GAN: Text to image

Ilias Alexandropoulos mtn2307
Vasiliki Rentoula mtn2317

University of Piraeus, NCSR “Demokritos”

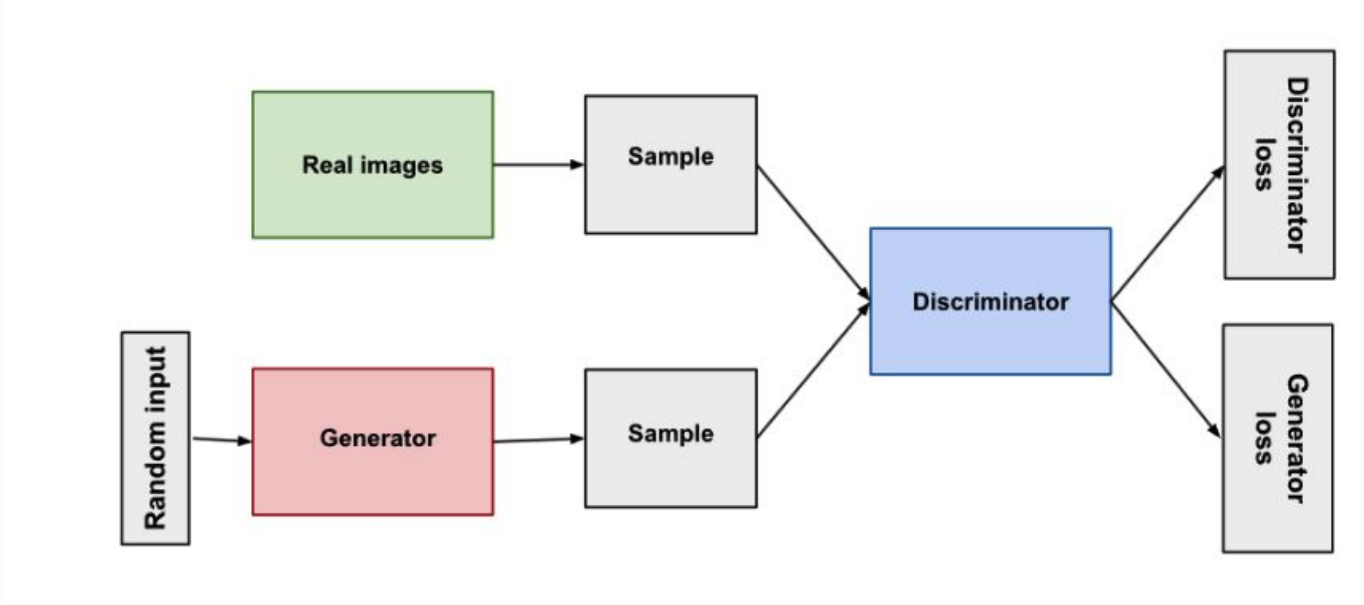
https://arxiv.org/pdf/1612.03242v1


Our task


The scope of this exercise is to create **text-to-image** (T2I) **bird images**. For this purpose, we chose the CUB-200-2011p [1] dataset retrieved from Kaggle. We aim to develop a GAN-based model that generates high-resolution, realistic bird images from textual descriptions.

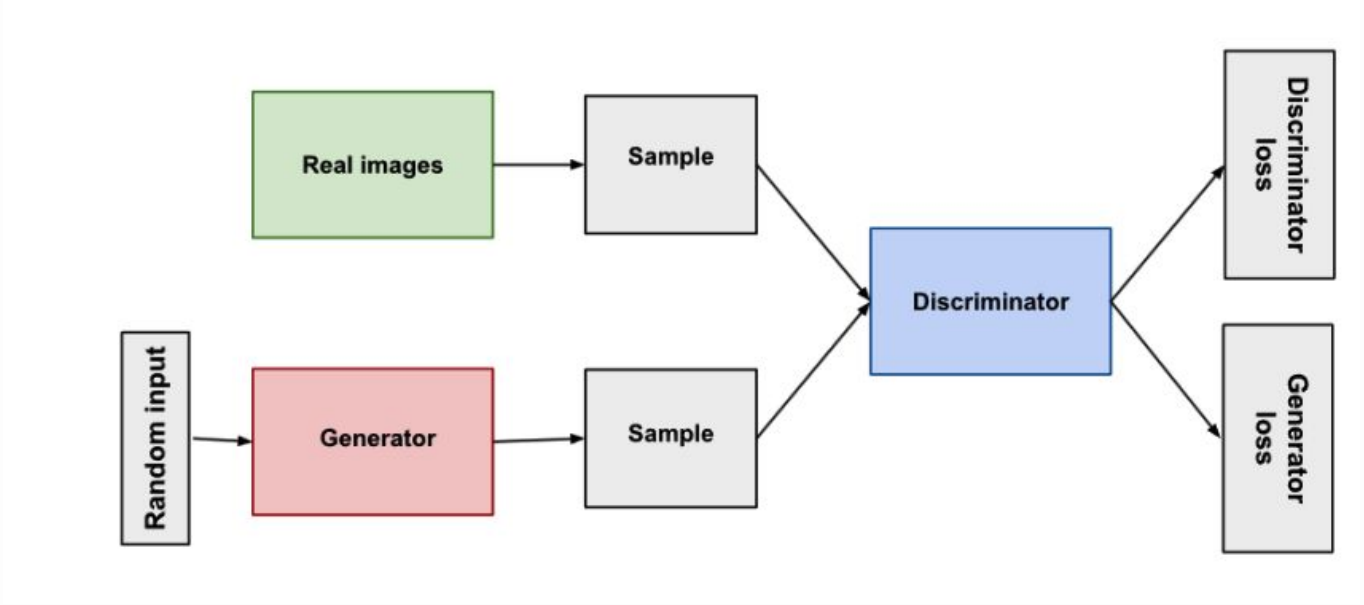
[1] https://www.kaggle.com/datasets/wenewone/cub2002011

GANS explained



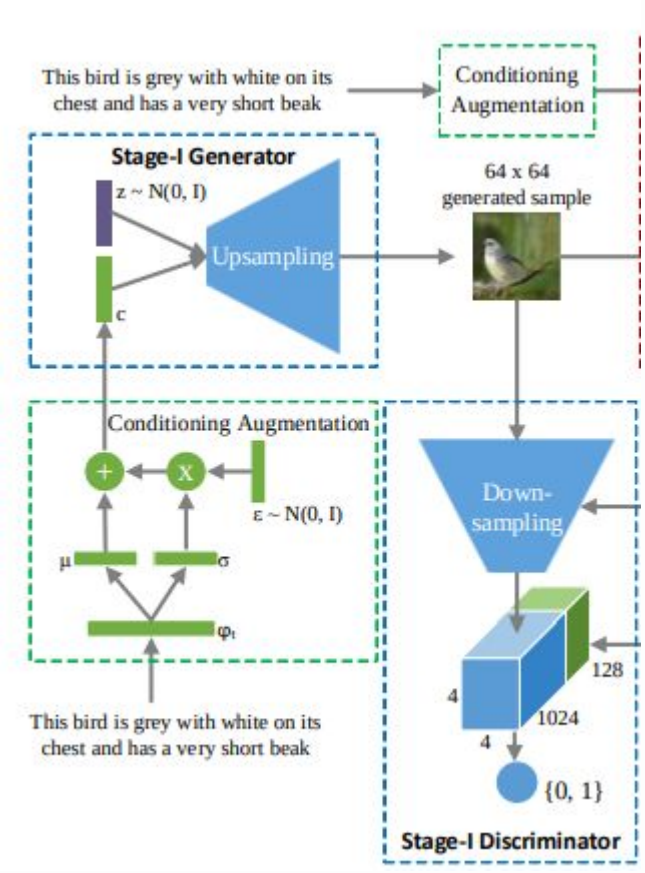
Generator
The generator learns to create fake data by incorporating feedback from the discriminator. The goal is to make the discriminator classify its output as real

GANS explained



Discriminator
The discriminator in a GAN is simply a classifier. It tries to distinguish real data from the data created by the generator

Conditional GAN



Data Prepossessing

01 Collect dataset

CUB-200-2011 contains 11,788 bird images divided into 200 categories. Each image has a text embedding and a bounding box.

02 Loading dataset

Loading the CUB dataset from the folder

03 Bounding Box extraction

Extract bounding boxes from the pre-defined bounding boxes of the dataset

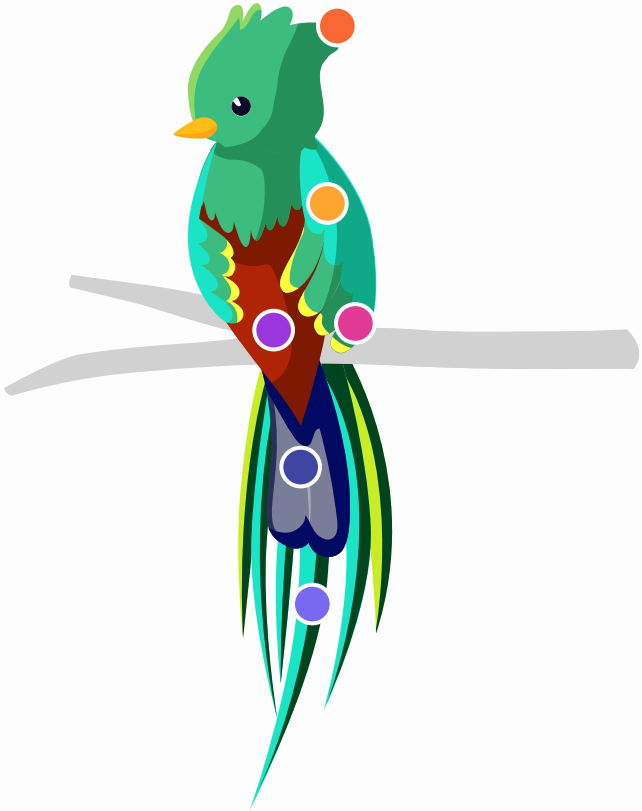
04 Image transformation

Random cropping (64x64)
Horizontal flipping
Normalization

05 Text embedding preparation

Load train pretrained embeddings along with their corresponding caption

06 Data loading & bathing



CNN

Generator: Transforms combined noise and text embedding into realistic images through progressive spatial enlargement via upsampling blocks.

- Convolutional transpose layers used to perform upsampling (nearest neighbor upsampling).

Discriminator: Evaluates the authenticity of generated images using downsampling blocks to process and classify images based on learned features

- Convolutional layers for performing downsampling.

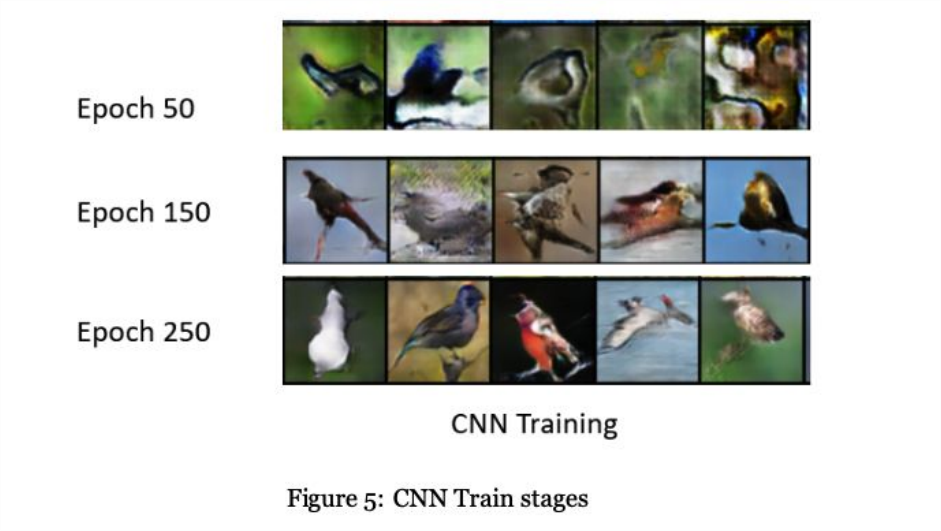


Figure 5: CNN Train stages

Resblock

Avoid the vanishing gradient problem in deep networks. Each block performs two consecutive 3x3 convolutions and adds the input directly to the output of these convolutions (residual connection), followed by a ReLU activation

Figure 6: Resblock Training

SE-Enhanced ResBlocks

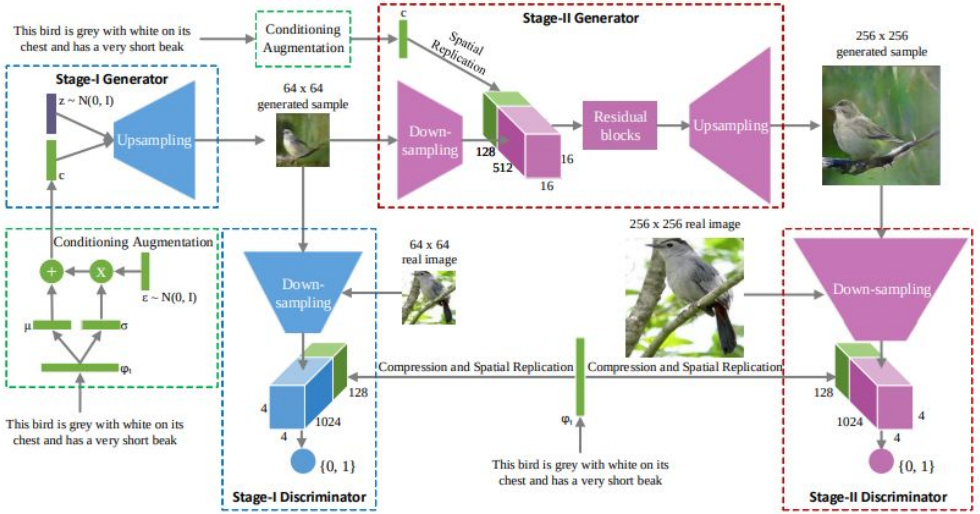
SE blocks recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels. This mechanism enhances the representational power of the network by allowing it to focus on more informative features.

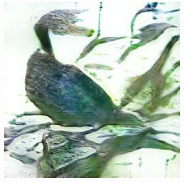
Figure 8: SE Blocks Training

Stage 2 Resblocks

Resblock stage 2 Training

Figure 7: Resblock stage 2



	light tan colored bird with a white head and an orange beak	small bird with grey feathers and a think black beak	this bird is black in color with a black beak and black eye rings.	a bird with a short, rounded beak which ends in a point, stark white eyes, and white throat.	the bird has a grey throat and white belly and breast.	a small bright yellow bird with black eyes, yellowish green back, and a large black spot on its crown.	this is a brown and white bird with a large downward pointing beak.
CNN							
Resblock							
SE block							
Stage2							

Results

Model	Inception Score	Human Score
CNN	4.74 ± 0.54	0.3
Resblock	4.61 ± 0.52	0.32
SE block	3.51 ± 0.46	0.22
Stage 2	4.85 ± 0.51	0.4

Table 1: Evaluation Metrics Table

Thank you!