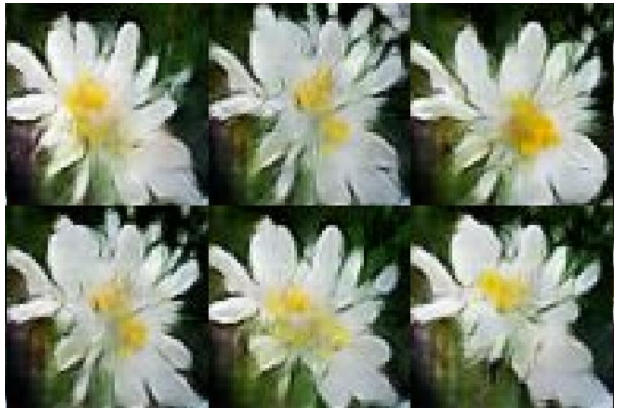
# 1 Introduction

**Captions are from the training set**

this magnificent fellow is almost all black with a red crest, and white cheek patch.



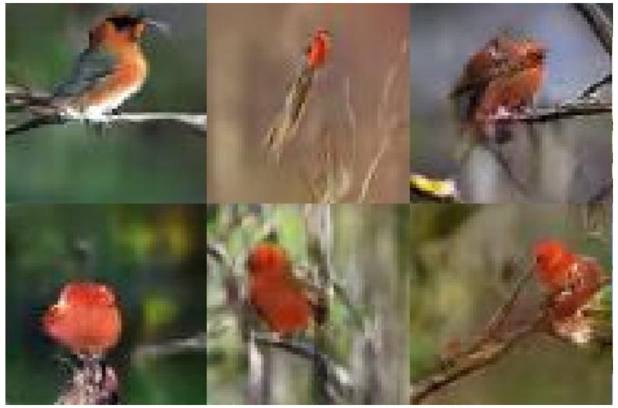
this white and yellow flower have thin white petals and a round yellow stamen.



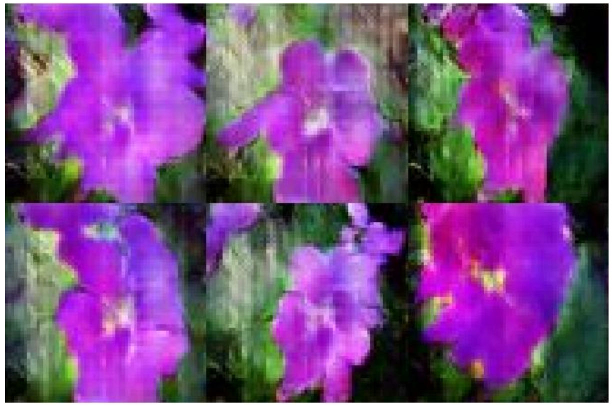
# 1 Introduction

**Captions are from Zero-shot(held out)**

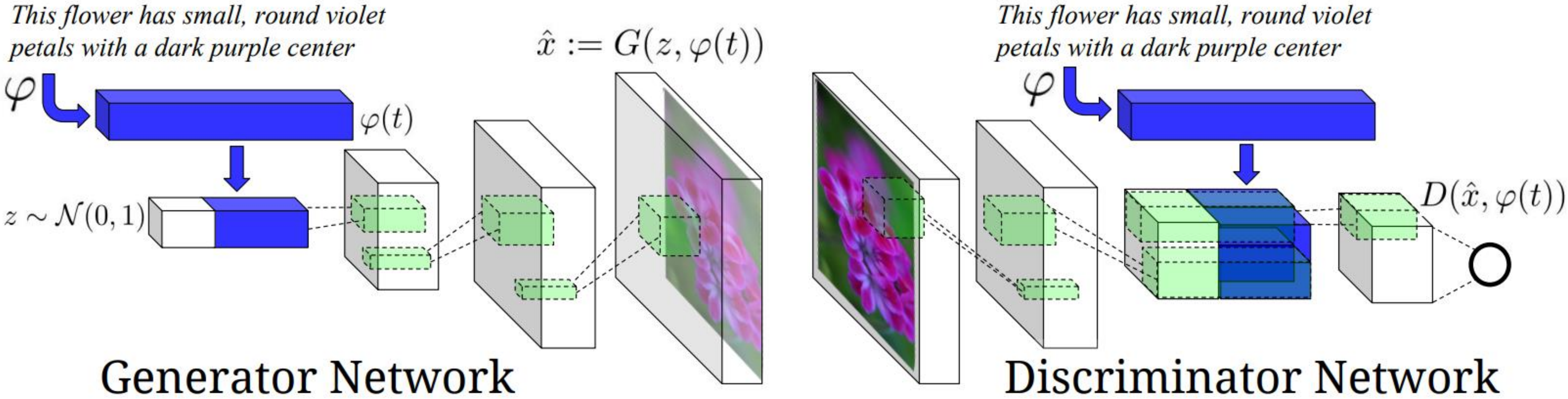
this small bird has a pink breast and crown, and black primaries and secondaries.



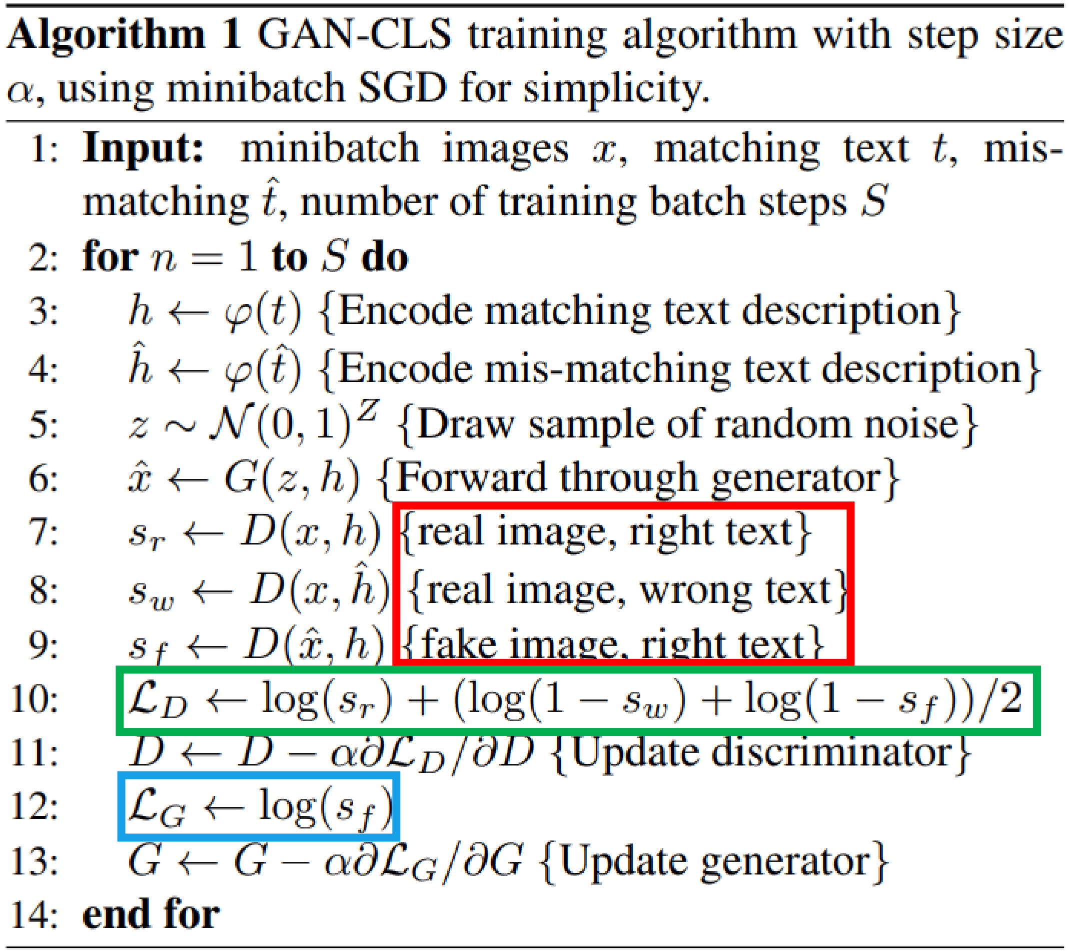
the flower has petals that are bright pinkish purple with white stigma.



# 2 GAN-INT-CLS


This flower has small, round violet petals with a dark purple center
$\varphi$
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
Generator Network
This flower has small, round violet petals with a dark purple center
$\varphi$
$D(\hat{x}, \varphi(t))$
Discriminator Network

# 2 GAN-INT-CLS



**Algorithm 1** GAN-CLS training algorithm with step size $\alpha$, using minibatch SGD for simplicity.

```
Input: minibatch images $x$, matching text $t$, mis-matching $\hat{t}$, number of training batch steps $S$
for $n = 1$ to $S$ do
    $h \leftarrow \varphi(t)$ {Encode matching text description}
    $\hat{h} \leftarrow \varphi(\hat{t})$ {Encode mis-matching text description}
    $z \sim \mathcal{N}(0,1)^Z$ {Draw sample of random noise}
    $\hat{x} \leftarrow G(z, h)$ {Forward through generator}
    $s_r \leftarrow D(x, h)$ {real image, right text}
    $s_w \leftarrow D(x, \hat{h})$ {real image, wrong text}
    $s_f \leftarrow D(\hat{x}, h)$ {fake image, right text}
    $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$
    $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ {Update discriminator}
    $\mathcal{L}_G \leftarrow \log(s_f)$
    $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ {Update generator}
end for
```

GAN-CLS

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

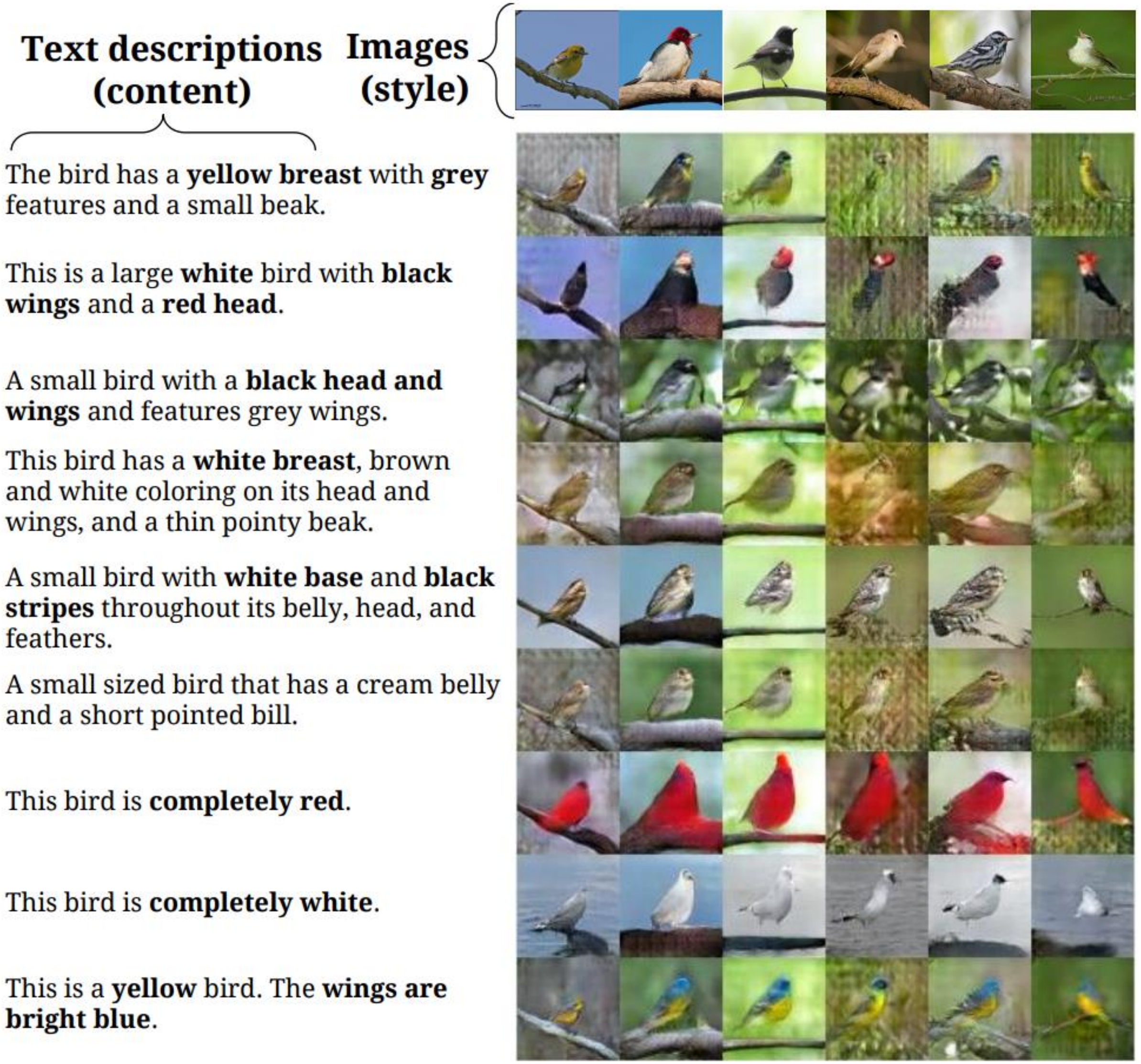
GAN-INT

$$\mathbb{E}_{t_1,t_2\sim p_{data}}[log(1-D(G(\beta t_1+(1-\beta)t_2)))]$$

Style Transfer

$$\mathbb{E}_{t,z\sim N(0,1)} ||z\text{-}S(\underbrace{G(z,\varphi(t))}_{\hat{x}})||_2^2$$

# 2 GAN-INT-CLS



Text descriptions (content)

Images (style)

The bird has a **yellow breast** with **grey** features and a small beak.

This is a large **white** bird with **black wings** and a **red head**.

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

This bird has a **white breast**, brown and white coloring on its head and wings, and a thin pointy beak.

A small bird with **white base** and **black stripes** throughout its belly, head, and feathers.

A small sized bird that has a cream belly and a short pointed bill.

This bird is **completely red**.

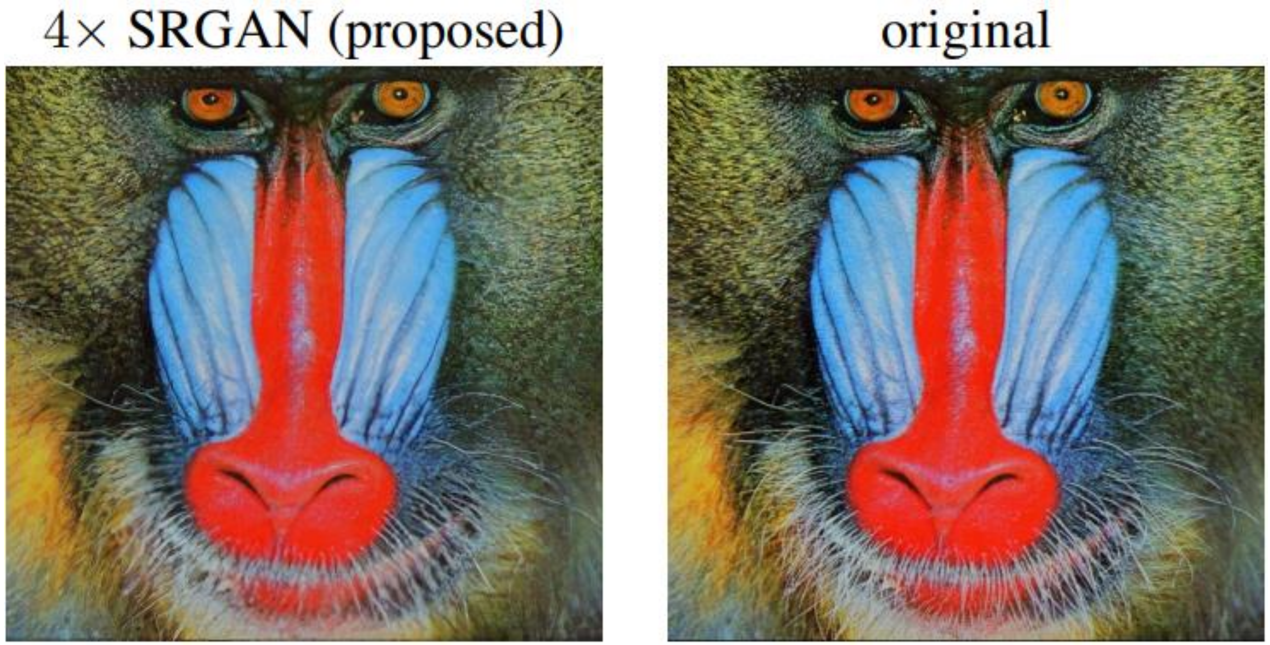
This bird is **completely white**.

This is a **yellow** bird. The **wings are bright blue**.

Transferring style from the top row (real) images to the content from the query text, with G acting as a deterministic decoder.

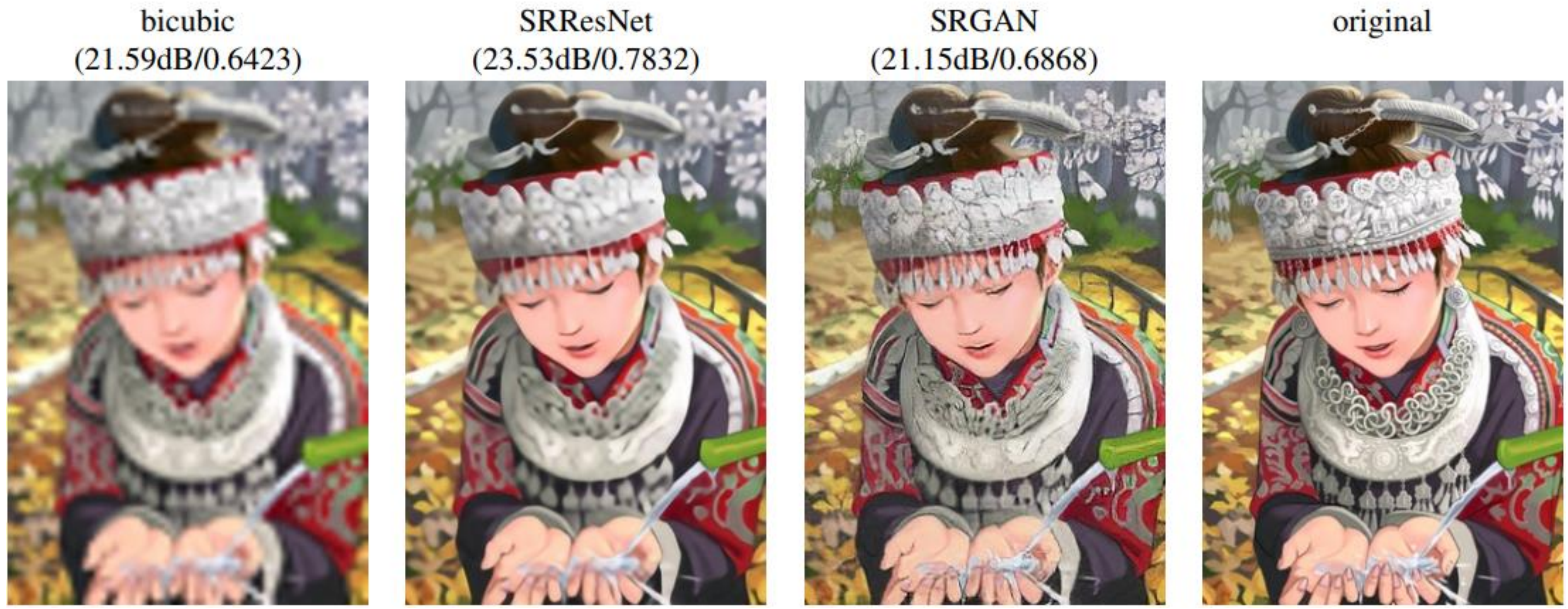
The bottom three rows are captions made up by us.

# 3 SRGAN



The task of estimating high-resolution (HR) images from low-resolution (LR) counterpart is referred to as super-resolution (SR).

# 3 SRGAN



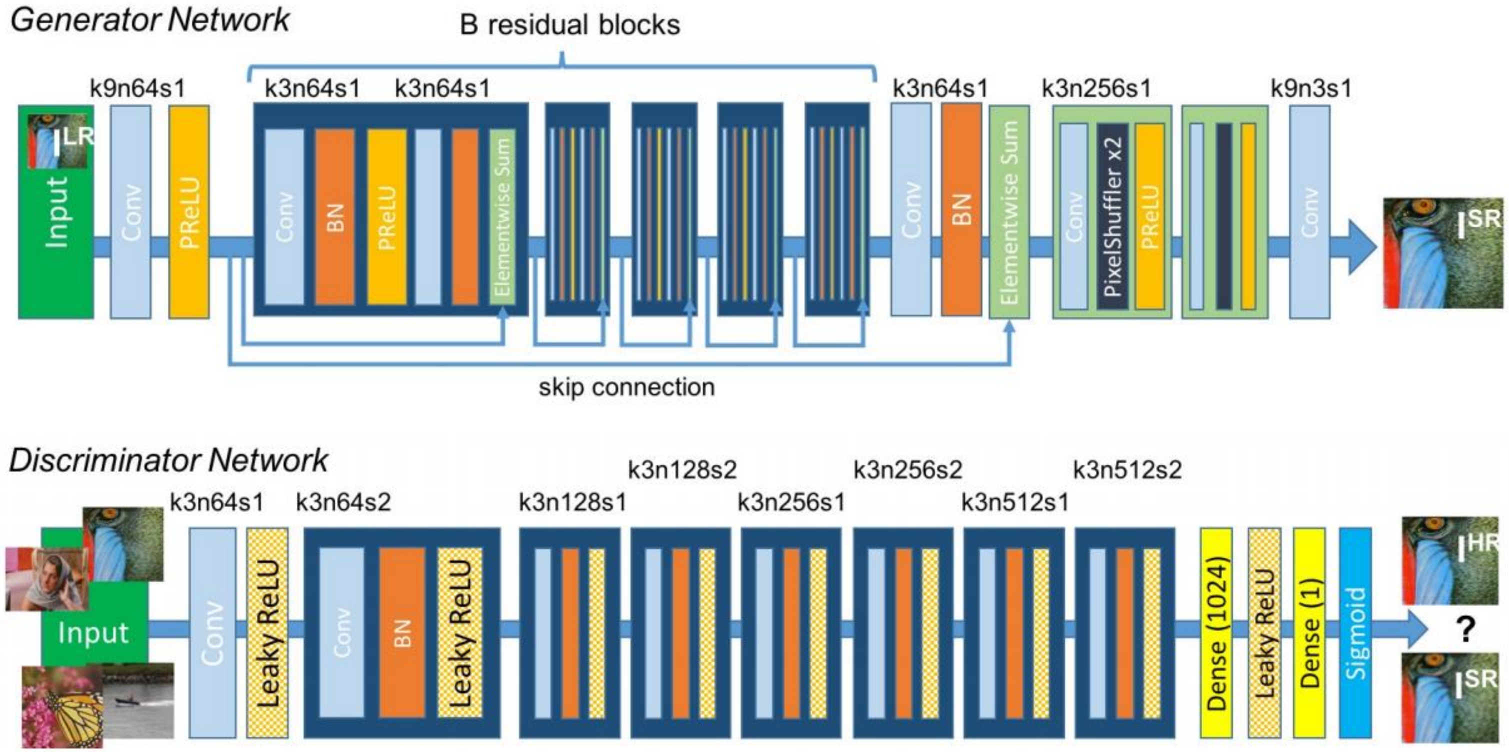
bicubic
(21.59dB/0.6423)

SRResNet
(23.53dB/0.7832)

SRGAN
(21.15dB/0.6868)

original

# 3 SRGAN



$$k \;=\; kernel\ size$$
$$n \;=\; number\ of\ feature\ maps$$
$$s \;=\; stride$$

$$\min_{\theta_G}\max_{\theta_D}\mathbb{E}_{I^{HR}\sim p_{train}(I^{HR})}\left[log\left(D_{\theta_D}(I^{HR})\right)\right] + \mathbb{E}_{I^{LR}\sim p_G(I^{LR})}\left[log\left(1 - D_{\theta_D}\left(G_{\theta_G}(I^{LR})\right)\right)\right]$$

$$\hat{\theta}_G = \underset{\theta_G}{argmin}\frac{1}{N}\sum_{n=1}^{N} l^{SR}\left(G_{\theta_G}(I_n^{LR}), l_n^{HR}\right)$$

# 3 SRGAN

$$\hat{\theta}_G = \underset{\theta_G}{argmin}\frac{1}{N}\sum_{n=1}^{N} l^{SR}\left(G_{\theta_G}(I_n^{LR}), l_n^{HR}\right)$$

$perceptual\ loss$

$$l^{SR} = \underbrace{l_x^{SR}}_{Content\ loss} + \underbrace{10^{-3} l_{Gen}^{SR}}_{adversarial\ loss}$$

$adversarial\ loss$

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -log\, D_{\theta_D}(G_{\theta_G}(I^{LR}))$$

$Content\ loss$

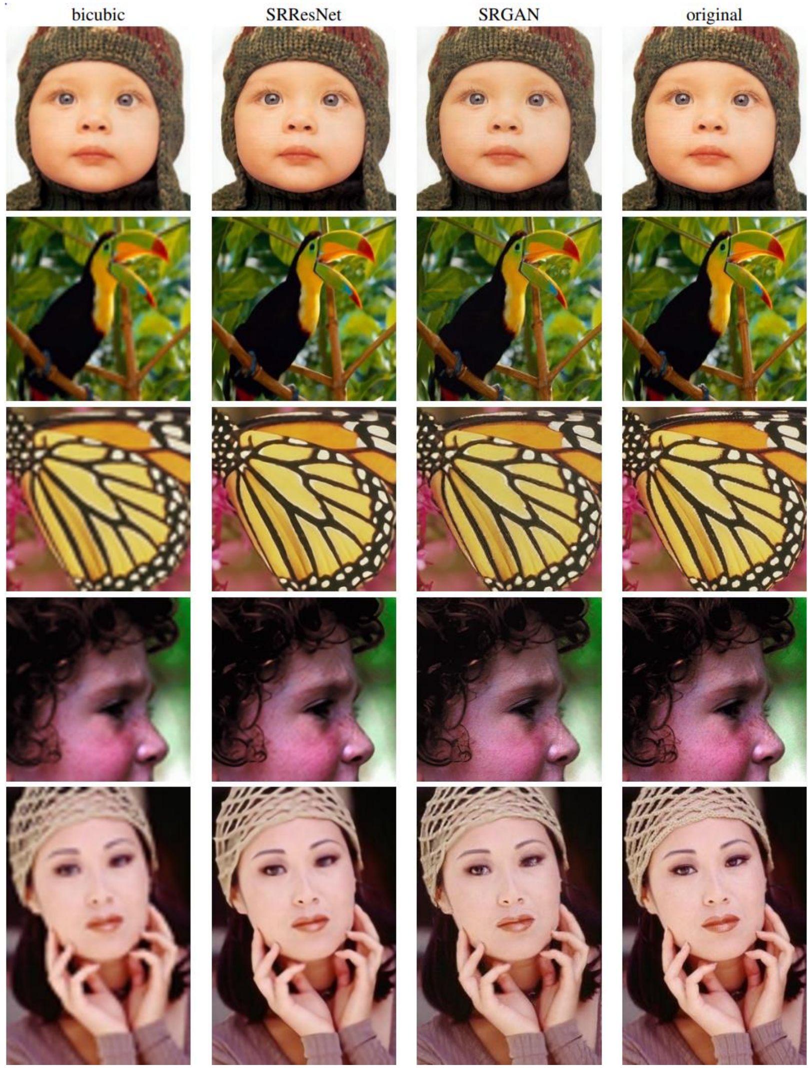
$$l_{MSE}^{SR} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2$$

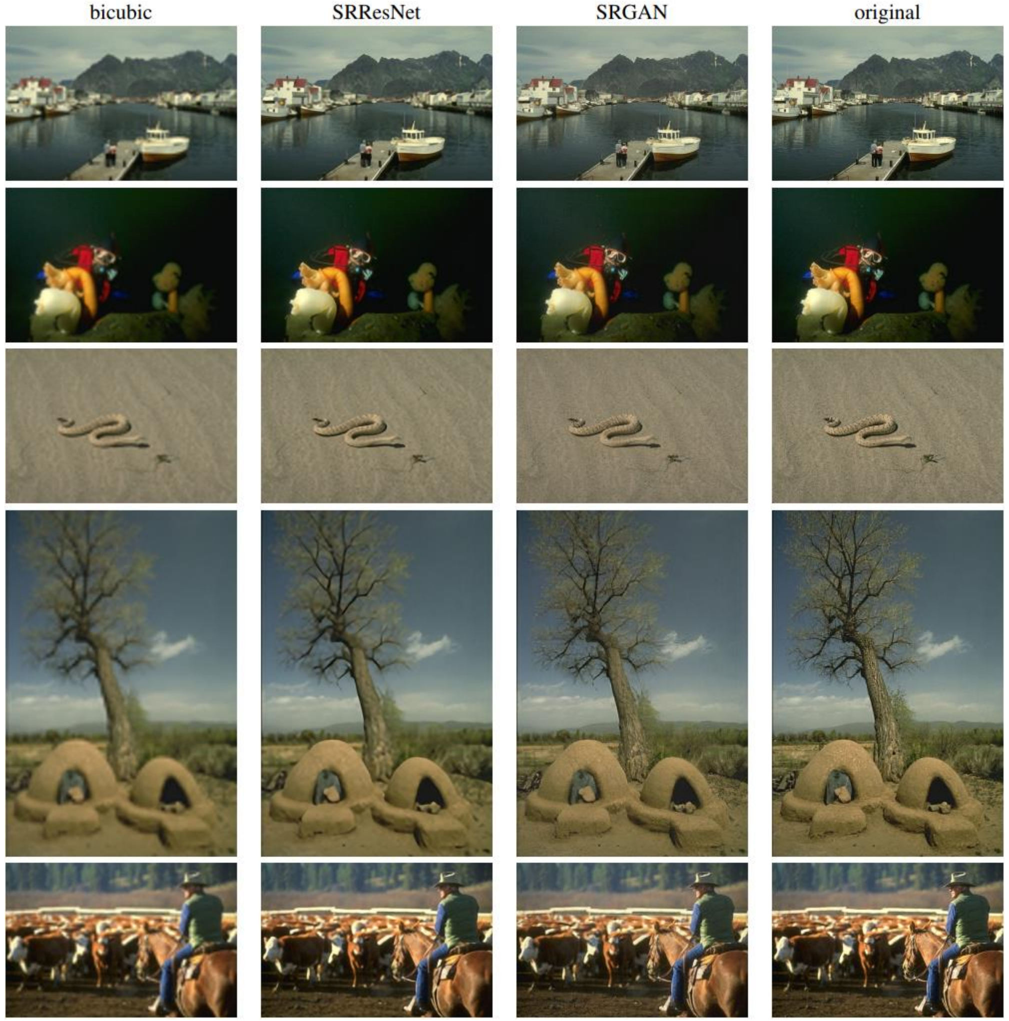
$$l_{VGG/i,j}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\emptyset_{i,j}(I^{HR})_{x,y}-\emptyset_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2$$

$r = down\ sampling\ factor$

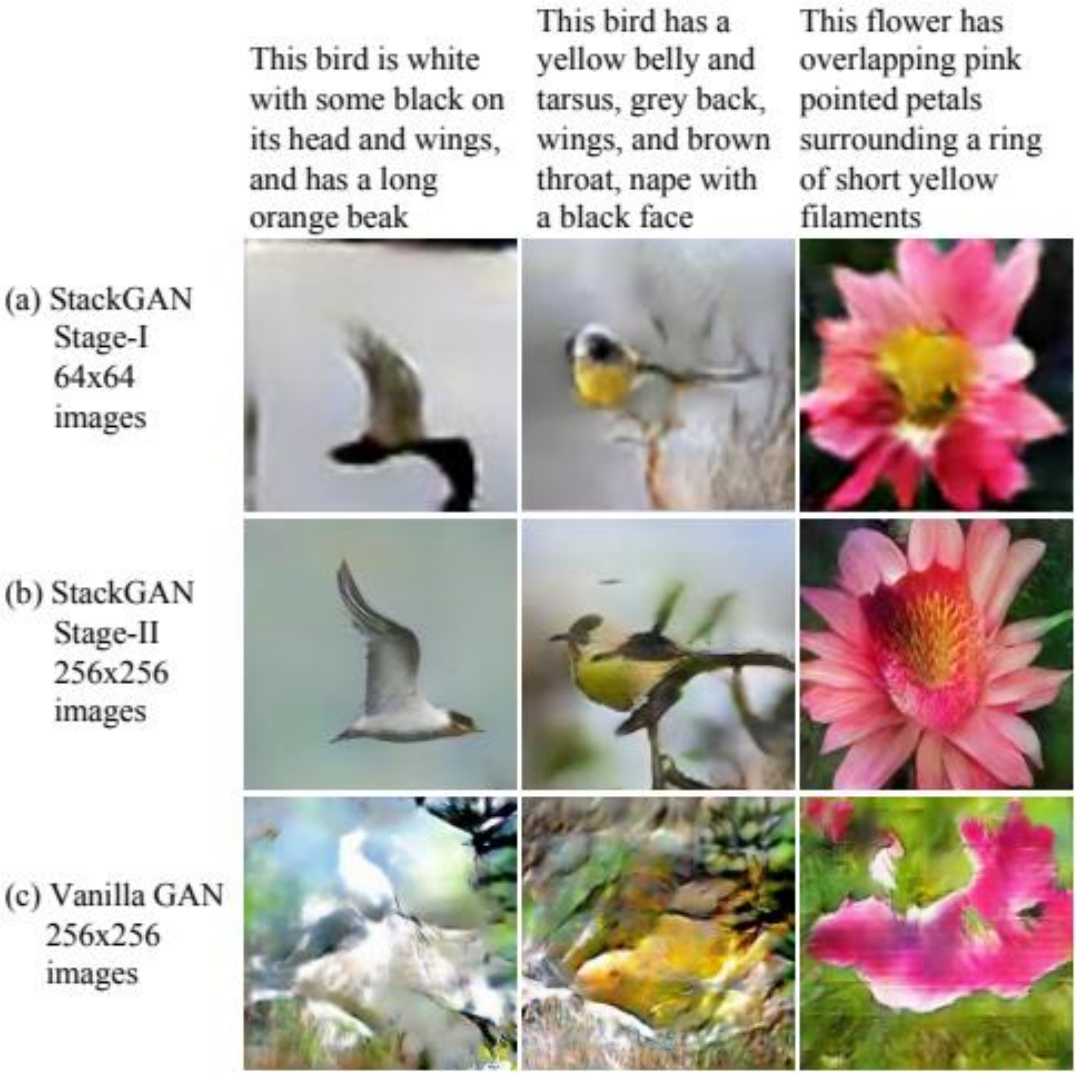
$\emptyset_{i,j}$ =the feature map obtained by the j-th convolution (after activation) before the i-th maxpooling layer within the VGG19 network,

## 3 SRGAN


bicubic
SRResNet
SRGAN
original


bicubic
SRResNet
SRGAN
original

# 4 StackGAN



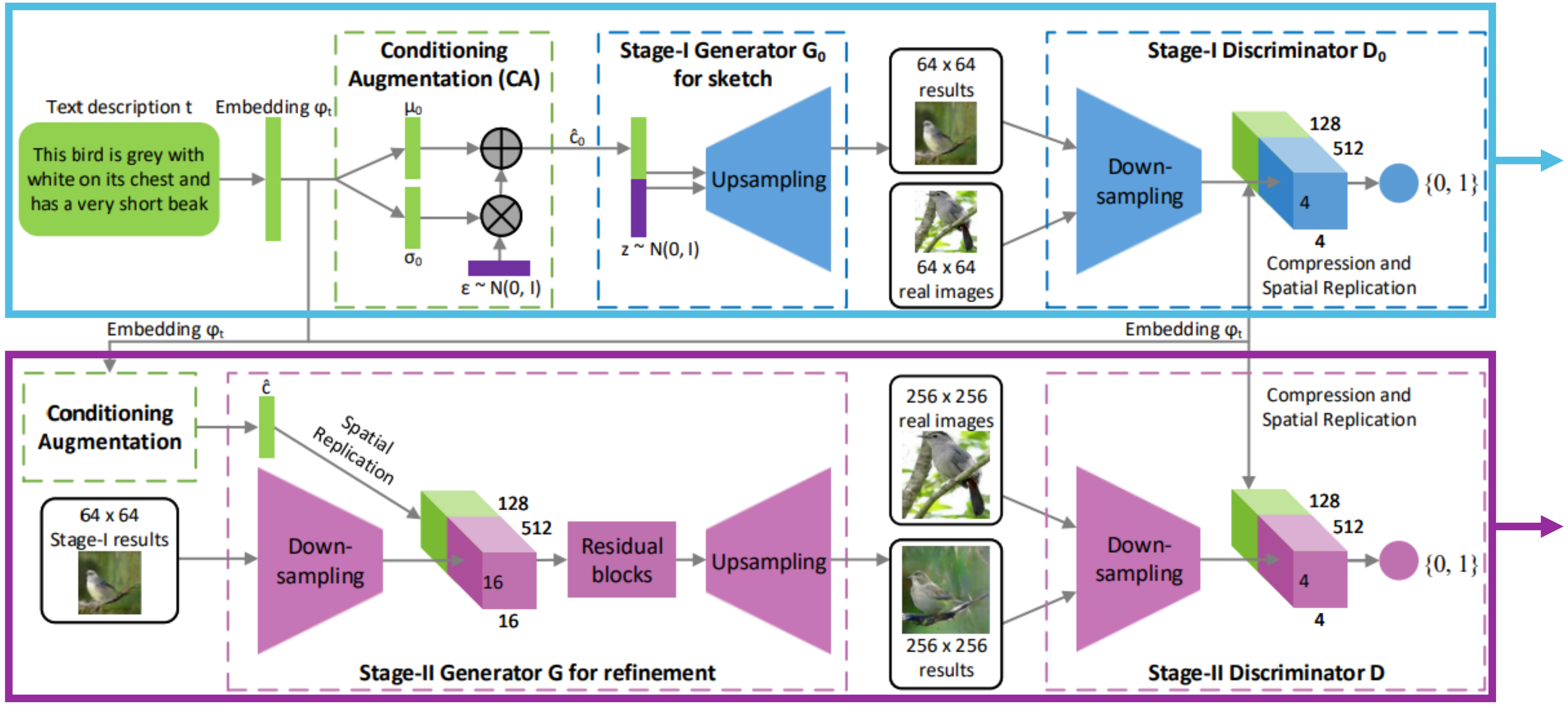
This bird is white with some black on its head and wings, and has a long orange beak

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

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

(a) StackGAN Stage-I 64x64 images
(b) StackGAN Stage-II 256x256 images
(c) Vanilla GAN 256x256 images

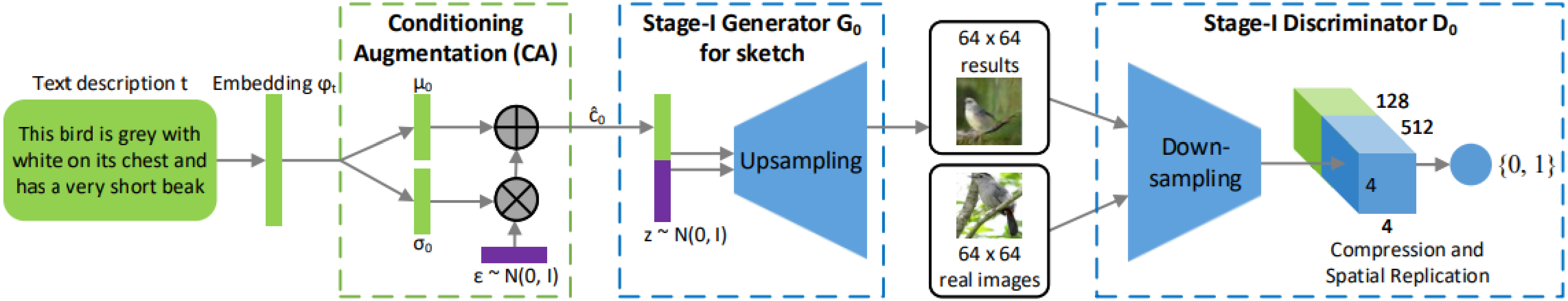
# 4 StackGAN



Stage-I GAN: 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 image.

Stage-II GAN: 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 photo-realistic image.

# 4 StackGAN



Text description t
This bird is grey with white on its chest and has a very short beak
Embedding φt
Conditioning Augmentation (CA)
μ0
σ0
ε ~ N(0, I)
ĉ0
Stage-I Generator G0 for sketch
z ~ N(0, I)
Upsampling
64 x 64 results
64 x 64 real images
Stage-I Discriminator D0
Down-sampling
128
512
4
4
Compression and Spatial Replication
{0, 1}

$$Conditioning\ Augmentation\ (CA)$$

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

$t$ : *text description*

$z$ : *noise vector from Gaussian Distribution*

$\varphi_t$ : *text embedding networks (pre – trained)*

$\hat{c}_0$ : *conditioning variable*

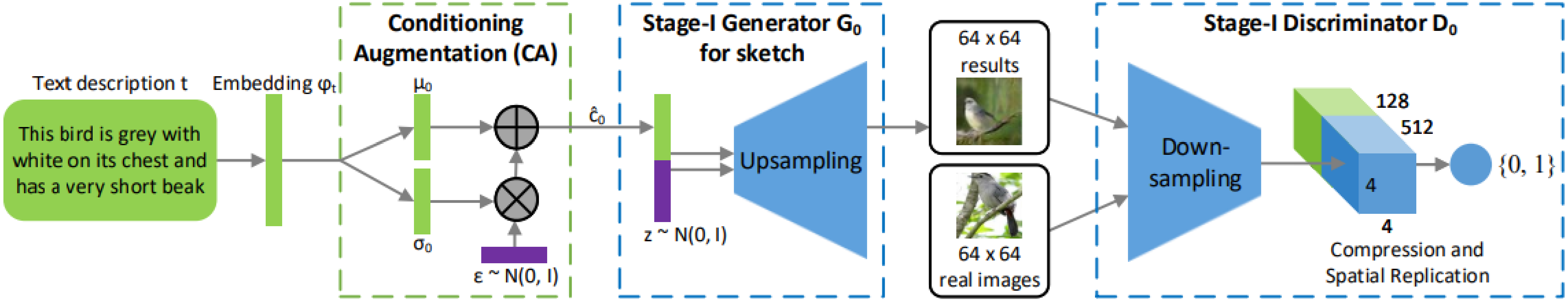
$\mathcal{N}(\mu(\varphi_t),\Sigma(\varphi_t))$ : *conditioning Gaussian distribution*

$\mathcal{N}(0,I)$ : *normal distribution*

$\Sigma(\varphi_t)$ : *diagonal covariance matrix*

$s_0$: *image generated by the Stage*-I

# 4 StackGAN


Text description t
This bird is grey with white on its chest and has a very short beak
Embedding $\varphi_t$
Conditioning Augmentation (CA)
$\mu_0$
$\sigma_0$
$\varepsilon$ ~ N(0, I)
$\hat{c}_0$
Stage-I Generator $G_0$ for sketch
z ~ N(0, I)
Upsampling
64 x 64 results
64 x 64 real images
Stage-I Discriminator $D_0$
Down-sampling
128
512
4
4
Compression and Spatial Replication
{0, 1}

$$L_{D_0} = \mathbb{E}_{(I_0,\mathrm{t})\sim p_{data}}[logD_0(I_0, \varphi_t)]$$
$$+\ \mathbb{E}_{z\sim p_{z,},t\sim p_{data,}}[\log(1 - D_0(G_0(z, \hat{c}_0, \varphi_t))]$$

$$L_{G_0} = \mathbb{E}_{z\sim p_{z,},t\sim p_{data,}}[\log(1 - D_0(G_0(z, \hat{c}_0, \varphi_t))]$$
$$+\ \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi_t), \Sigma_0(\varphi_t))\|\mathcal{N}(0, I))$$

$t$ : *text description*

$z$ : *noise vector from Gaussian Distribution*

$\varphi_t$ : *text embedding networks (pre − trained)*

$\hat{c}_0$ : *conditioning variable*

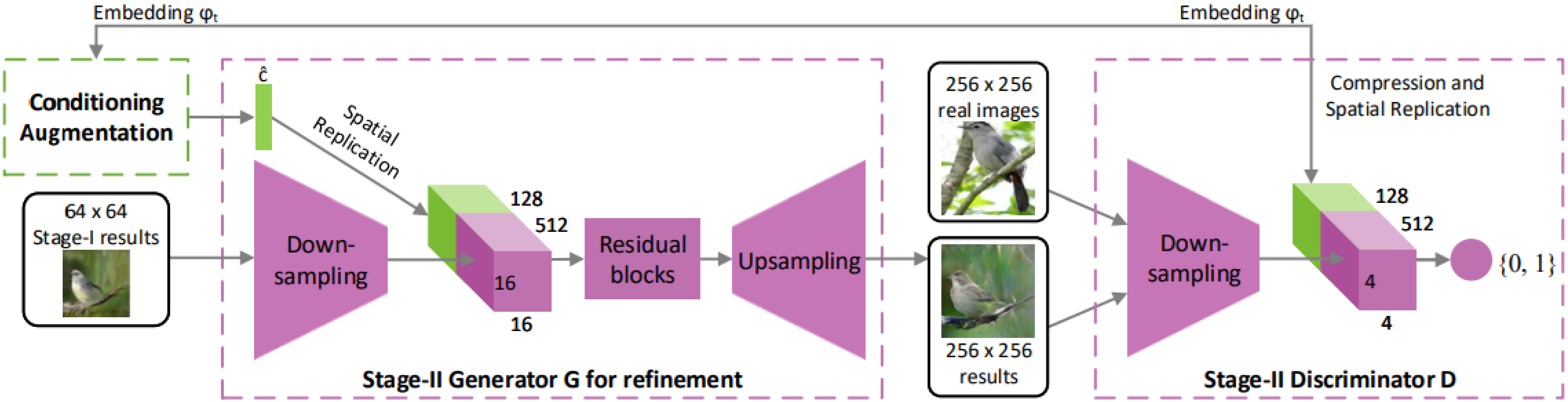
$\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t))$ : *conditioning Gaussian distribution*

$\mathcal{N}(0, I)$ : *normal distribution*

$\Sigma(\varphi_t)$ : *diagonal covariance matrix*

$s_0$: *image generated by the Stage*-I

# 4 StackGAN



$$L_D = \mathbb{E}_{(I,\mathrm{t})\sim p_{data}}[logD(I,\varphi_t)] + \mathbb{E}_{s_0\sim p_{G_0},,t\sim p_{data},}[\log(1-D(G\ (s_0,\hat{c}_0),\varphi_t))]$$

$$L_G = \mathbb{E}_{s_0\sim p_{G_0},,t\sim p_{data},}[\log(1-D(G\ (s_0,\hat{c}),\varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t),\textstyle\sum(\varphi_t))\|\mathcal{N}(0,I))$$

$t$ : *text description*

$z$ : *noise vector from Gaussian Distribution*

$\varphi_t$ : *text embedding networks (pre − trained)*

$\hat{c}_0$ : *conditioning variable*

$\mathcal{N}(\mu(\varphi_t),\sum(\varphi_t))$ : *conditioning Gaussian distribution*

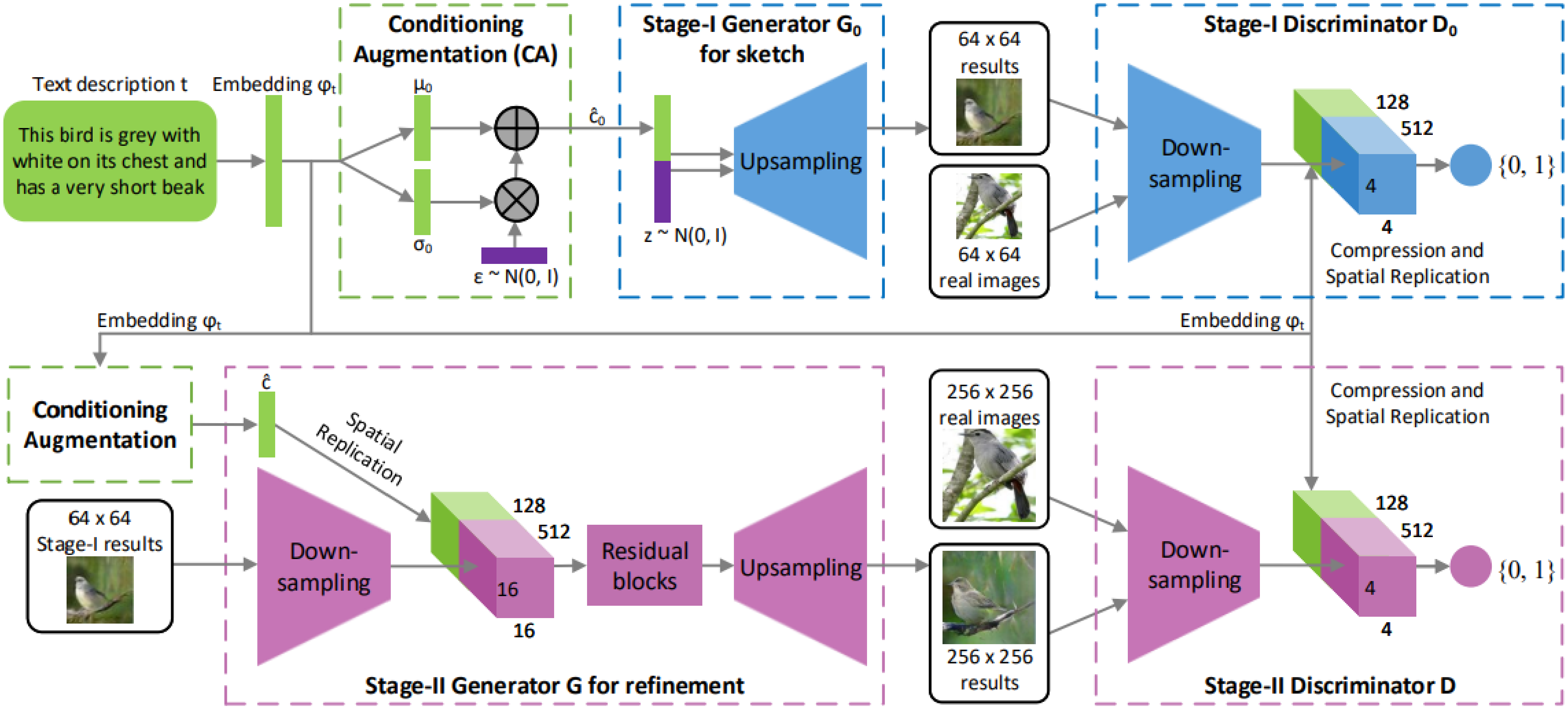
$\mathcal{N}(0,I)$ : *normal distribution*

$\sum(\varphi_t)$ : *diagonal covariance matrix*

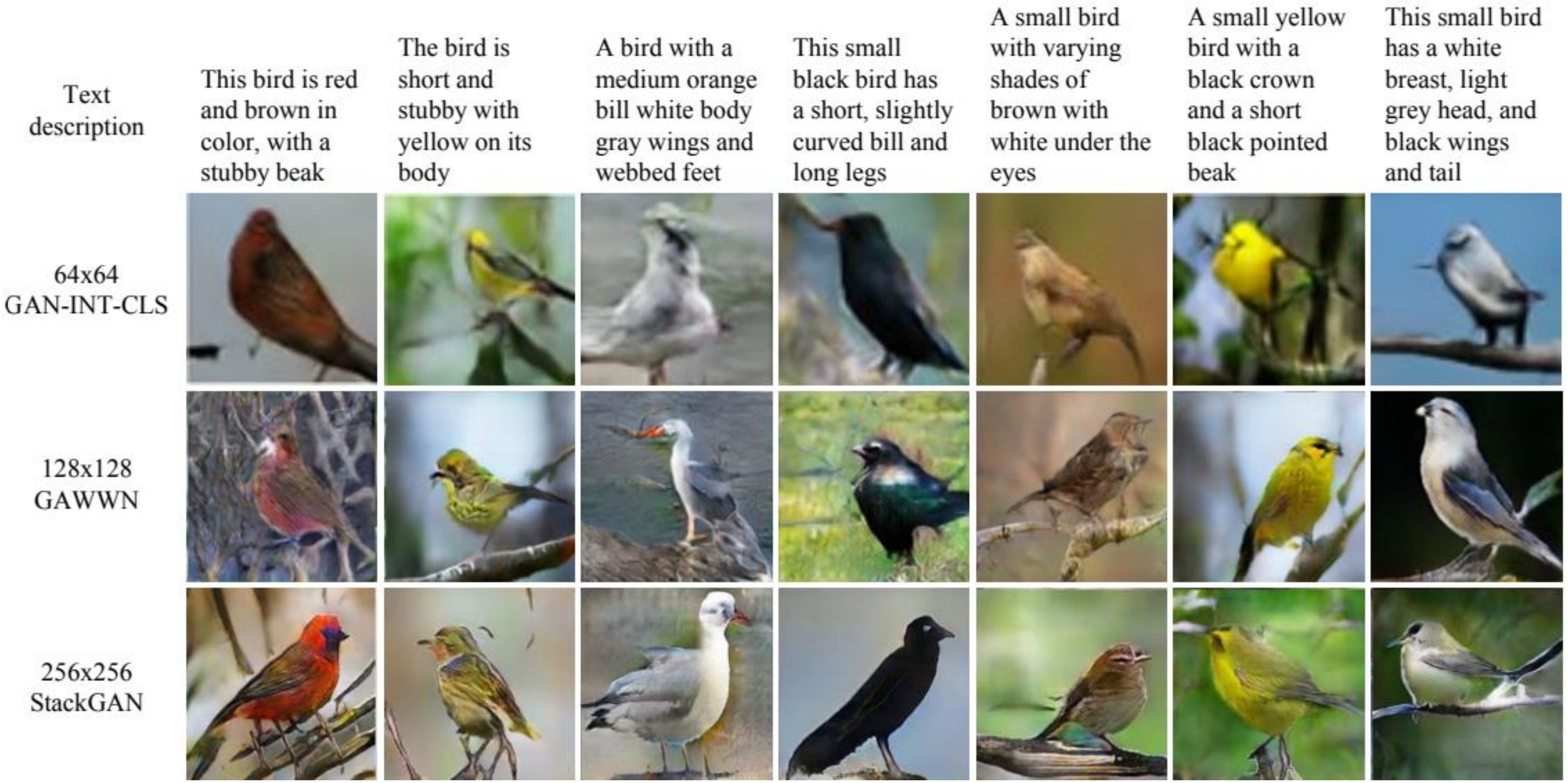
$s_0$: *image generated by the Stage*-I

4

# StackGAN


Text description t
This bird is grey with white on its chest and has a very short beak
Embedding φt
Conditioning Augmentation (CA)
μ0
σ0
ε ~ N(0, I)
ĉ0
Stage-I Generator G0 for sketch
z ~ N(0, I)
Upsampling
64 x 64 results
64 x 64 real images
Stage-I Discriminator D0
Down-sampling
128
512
4
4
{0, 1}
Compression and Spatial Replication
Embedding φt
Embedding φt
Conditioning Augmentation
ĉ
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

# 5 Result



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

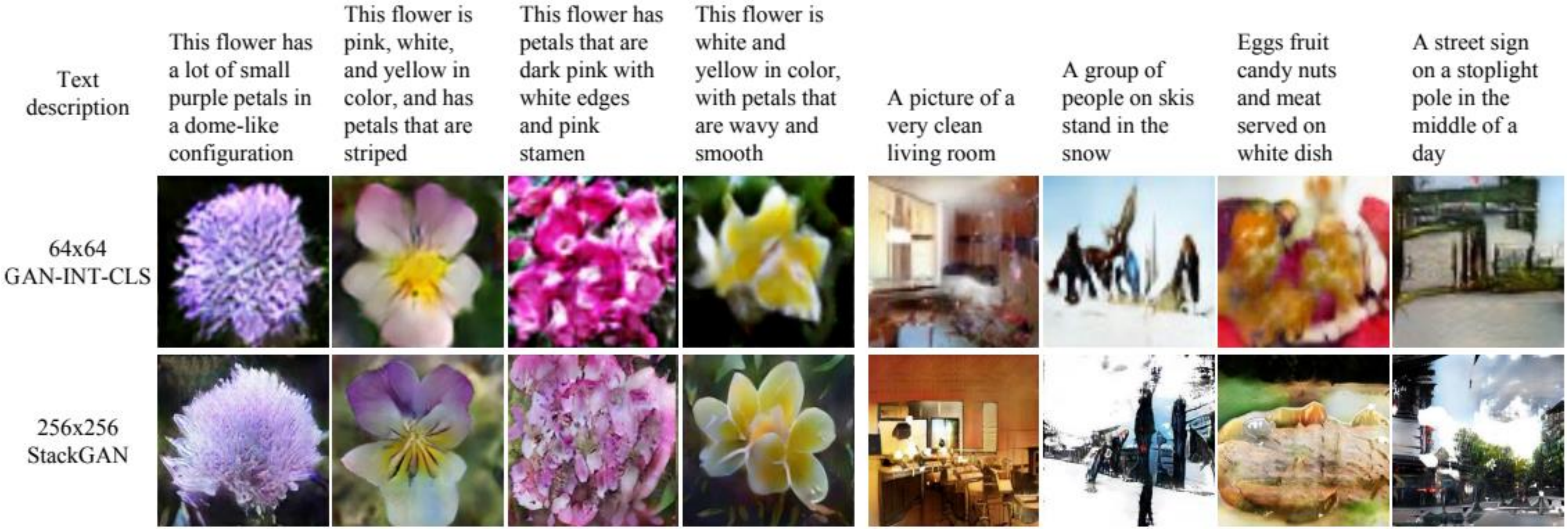
A small bird with varying shades of brown with white under the eyes

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

Example results by our StackGAN, GAWWN, and GAN-INT-CLS conditioned on text descriptions from CUB test set

# 5 Result



This flower has a lot of small purple petals in a dome-like configuration

This flower is pink, white, and yellow in color, and has petals that are striped

This flower has petals that are dark pink with white edges and pink stamen

This flower is white and yellow in color, with petals that are wavy and smooth

A picture of a very clean living room

A group of people on skis stand in the snow

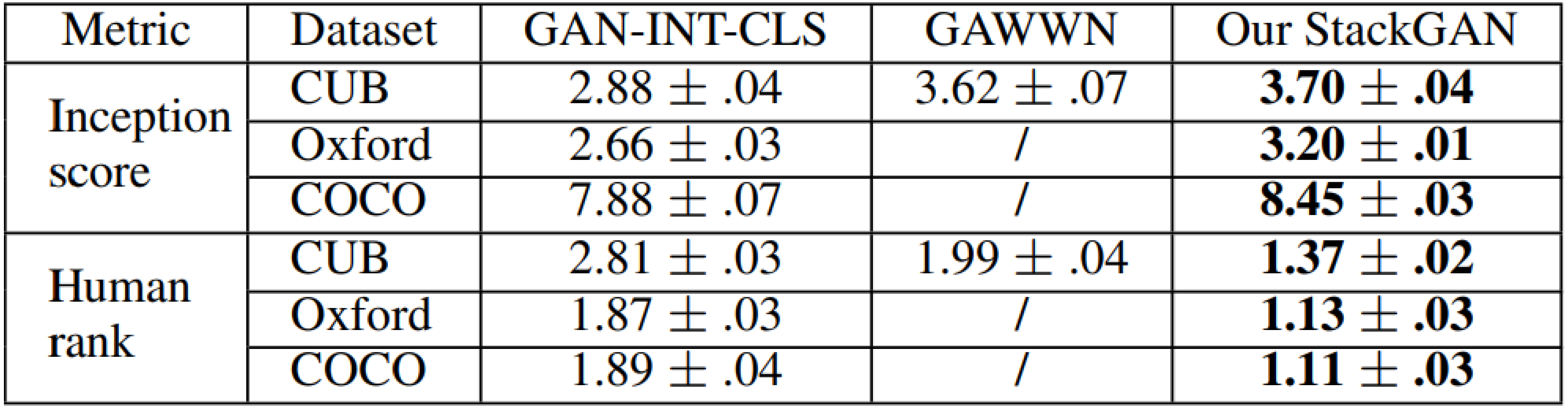
Eggs fruit candy nuts and meat served on white dish

A street sign on a stoplight pole in the middle of a day

Oxford-102

MS COCO

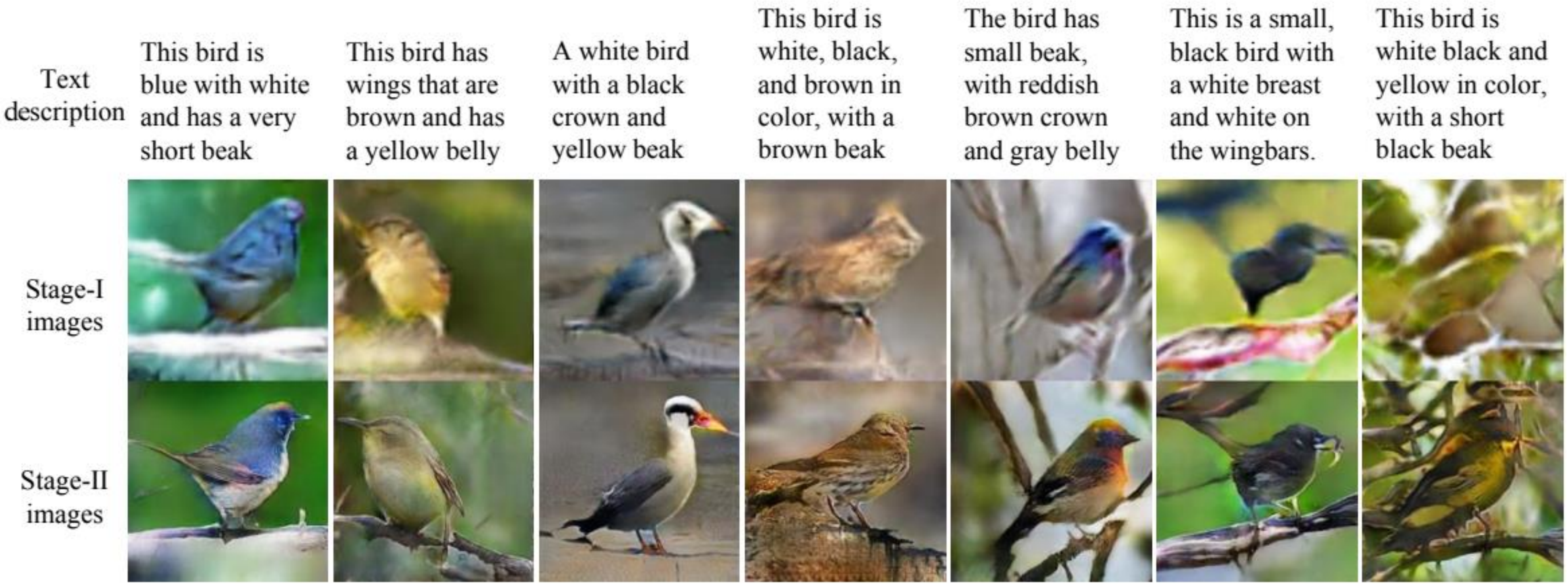
# 5 Result



| Metric | Dataset | GAN-INT-CLS | GAWWN | Our StackGAN |
|---|---|---|---|---|
| Inception score | CUB | 2.88 ± .04 | 3.62 ± .07 | **3.70 ± .04** |
| | Oxford | 2.66 ± .03 | / | **3.20 ± .01** |
| | COCO | 7.88 ± .07 | / | **8.45 ± .03** |
| Human rank | CUB | 2.81 ± .03 | 1.99 ± .04 | **1.37 ± .02** |
| | Oxford | 1.87 ± .03 | / | **1.13 ± .03** |
| | COCO | 1.89 ± .04 | / | **1.11 ± .03** |

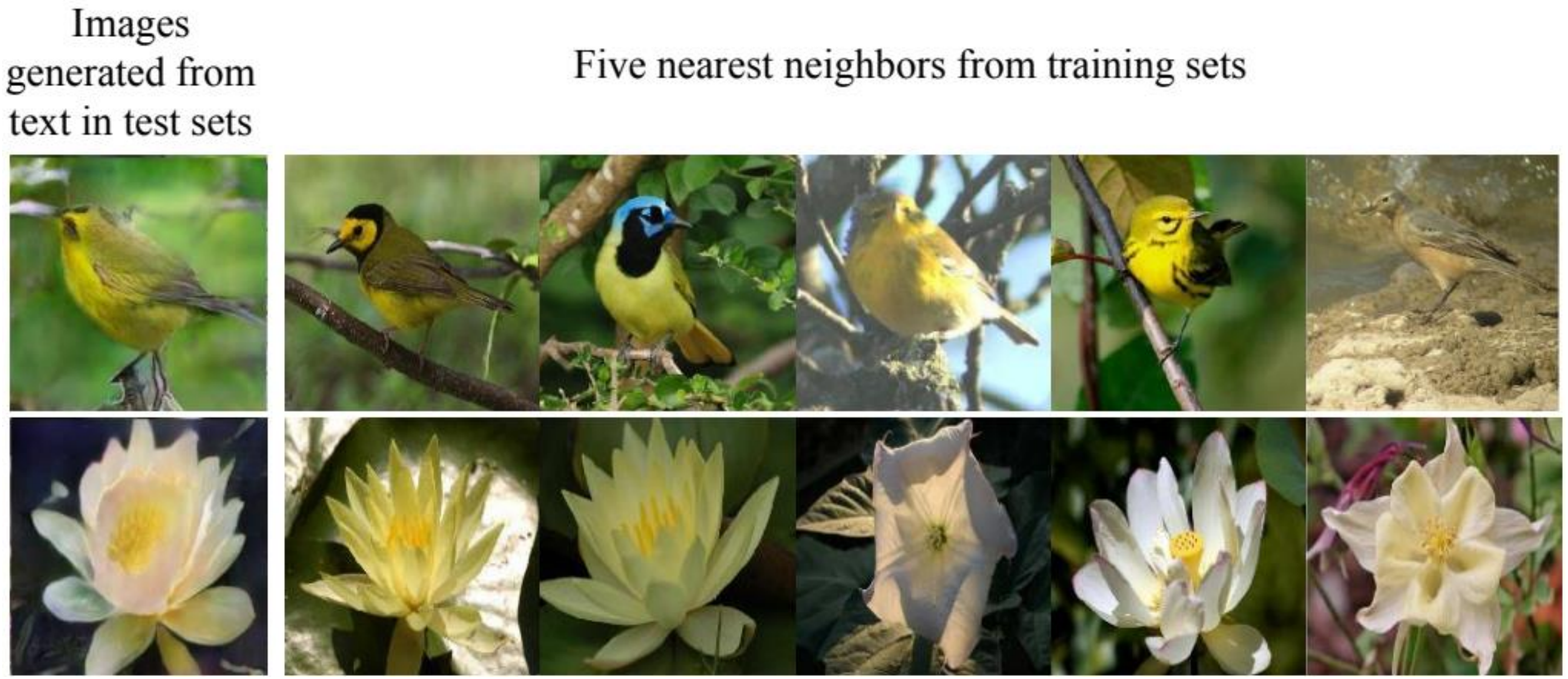
Inception scores and average human ranks of StackGAN, GAWWN, and GAN-INT-CLS on CUB, Oxford102, and MS-COCO datasets.

# 5 Result


Text description
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
Stage-I images
Stage-II images

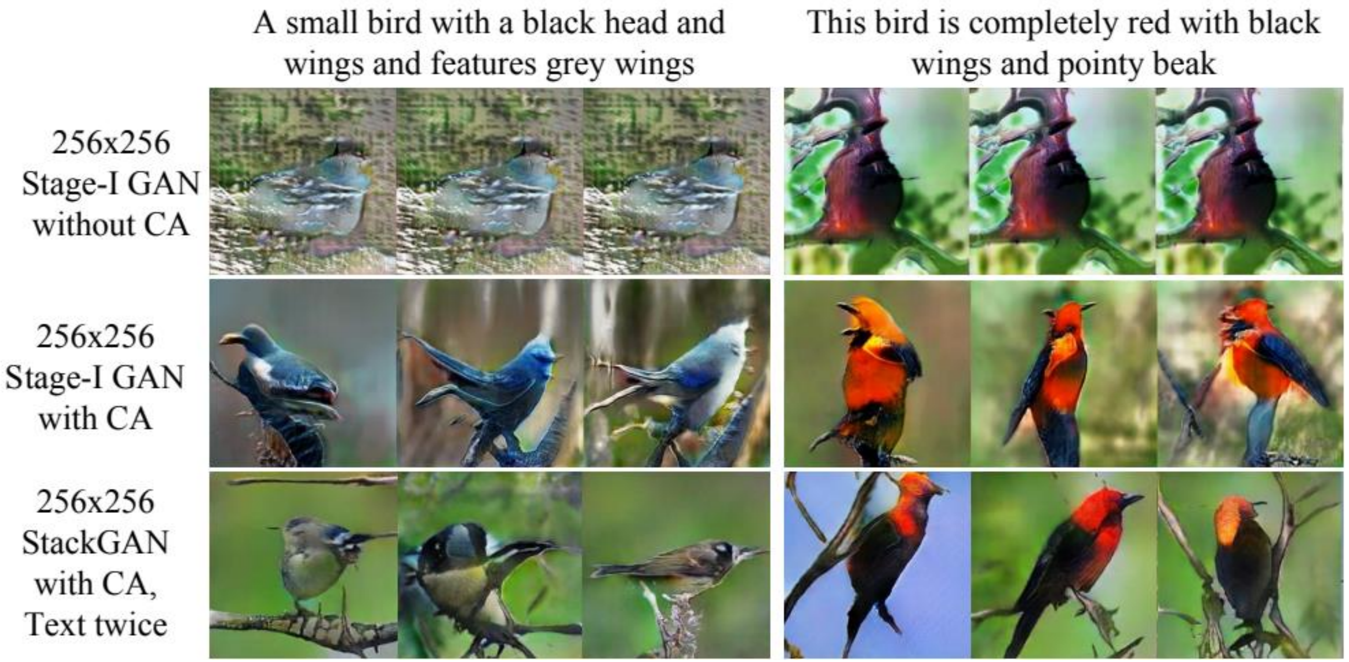
Samples generated by StackGAN from unseen texts in CUB test set.
Each column lists the text description, images generated from the text by Stage-I and Stage-II

# 5 Result


Images generated from text in test sets
Five nearest neighbors from training sets

For generated images, retrieving their nearest training images by utilizing Stage-II discriminator to extract visual features.

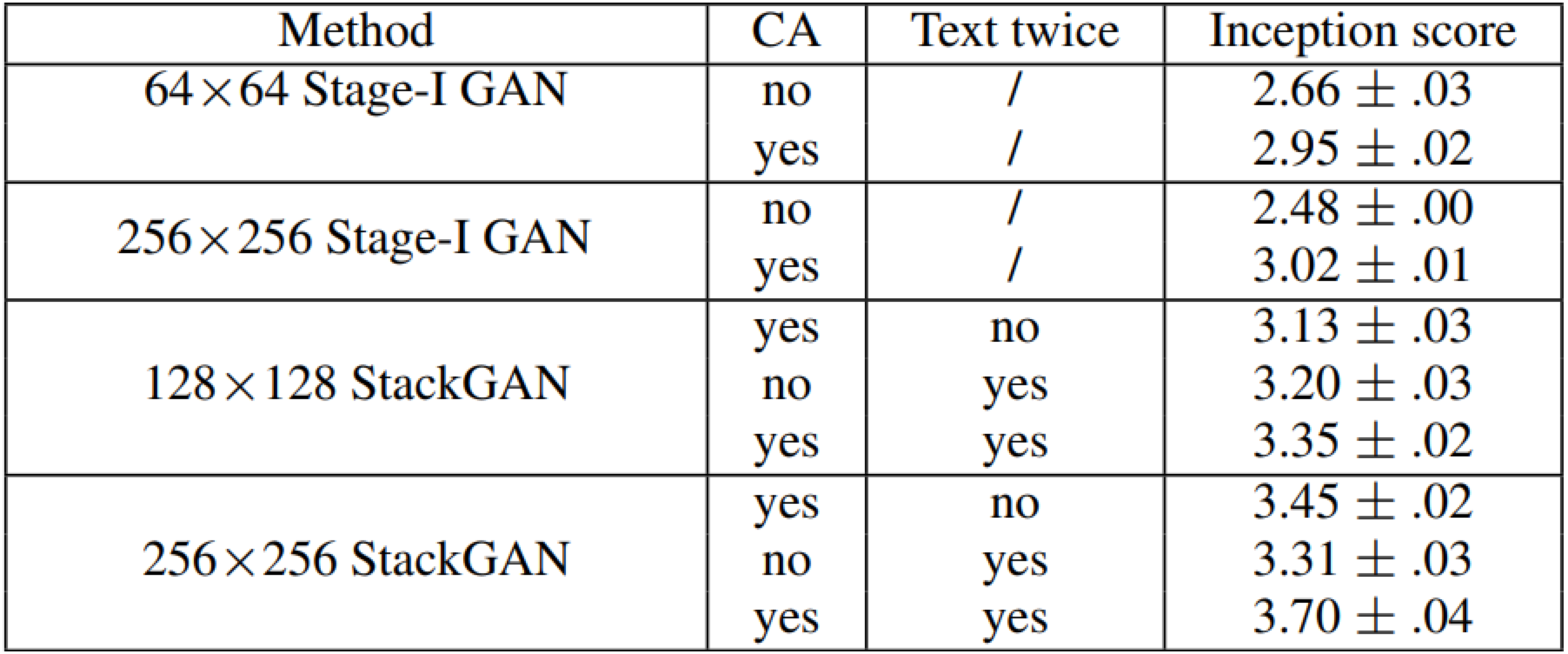
# 5 Result


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

Conditioning Augmentation (CA) helps stabilize the training of conditional GAN and improves the diversity of the generated samples. (Row 1) without CA, Stage-I GAN fails to generate plausible 256×256 samples.

Although different noise vector z is used for each column, the generated samples collapse to be the same for each input text description. (Row 2-3) with CA but fixing the noise vectors z, methods are still able to generate birds with different poses and viewpoints.

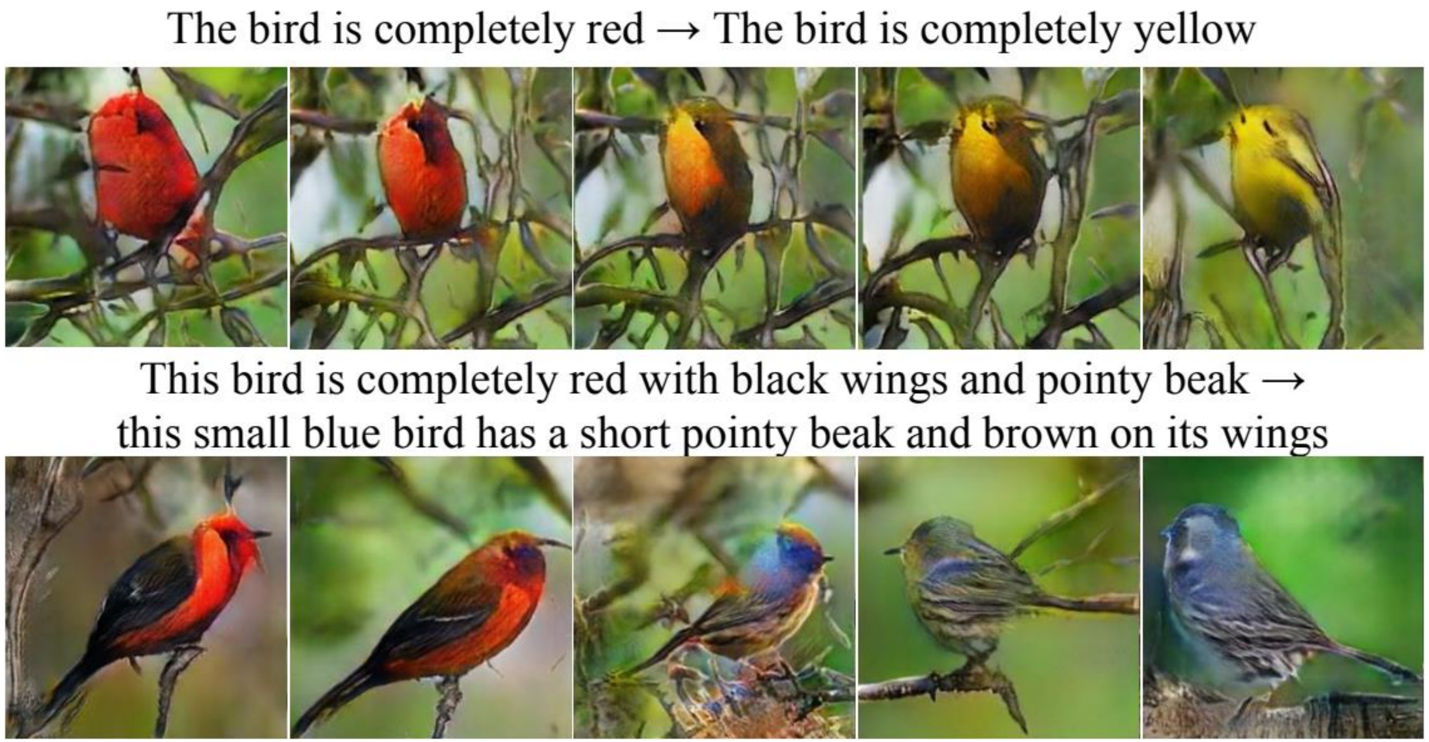
# 5 Result



| Method | CA | Text twice | Inception score |
|---|---|---|---|
| 64×64 Stage-I GAN | no | / | 2.66 ± .03 |
| | yes | / | 2.95 ± .02 |
| 256×256 Stage-I GAN | no | / | 2.48 ± .00 |
| | yes | / | 3.02 ± .01 |
| 128×128 StackGAN | yes | no | 3.13 ± .03 |
| | no | yes | 3.20 ± .03 |
| | yes | yes | 3.35 ± .02 |
| 256×256 StackGAN | yes | no | 3.45 ± .02 |
| | no | yes | 3.31 ± .03 |
| | yes | yes | 3.70 ± .04 |

Inception scores calculated with 30,000 samples generated by different baseline models of StackGAN

# 5 Result


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

(Left to right) Images generated by interpolating two sentence embeddings.
Gradual appearance changes from the first sentence’s meaning to that of the second sentence can be observed. The noise vector z is fixed to be zeros for each row.

6

# Reference

- http://proceedings.mlr.press/v48/reed16.pdf
- https://arxiv.org/pdf/1609.04802.pdf
- https://arxiv.org/pdf/1612.03242.pdf
- https://www.slideshare.net/WoojinJeong5/review-srgan
- https://hichoe95.tistory.com/47
- https://leedakyeong.tistory.com/entry/%EB%85%BC%EB%AC%B8Photo-Realistic-Single-Image-Super-Resolution-Using-a-Generative-Adversarial-NetworkSRGAN