Text to Image Synthesis with StackGAN

Vision & Perception

SAPIENZA
UNIVERSITÀ DI ROMA

Davide Ceriola, Fabio Fraschetti

Department of Computer, Control, and Management Engineering

18/10/2023

What is the project about?

Goal:

What is the project about?

Goal:

- Create a Network in order to generate 128x128 images starting from a text description.

What is the project about?

Goal:

- Create a Network in order to generate 128x128 images starting from a text description.

Proposal:

What is the project about?

Goal:

- Create a Network in order to generate 128x128 images starting from a text description.

Proposal:

- Implementing a stacked GAN to decompose the hard problem into more manageable sub-problems.

what is a GAN (Generative Adversarial Network)?

The GANs architecture consist of two neural networks:

what is a GAN (Generative Adversarial Network)?

The GANs architecture consist of two neural networks:

- **Generator**: Creates new data samples by learning from random noise.

what is a GAN (Generative Adversarial Network)?

The GANs architecture consist of two neural networks:

- **Generator**: Creates new data samples by learning from random noise.
- **Discriminator**: Attempts to distinguish between real and generated data.

what is a GAN (Generative Adversarial Network)?

The GANs architecture consist of two neural networks:

- **Generator**: Creates new data samples by learning from random noise.
- **Discriminator**: Attempts to distinguish between real and generated data.

During training, the generator and discriminator compete against each other, leading to the improvement of both networks.

stacked GAN

What is a Stack GAN?

The Stack GAN architecture consists of multiple GANs stacked in layers.

stacked GAN

What is a Stack GAN?

The Stack GAN architecture consists of multiple GANs stacked in layers.

Why using a Stack GAN?

stacked GAN

What is a Stack GAN?

The Stack GAN architecture consists of multiple GANs stacked in layers.

Why using a Stack GAN?

- Can model highly complex and diverse datasets by leveraging the hierarchical representation of features.

stacked GAN

What is a Stack GAN?

The Stack GAN architecture consists of multiple GANs stacked in layers.

Why using a Stack GAN?

- Can model highly complex and diverse datasets by leveraging the hierarchical representation of features.
- Different layers can focus on generating specific details, leading to fine-grained and realistic samples.

stacked GAN

What is a Stack GAN?

The Stack GAN architecture consists of multiple GANs stacked in layers.

Why using a Stack GAN?

- Can model highly complex and diverse datasets by leveraging the hierarchical representation of features.
- Different layers can focus on generating specific details, leading to fine-grained and realistic samples.
- It stabilize the training process, making it easier to train deep generative models.

About Implementation

model

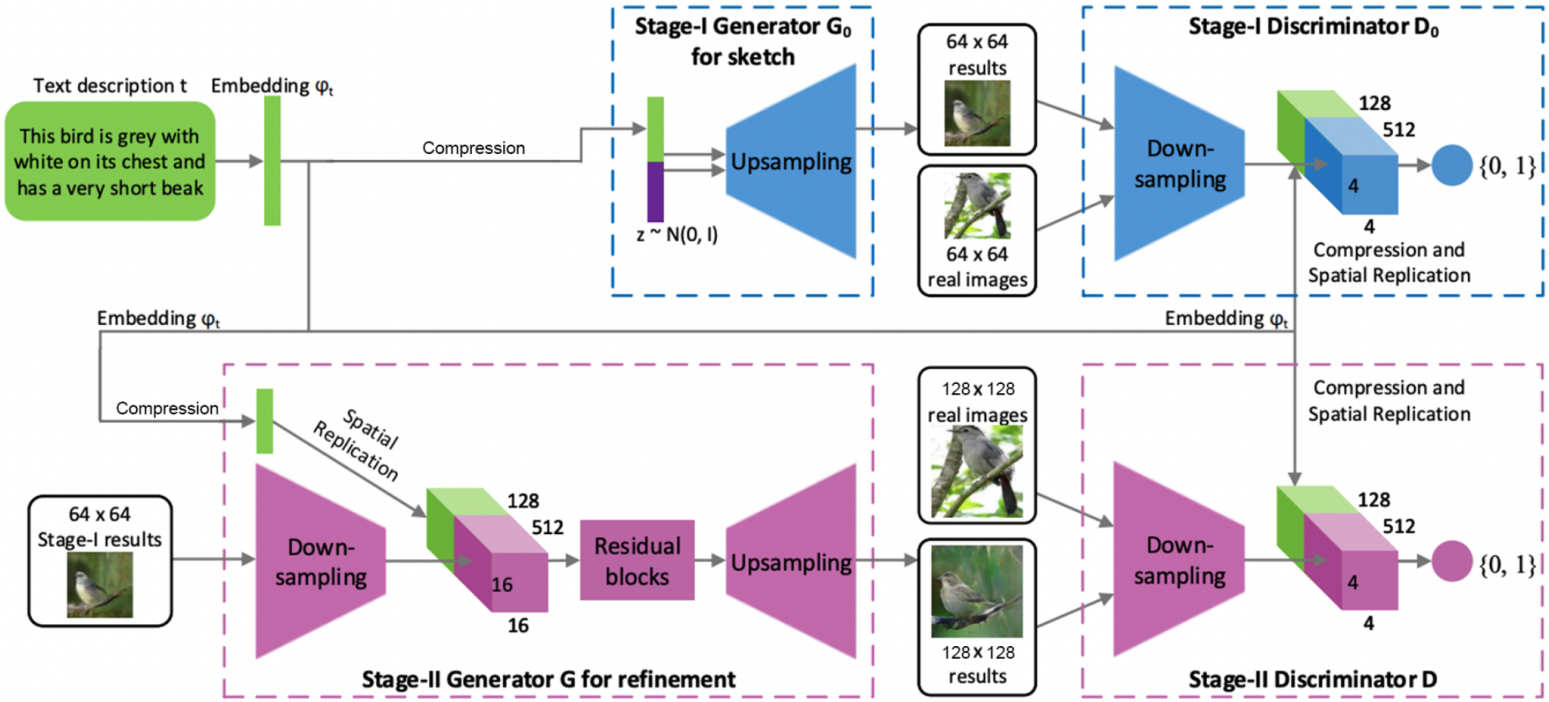


Figure: Model

About Implementation

training

- Datasets:

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each
  - *Birds*: containing embeddings of text from a pre-trained network;

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each
  - *Birds*: containing embeddings of text from a pre-trained network;
- We used the BCE (Binary Cross Entropy) as criterion

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each
  - *Birds*: containing embeddings of text from a pre-trained network;
- We used the BCE (Binary Cross Entropy) as criterion
  - $\mathcal{L}_D = \mathbb{E}_{(I,t)\sim p_{data}}[log D(I, \varphi_t)] + \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each
  - *Birds*: containing embeddings of text from a pre-trained network;
- We used the BCE (Binary Cross Entropy) as criterion
  - $\mathcal{L}_D = \mathbb{E}_{(I,t)\sim p_{data}}[log D(I, \varphi_t)] + \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$
  - $\mathcal{L}_G = \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each
  - *Birds*: containing embeddings of text from a pre-trained network;
- We used the BCE (Binary Cross Entropy) as criterion
  - $\mathcal{L}_D = \mathbb{E}_{(I,t)\sim p_{data}}[log D(I, \varphi_t)] + \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$
  - $\mathcal{L}_G = \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$
- We used the Adam optimizer
  - *learning rate* $= 0.0002$ (decays by 50% every 20 epochs)
  - *betas* $= (0.5, 0.999)$

About Implementation

training

- Datasets:
  - *CUB-200-2011*:contains 200 spaces of birds with several images for each
  - *Birds*: containing embeddings of text from a pre-trained network;
- We used the BCE (Binary Cross Entropy) as criterion
  - $\mathcal{L}_D = \mathbb{E}_{(I,t)\sim p_{data}}[log D(I, \varphi_t)] + \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$
  - $\mathcal{L}_G = \mathbb{E}_{z\sim p_z, t\sim p_{data}}[log(1 - D(G(z, \hat{c}), \varphi_t))]$
- We used the Adam optimizer
  - *learning rate* $= 0.0002$ (decays by 50% every 20 epochs)
  - *betas* $= (0.5, 0.999)$
- We trained both GANs for 150 epochs with 64 batch size.

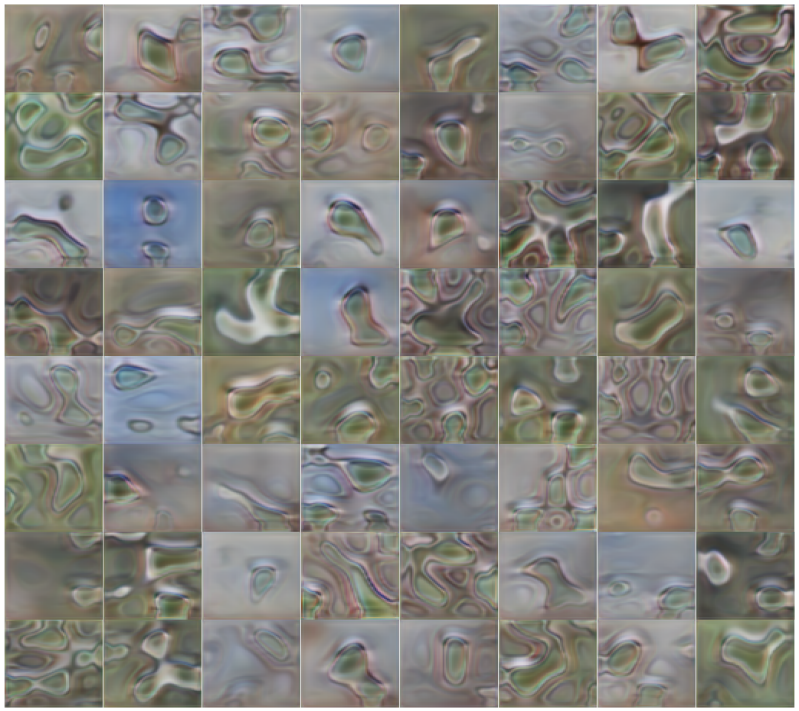
Loss during training

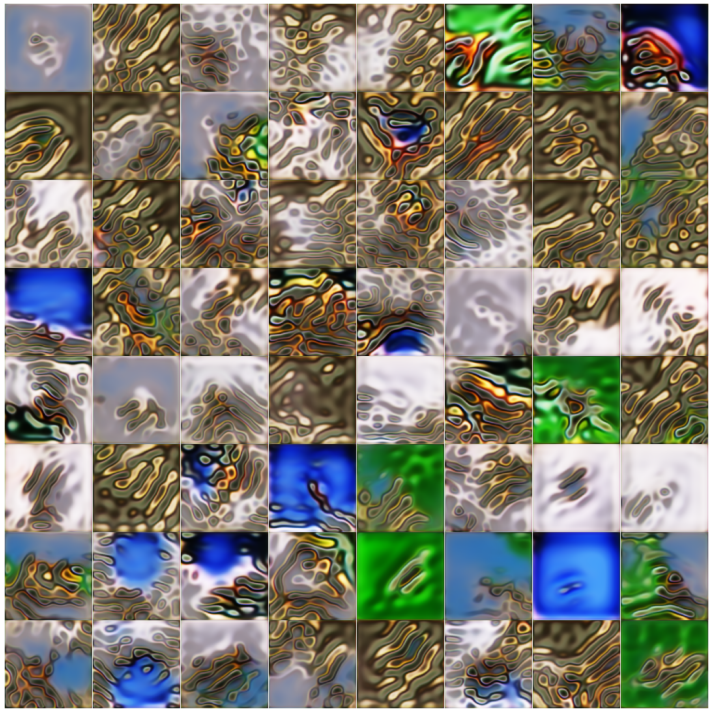
(a) first stack GAN.

(b) second stack GAN.

Figure: evolution of the losses during the training of the networks

Images during training





Generated image Examples

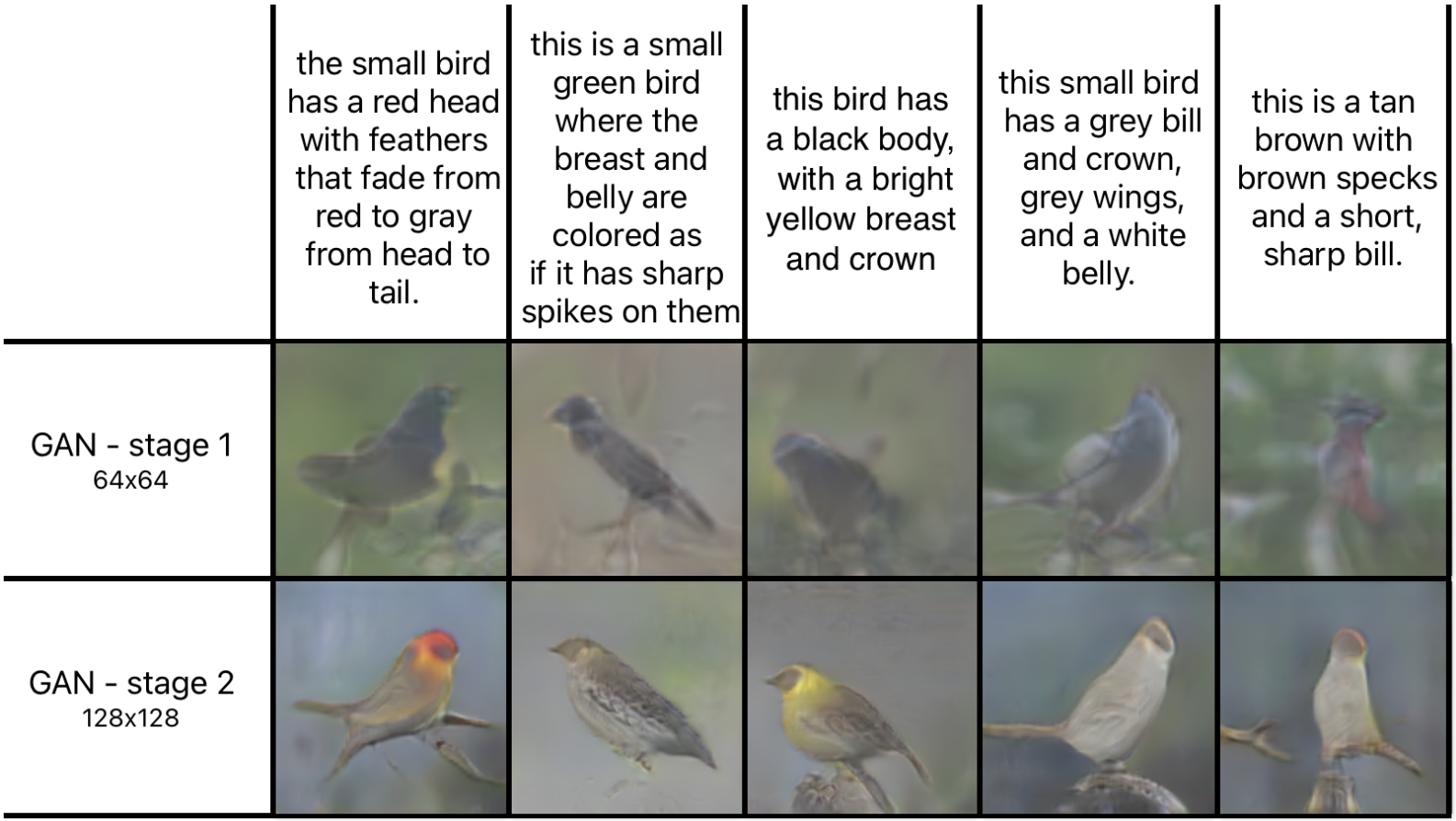


Figure: stackGAN results

Final considerations

Summarizing:

Final considerations

Summarizing:

- We proposed a Stacked GAN for synthesizing 128x128 images from text;

Final considerations

Summarizing:

- We proposed a Stacked GAN for synthesizing 128x128 images from text;
  - stack-GAN 1: sketches basic color and shape;

Final considerations

Summarizing:

- We proposed a Stacked GAN for synthesizing 128x128 images from text;
  - stack-GAN 1: sketches basic color and shape;
  - stack-GAN 2: corrects the defects and adds more details.

Final considerations

Summarizing:

- We proposed a Stacked GAN for synthesizing 128x128 images from text;
  - stack-GAN 1: sketches basic color and shape;
  - stack-GAN 2: corrects the defects and adds more details.
- We demonstrated that this architecture is effective for generating high-resolution images.

Final considerations

Summarizing:

- We proposed a Stacked GAN for synthesizing 128x128 images from text;
  - stack-GAN 1: sketches basic color and shape;
  - stack-GAN 2: corrects the defects and adds more details.
- We demonstrated that this architecture is effective for generating high-resolution images.
- This feature could be appreciated even more by modifying the network to generate higher resolution images (ex. 256x256).

Thank you for your attention