MirrorGAN: Learning Text-to-image Generation by Redescription

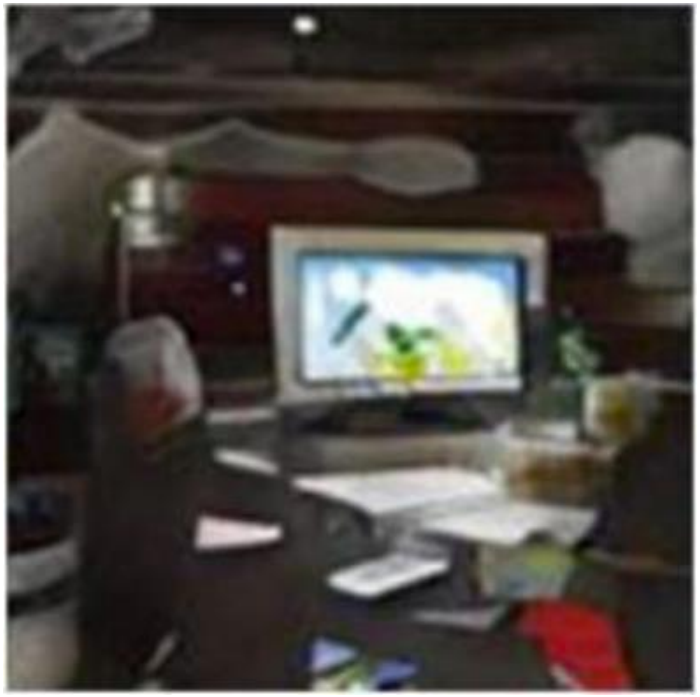
Sanghyuck Na
Jan, 15, 2021
Dongguk University
Artificial Intelligence Laboratory
shna@Dongguk.edu

0

Contents

1 Introduction

Text to image generation

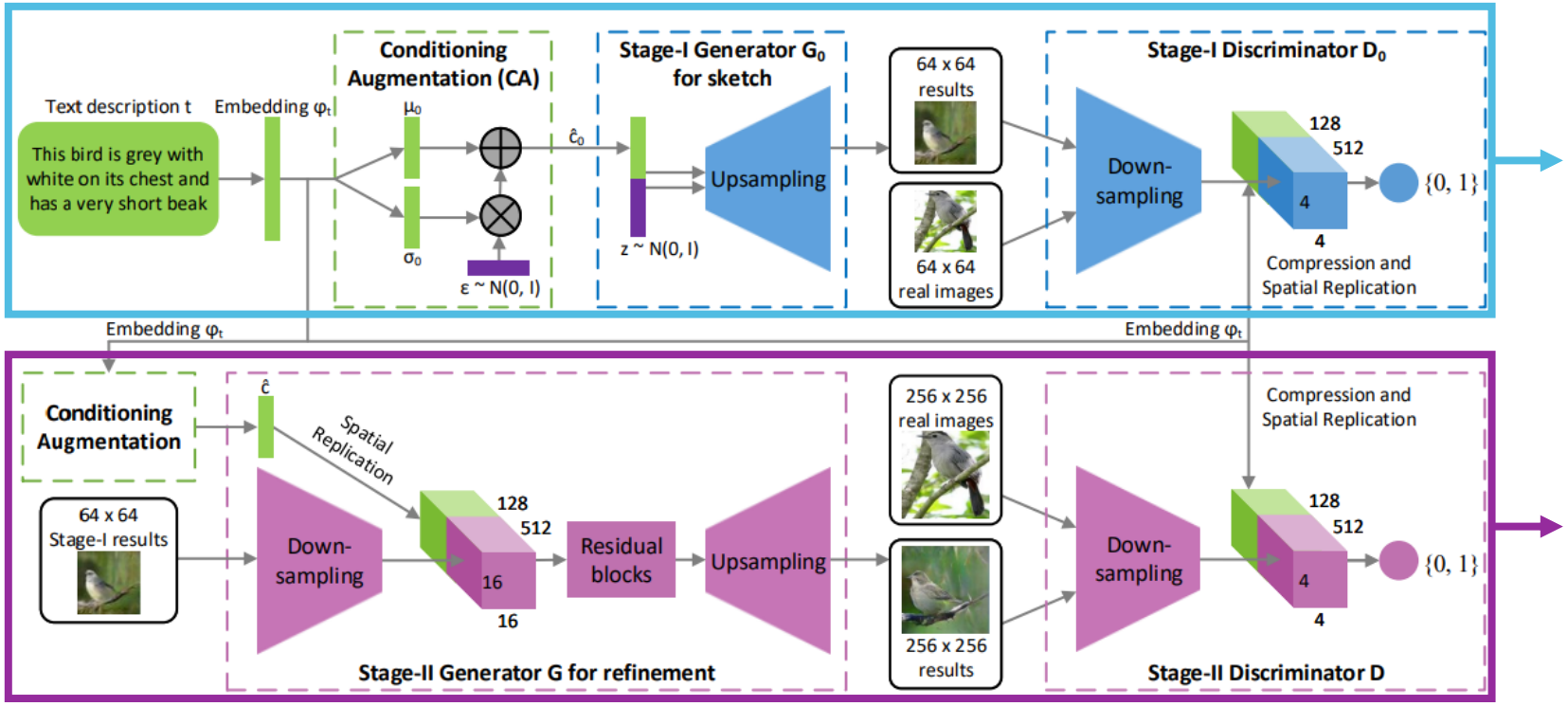


There is a lot of electrical sitting on the table.

Image to text generation

Green: Human ground truth.
Red: Top-scoring sentence from training set.
Blue: Generated sentence.

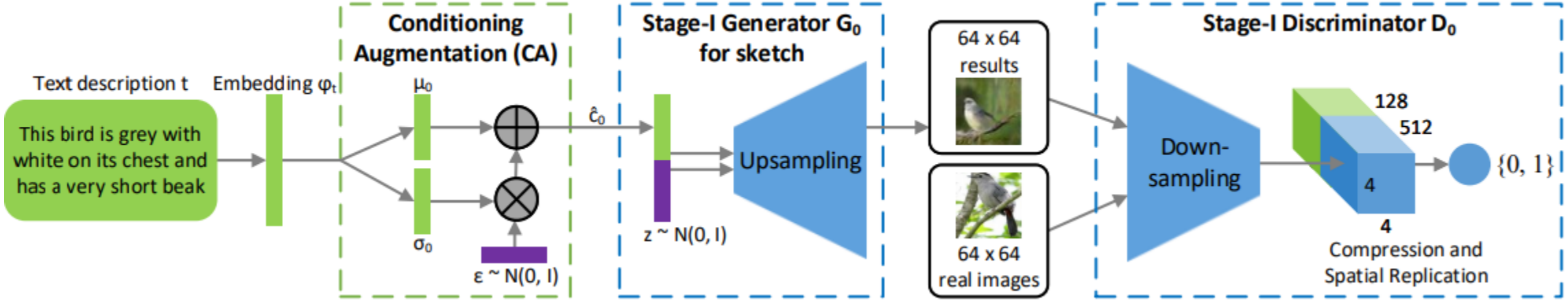
2 StackGAN



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.

2 StackGAN

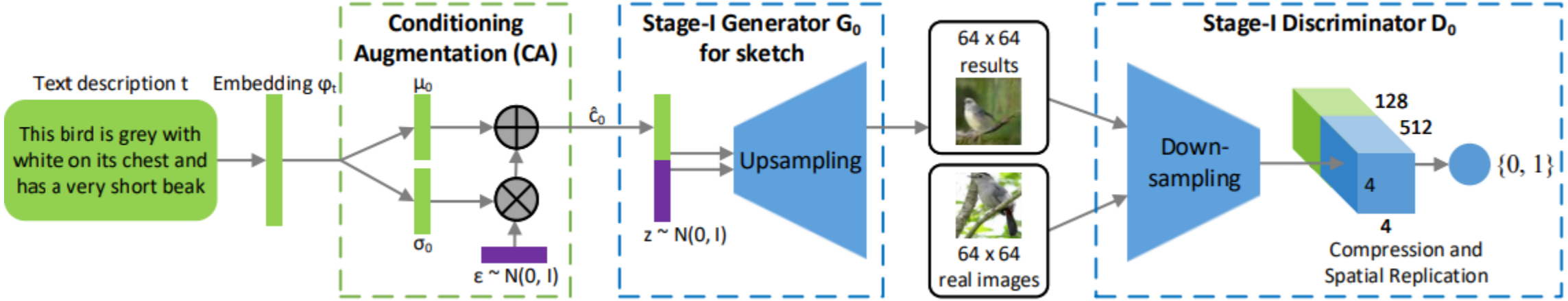


$$Conditioning\ Augmentation\ (CA)$$

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

t : *text description*
z : *noise vector from Gaussian Distribution*
φ_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

2 StackGAN

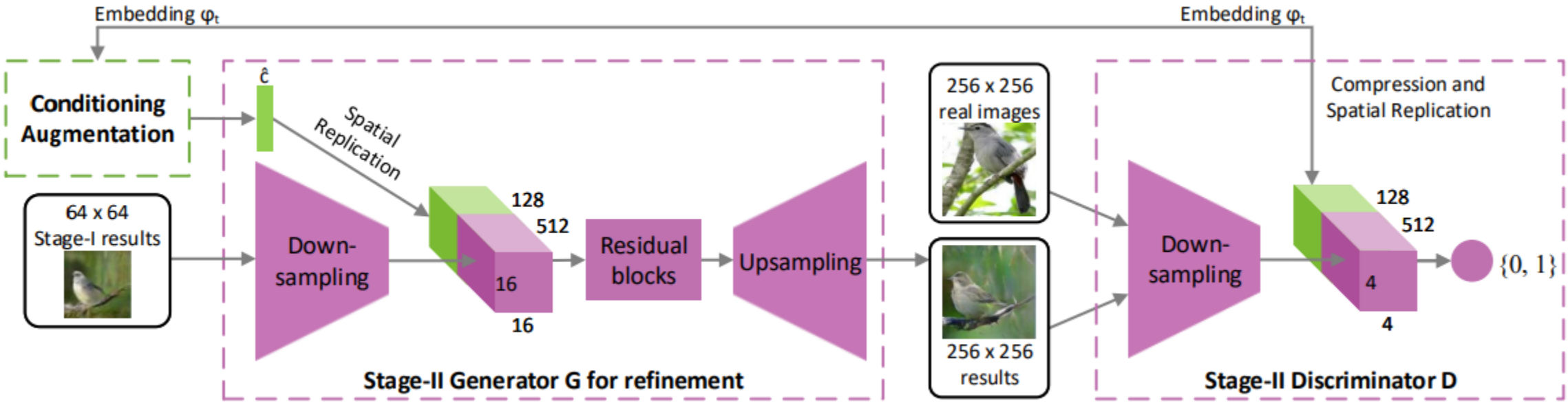


$$L_{D_0} = \mathbb{E}_{(I_0,t)\sim p_{data}}[logD_0(I_0, \varphi_t)] + \mathbb{E}_{z\sim p_{z,},t\sim p_{data,}}[\log(1 - D_0(G_0(z, \hat{c}_0, \varphi_t))]$$

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

t : *text description*
z : *noise vector from Gaussian Distribution*
φ_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

2 StackGAN

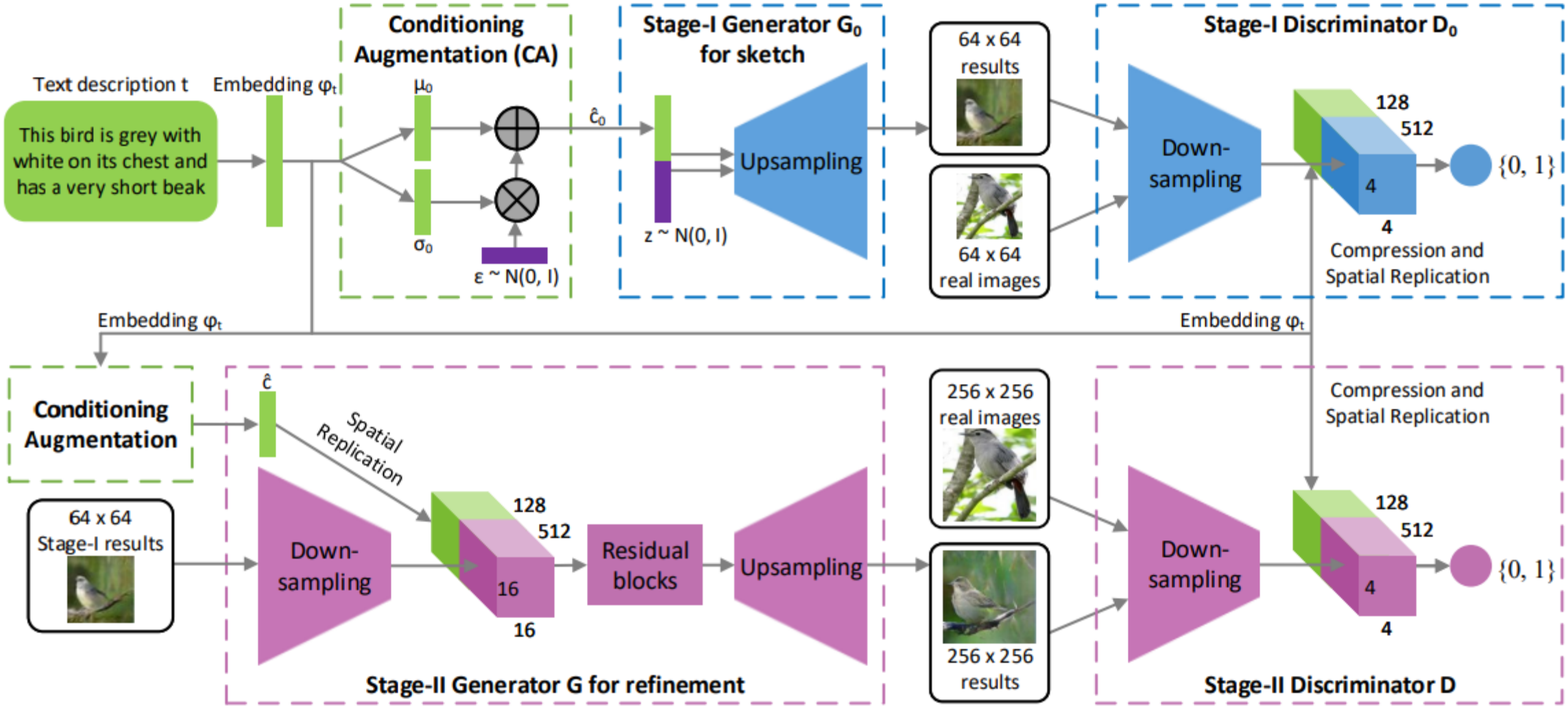


$$L_D = \mathbb{E}_{(I,t)\sim p_{data}}[logD(I,\varphi_t)] + \mathbb{E}_{s_0\sim p_{G_0},t\sim p_{data,}}[\log(1 - D(G\ (s_0,\hat{c}_0),\varphi_t))]$$

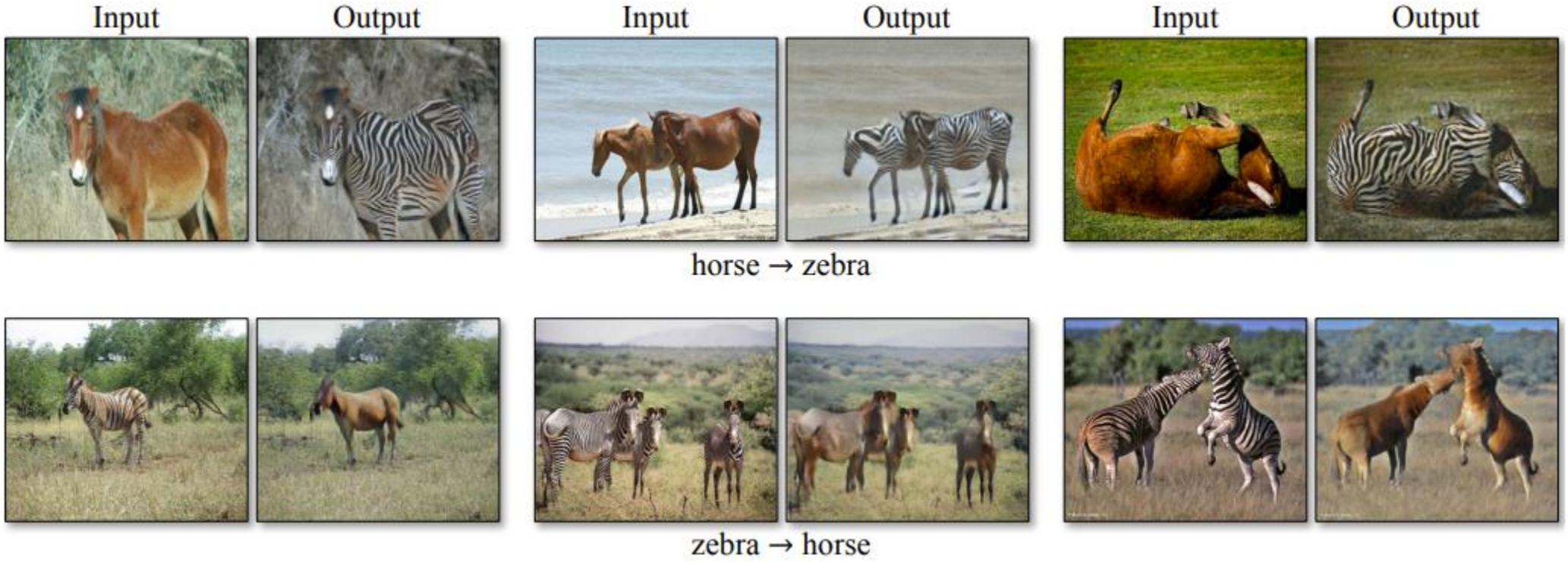
$$L_G = \mathbb{E}_{s_0\sim p_{G_0},t\sim p_{data,}}[\log(1 - D(G\ (s_0,\hat{c}),\varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t),\Sigma(\varphi_t))\|\mathcal{N}(0,I))$$

t : *text description*
z : *noise vector from Gaussian Distribution*
φ_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

2 StackGAN

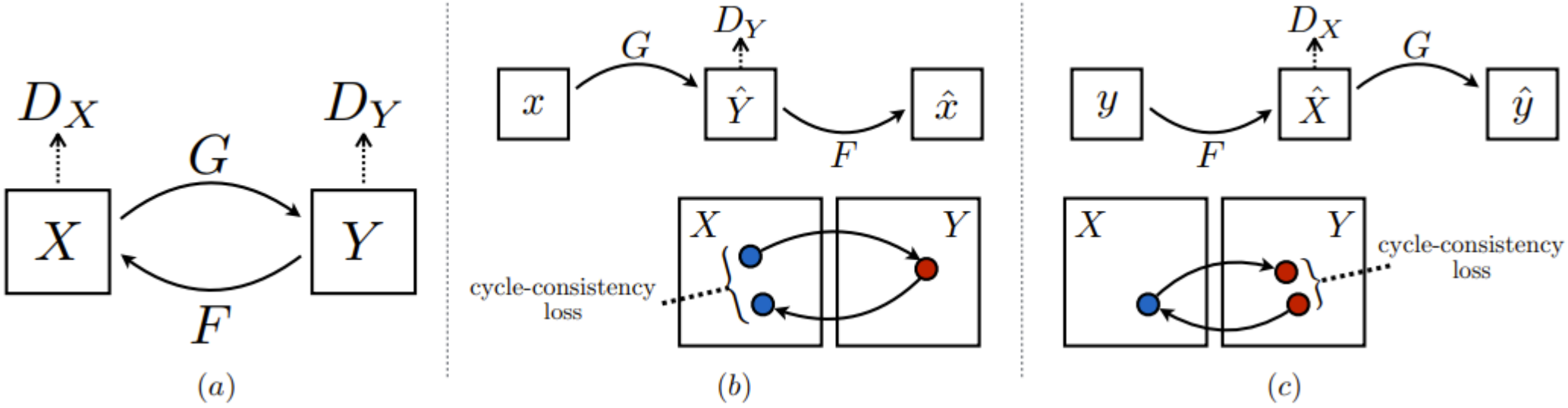

Text description t
This bird is grey with white on its chest and has a very short beak
Embedding φ_t
Conditioning Augmentation (CA)
μ_0
σ_0
$\hat{c}_0$
$\varepsilon \sim N(0, I)$
Stage-I Generator G_0 for sketch
$z \sim N(0, I)$
Upsampling
64 x 64 results
64 x 64 real images
Stage-I Discriminator D_0
Down-sampling
128
512
4
4
{0, 1}
Compression and Spatial Replication
Embedding φ_t
Embedding φ_t
Conditioning Augmentation
$\hat{c}$
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

3 CycleGAN



Given any two unordered image collections, CycleGAN translates image to image.

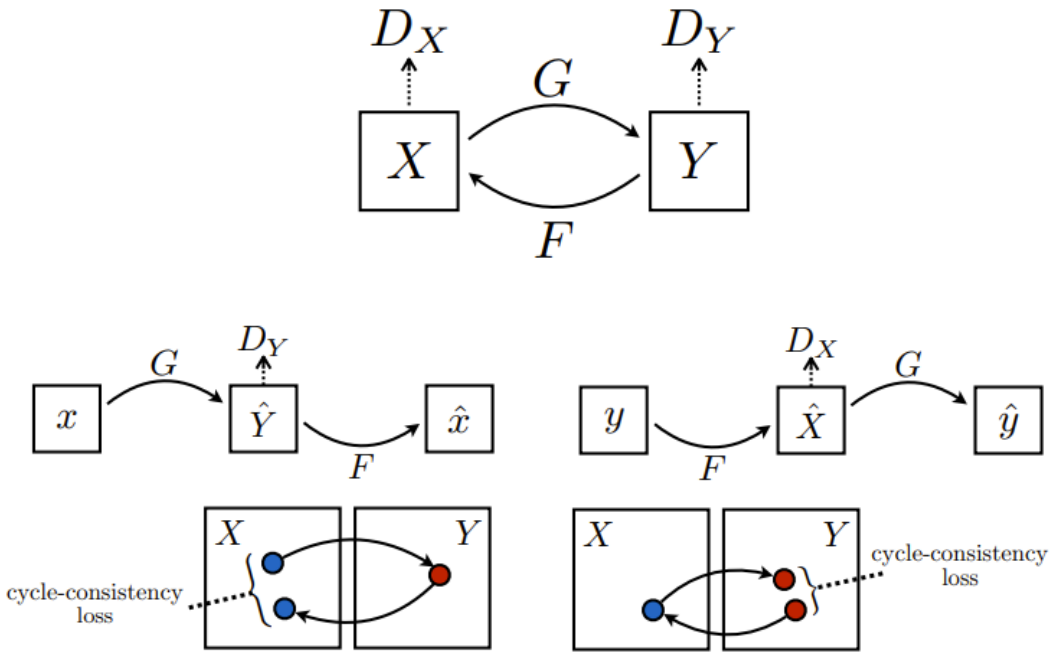
3 CycleGAN



$$G^*, F^* = arg \min_{G,F} \max_{D_X D_Y} L(G, F, D_X, D_Y)$$

CycleGAN contains two mapping functions and Associated adversarial discriminator D_X and D_Y.

3 CycleGAN



Adversarial loss

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y\sim p_{data}(y)}[logD_Y(y) + \mathbb{E}_{x\sim p_{data}(x)}[logD_Y((G(x))$$

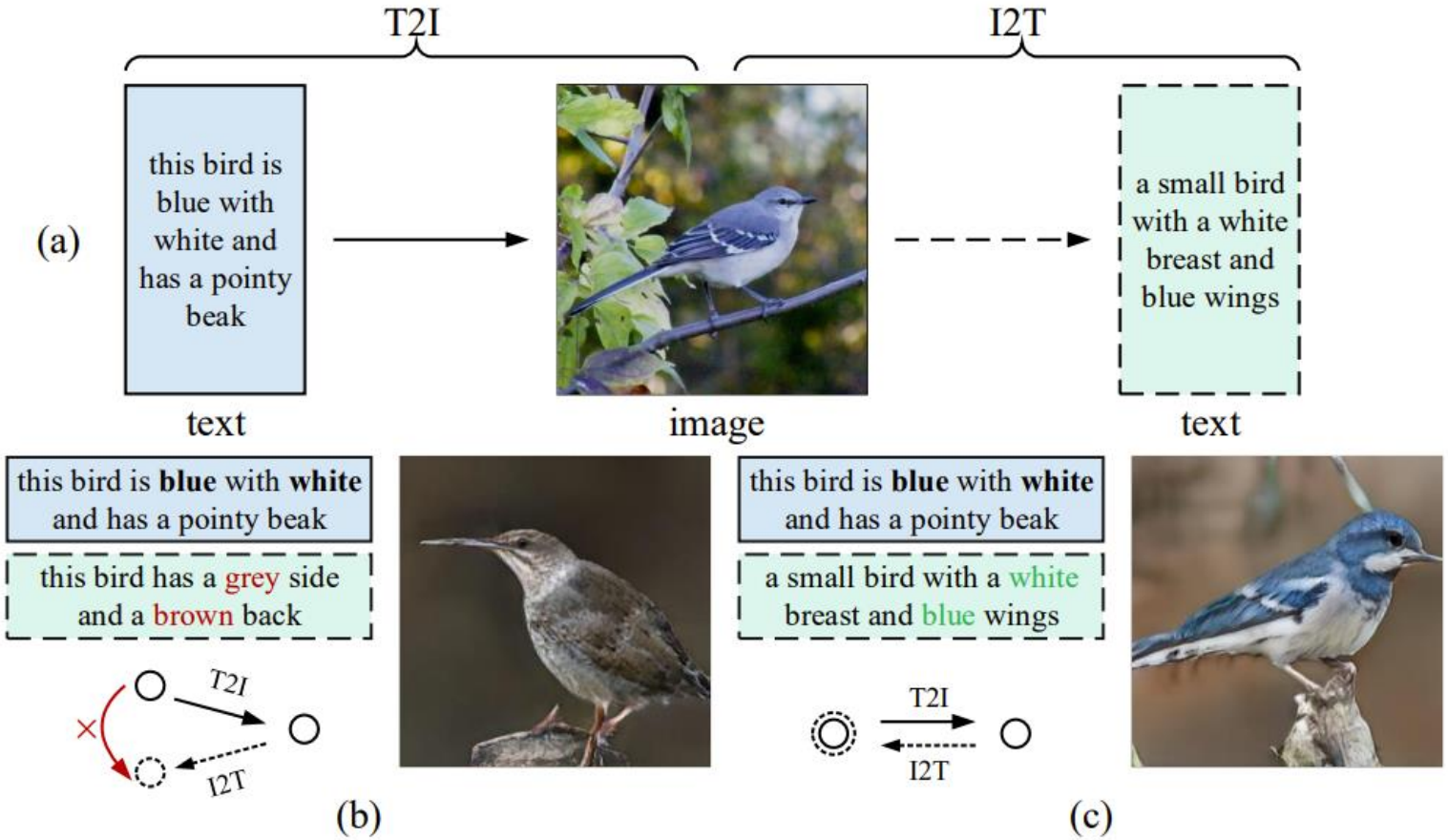
Cycle consistency loss

$$L_{cyc}(G, F) = \mathbb{E}_{x\sim p_{data}(x)}\left[\left\|F(G(x)) - x\right\|_1\right] + \mathbb{E}_{y\sim p_{data}(y)}[\left\|G(F(y)) - y\right\|_1]$$

Full objective

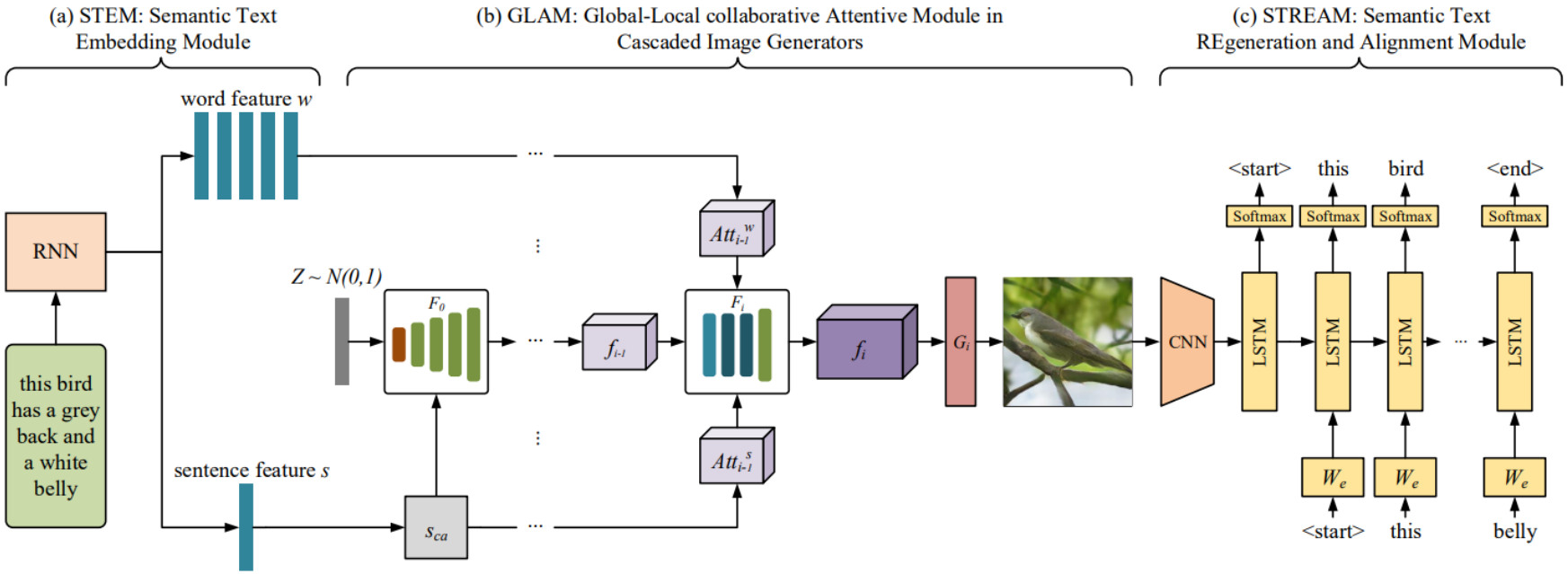
$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F)$$

MirrorGAN



MirrorGAN has two goals : T2I and I2T.
First, T2I : An image generates from a text description.
Second, I2T : A re-description is generated from the image created by T2I.

MirrorGAN

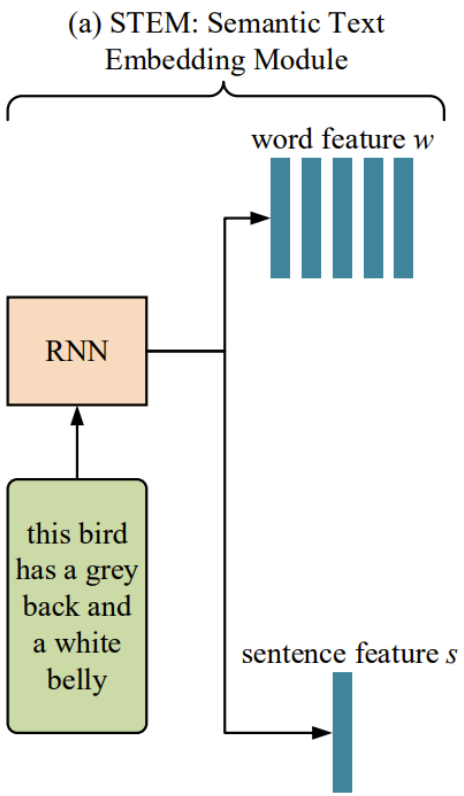


MirrorGAN has three modules.

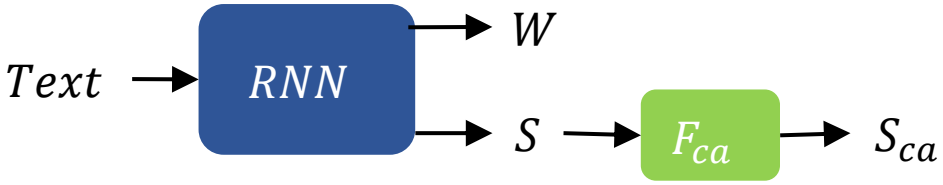
First, STEM : Semantic Text Embedding Module.
Second, GLAM : Global-Local collaborative Attentive Module in Cascaded Image Generators.
Third, STREAM : Semantic Text REgeneration and Alignment Module.

MirrorGAN

STEM : Semantic Text Embedding Module



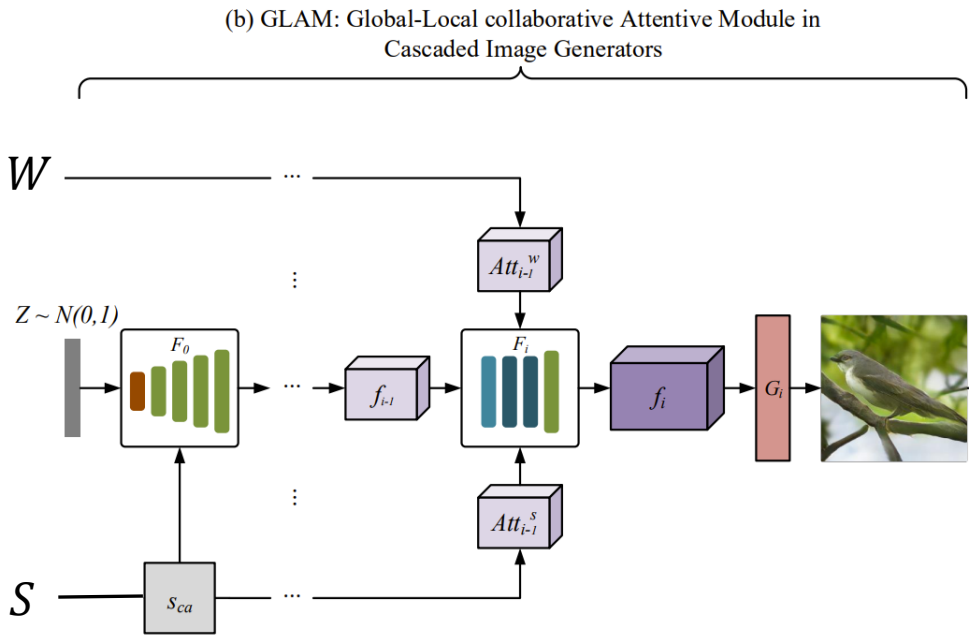
- RNN is used to extract word and sentence features from the given text.
- Sentence features apply conditioning augmentation.



$$w, s = RNN(T) \qquad S_{ca} = F_{ca}(s)$$

MirrorGAN

GLAM : Global-Local collaborative Attentive Module in Cascaded Image Generators



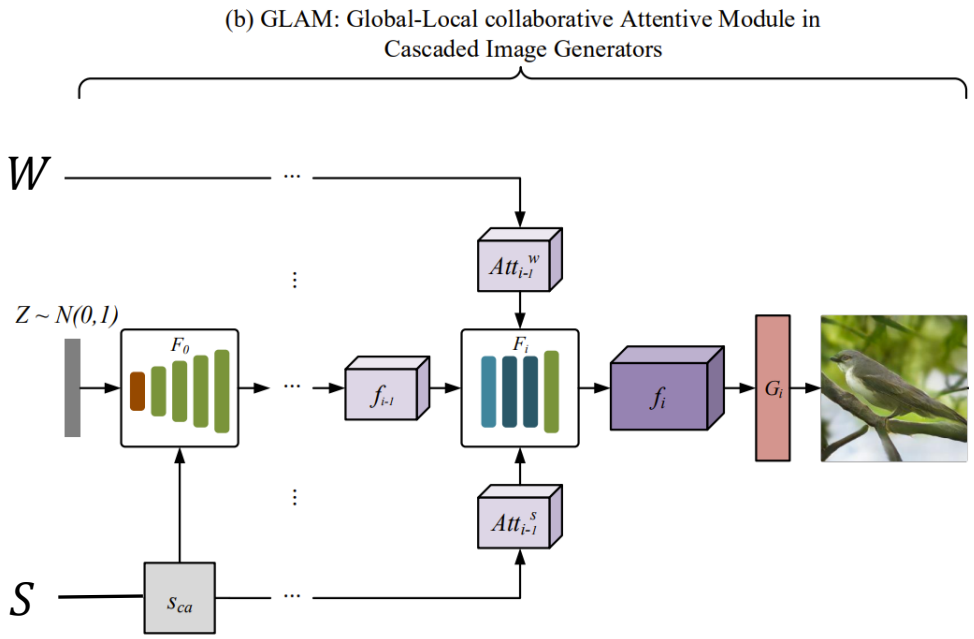
- Image generate through attention of $word \leftrightarrow local\ image$ and $sentence \leftrightarrow global\ image$
- There is construct a multi-stage cascaded generator by stacking three image generation networks sequentially.

$$Visual\ feature\ transfomers = \{F_0, F_1, \dots F_{m-1}\}$$

$$Image\ generator = \{G_0, G_1, \dots G_{m-1}\}$$

MirrorGAN

GLAM : Global-Local collaborative Attentive Module in Cascaded Image Generators



Att_{i-1}^w *is context attn score between words and images.*
Att_{i-1}^s *is context attn score between sentences and images.*
V_{i-1} *is the term for* s_{ca} *mapping to a common semantic space.*

$$f_0 = F_0(z, s_{ca})$$

$$f_i = F_i\left(f_{i-1}, F_{att_i}(f_{i-1}, w, s_{ca})\right), i \in \{1, 2, \dots, m-1\}$$

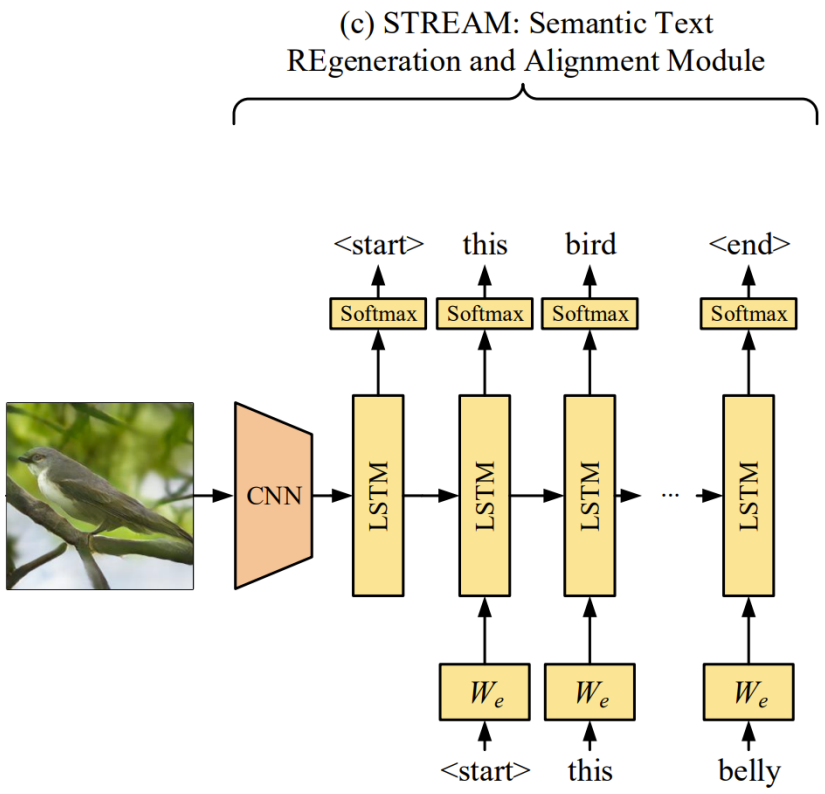
$$I_i = G_i(f_i), i \in \{1, 2, \dots, m-1\}$$

$$Att_{i-1}^w = \sum_{l=0}^{L-1} (U_{i-1}w^l)(softmax\left(f_{i-1}^T(U_{i-1}w^l)\right))^T$$

$$Att_{i-1}^s = (V_{i-1}s_{ca}) \circ (softmax(f_{i-1} \circ (V_{i-1}s_{ca})))$$

$$F_{att_i}(f_{i-1}, w, s_{ca}) = concat(Att_{i-1}^w, Att_{i-1}^s)$$

MirrorGAN



STREAM : Semantic Text Regeneration and Alignment Module

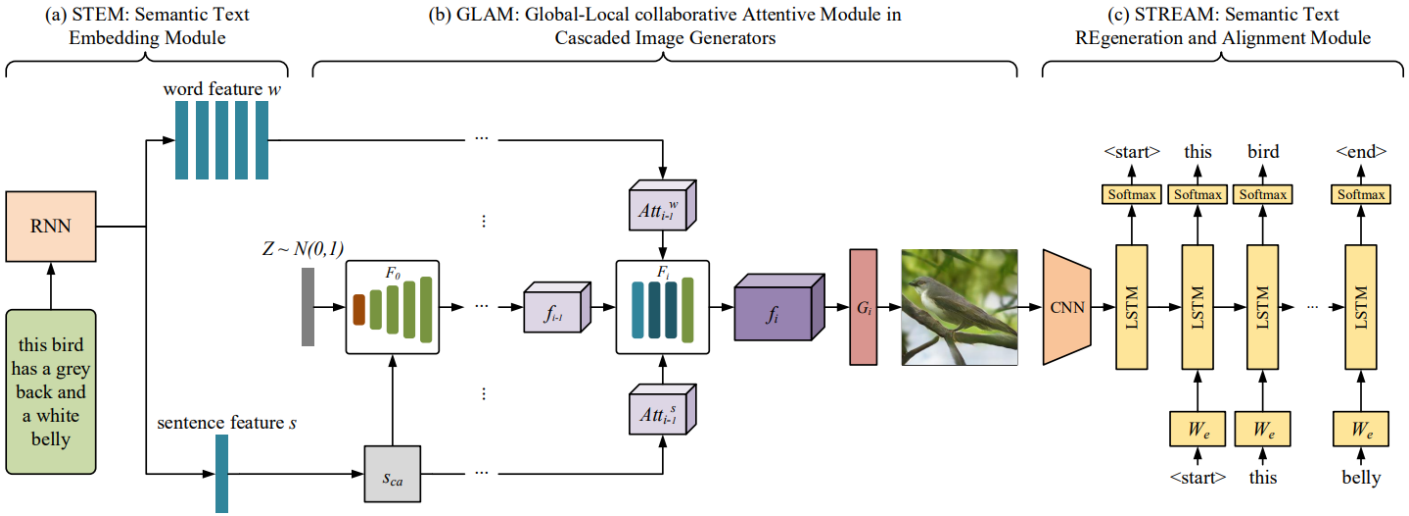
- Regenerate text description from generation image.
- CNN was pretrained on imageNet.

$$x_{-1} = CNN(I_{m-1})$$

$$x_t = W_e T_t, t \in \{0, \dots L-1\}$$

$$p_{t+1} = RNN(x_t), t \in \{0, \dots L-1\}$$

MirrorGAN

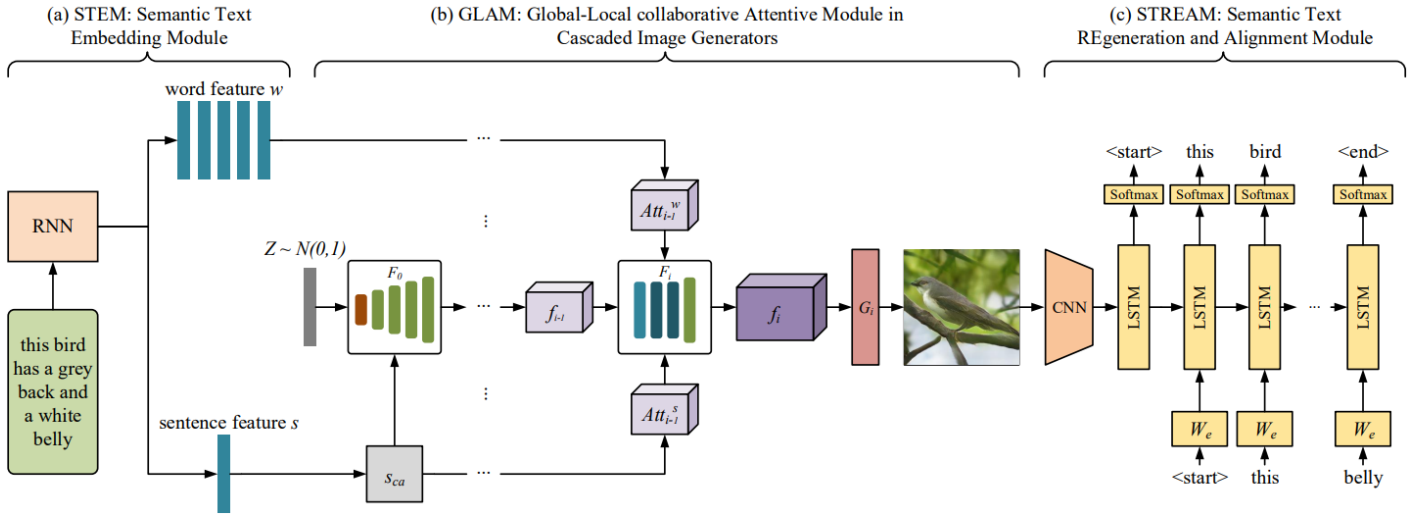


$$L_{G_i} = -\frac{1}{2}\mathbb{E}_{I_i \sim P_{I_i}}\left[\log\left(D_i(I_i)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i \sim P_{I_i}}\left[\log\left(D_i(I_i, s)\right)\right]$$

$$L_{stream} = -\sum_{t=0}^{L-1} logp_t(T_t)$$

$$L_G = \sum_{i=0}^{m-1} L_{G_i} + \lambda L_{stream}$$

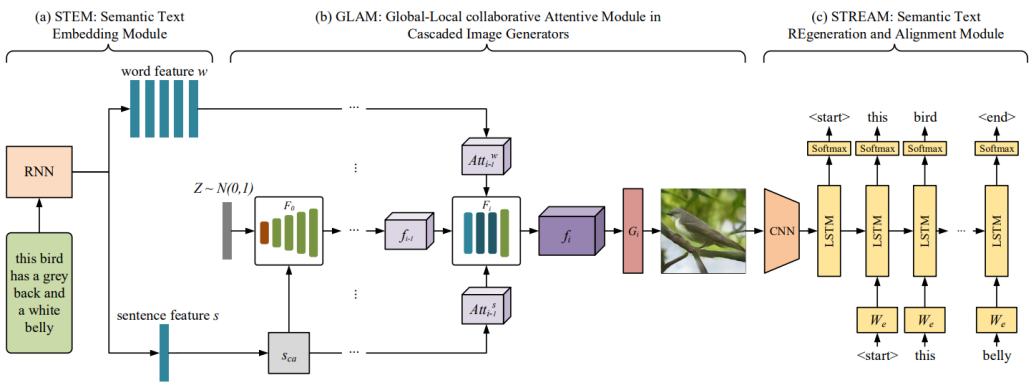
MirrorGAN



$$L_{D_i} = -\frac{1}{2}\mathbb{E}_{I_i^{GT}\sim P_{I_i^{GT}}}\left[\log\left(D_i\left(I_i^{GT}\right)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i\sim P_{I_i}}\left[\log\left(1 - D_i(I_i)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i^{GT}\sim P_{I_i^{GT}}}\left[\log\left(D_i\left(I_i^{GT}, s\right)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i\sim P_{I_i}}\left[\log\left(1 - D_i(I_i, s)\right)\right]$$

$$L_D = \sum_{i=0}^{m-1} L_{D_i}$$

MirrorGAN



$$L_{D_i} = -\frac{1}{2}\mathbb{E}_{I_i^{GT}\sim P_{I_i^{GT}}}\left[\log\left(D_i\left(I_i^{GT}\right)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i\sim P_{I_i}}\left[\log\left(1 - D_i(I_i)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i^{GT}\sim P_{I_i^{GT}}}\left[\log\left(D_i\left(I_i^{GT}, s\right)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i\sim P_{I_i}}\left[\log\left(1 - D_i(I_i, s)\right)\right]$$

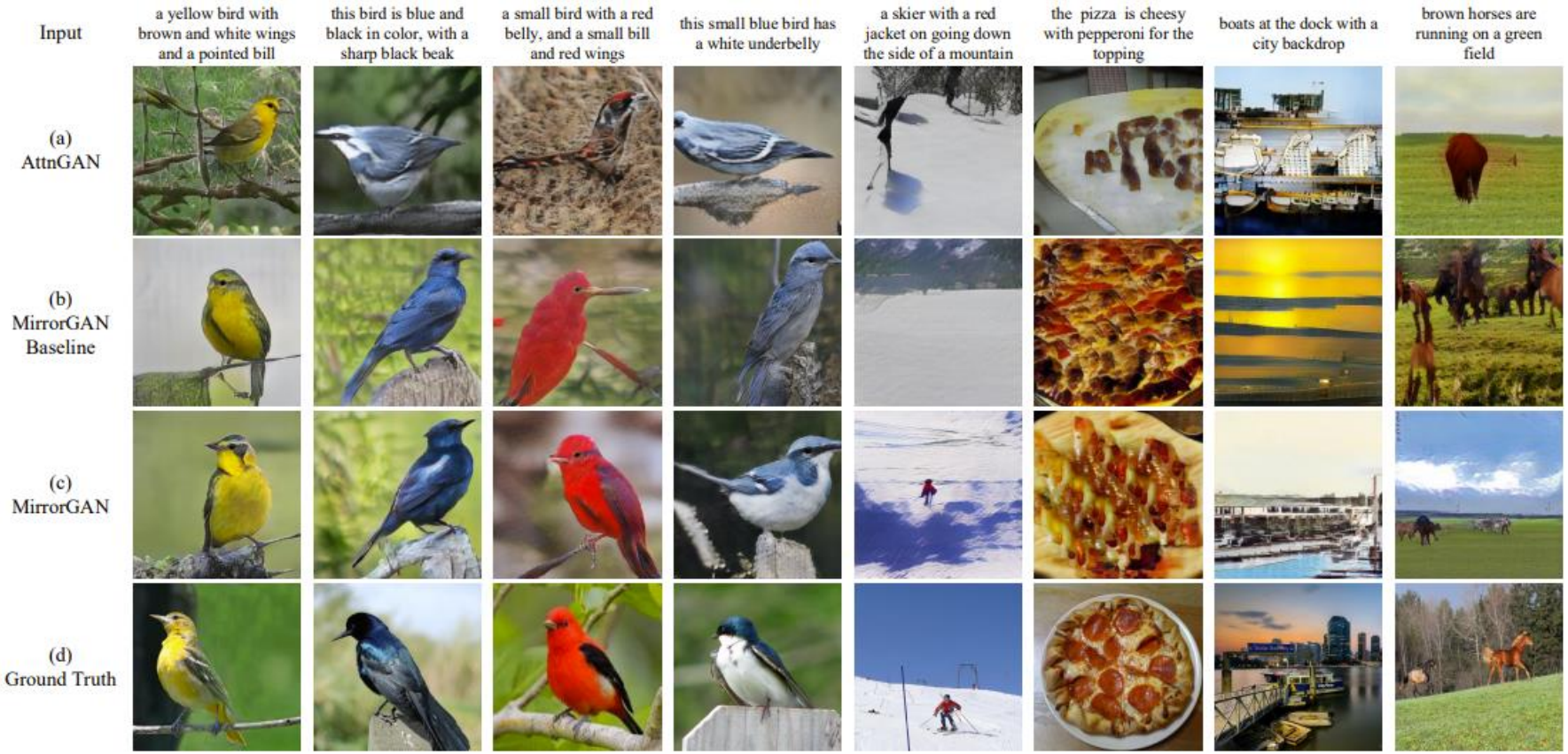
$$L_D = \sum_{i=0}^{m-1} L_{D_i}$$

$$L_{G_i} = -\frac{1}{2}\mathbb{E}_{I_i\sim P_{I_i}}\left[\log\left(D_i(I_i)\right)\right] - \frac{1}{2}\mathbb{E}_{I_i\sim P_{I_i}}\left[\log\left(D_i(I_i, s)\right)\right]$$

$$L_{stream} = -\sum_{t=0}^{L-1} \log p_t(T_t)$$

$$L_G = \sum_{i=0}^{m-1} L_{G_i} + \lambda L_{stream}$$

5 Result


Input
a yellow bird with brown and white wings and a pointed bill
this bird is blue and black in color, with a sharp black beak
a small bird with a red belly, and a small bill and red wings
this small blue bird has a white underbelly
a skier with a red jacket on going down the side of a mountain
the pizza is cheesy with pepperoni for the topping
boats at the dock with a city backdrop
brown horses are running on a green field
(a) AttnGAN
(b) MirrorGAN Baseline
(c) MirrorGAN
(d) Ground Truth

5 Result

Statistics of datasets

Dataset	CUB [28]		COCO [14]	
	train	test	train	test
#samples	8,855	2,933	80k	40k
caption/image	10	10	5	5

Inception Score

Model	CUB	COCO
GAN-INT-CLS [24]	2.88 ± 0.04	7.88 ± 0.07
GAWWN [25]	3.62 ± 0.07	-
StackGAN [38]	3.70 ± 0.04	8.45 ± 0.03
StackGAN++ [39]	3.82 ± 0.06	-
PPGN [20]	-	9.58 ± 0.21
AttnGAN [34]	4.36 ± 0.03	25.89 ± 0.47
MirrorGAN	**4.56 ± 0.05**	**26.47 ± 0.41**

Inception Score evaluate quality and diversity from generated image.

R-precision

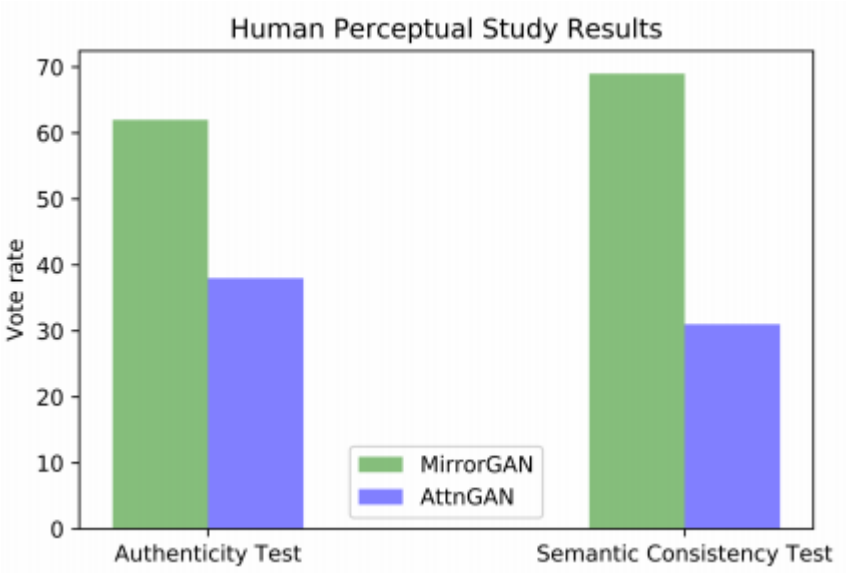
Dataset	CUB			COCO		
top-k	k=1	k=2	k=3	k=1	k=2	k=3
AttnGAN [34]	53.31	54.11	54.36	72.13	73.21	76.53
MirrorGAN	**57.67**	**58.52**	**60.42**	**74.52**	**76.87**	**80.21**

There is cosine similarities between the generated image vectors and text vectors.

Evaluation Metric	Inception Score		R-precision (top-1)	
	CUB	COCO	CUB	COCO
MirrorGAN w/o GA, λ=0	$3.91 \pm .09$	$19.01 \pm .42$	39.09	50.69
MirrorGAN w/o GA, λ=20	$4.47 \pm .07$	$25.99 \pm .41$	55.67	73.28
MirrorGAN, λ=5	$4.01 \pm .06$	$21.85 \pm .43$	32.07	52.55
MirrorGAN, λ=10	$4.30 \pm .07$	$24.11 \pm .31$	43.21	63.40
MirrorGAN, λ=20	**$4.54 \pm .17$**	**$26.47 \pm .41$**	**57.67**	**74.52**

GA = Global Attention
λ expresses the importance of *STREAM*

5 Result

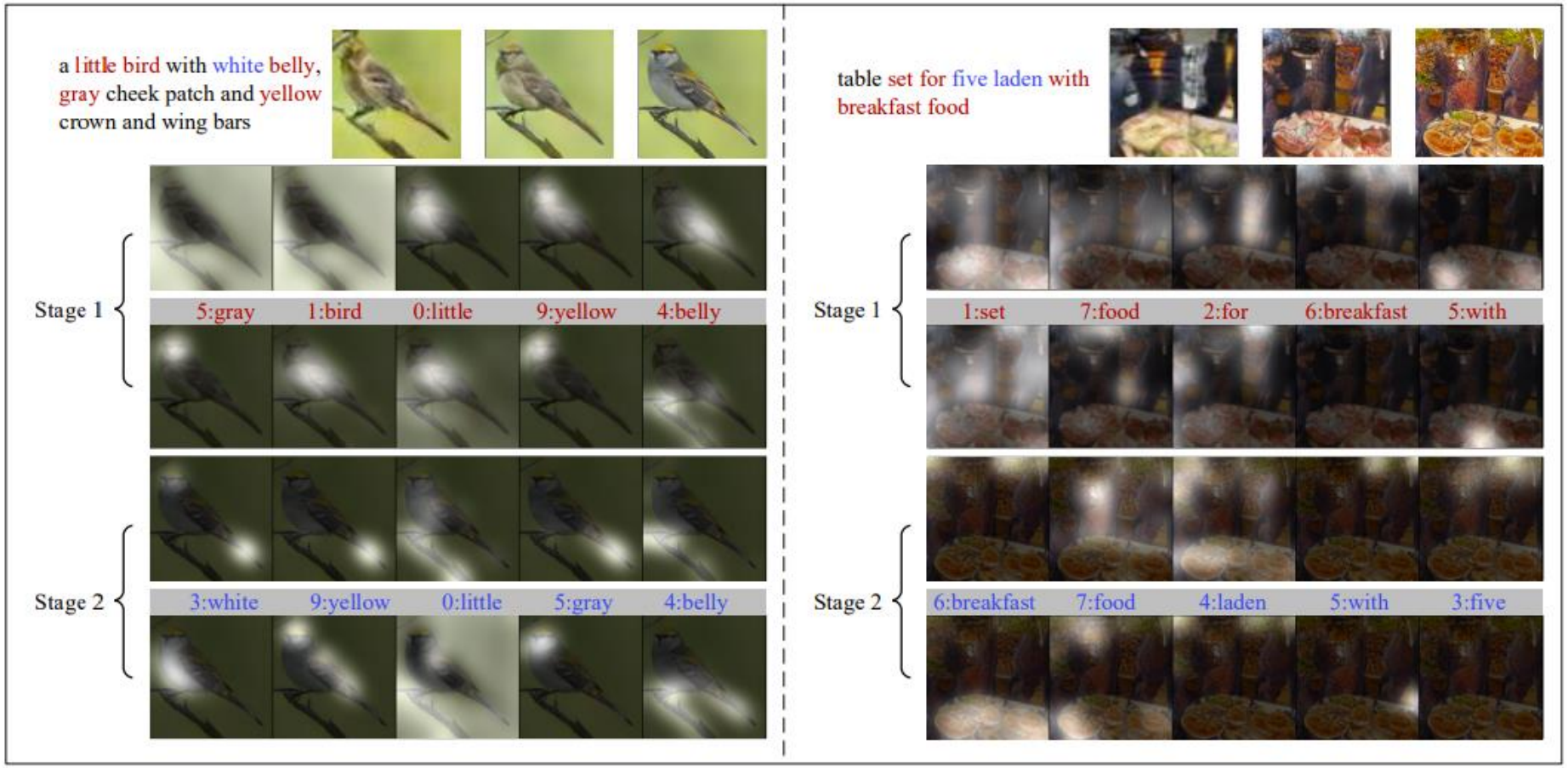


Results of Human perceptual test

A higher value of the Authenticity Test means more convincing images.

A higher value of the Semantic Consistency Test means a closer semantics between input text and generated images.

5 Result


a little bird with white belly, gray cheek patch and yellow crown and wing bars
Stage 1
5:gray
1:bird
0:little
9:yellow
4:belly
Stage 2
3:white
9:yellow
0:little
5:gray
4:belly
table set for five laden with breakfast food
Stage 1
1:set
7:food
2:for
6:breakfast
5:with
Stage 2
6:breakfast
7:food
4:laden
5:with
3:five

5 Result

Images generated by MirrorGAN by modifying the text descriptions by a single word and the corresponding top-2 attention maps in the last stage

Reference

- https://arxiv.org/pdf/1903.05854v1.pdf
- https://arxiv.org/pdf/1612.03242.pdf
- https://arxiv.org/pdf/1711.10485.pdf
- https://arxiv.org/pdf/1703.10593.pdf
- https://cs.stanford.edu/people/karpathy/cvpr2015.pdf
- https://drive.google.com/file/d/1luXzEwW1NYX0s2SOuTLMetnIv0l5RR7M/view