



Mohamed Sharaf (54)  
Mahmoud Tarek (63)



---

## Introduction

“A picture is worth a thousand words!”. What if we can generate a realistic image from its description? In recent years, a lot of research and interest focused on text-to-image generation. In this proposal, we define the problem and its challenges, we review the related work in the literature with the goal of building on them.

## Problem Statement

Text-to-image generation is to translate text in the form of human-written description into image that is indistinguishable from realistic one. Examples of input/output are illustrated in the following table.

Input	Output
a flower with long pink petals and raised orange stamen.	 <a href="#">[1]</a>
a sheep standing in an open grass field.	 <a href="#">[1]</a>

---

## Motivation

Text-to-image generation has many practical and useful applications including, but not limited to:

- Computer-aided tools
- Language learning
- Literacy development
- Storytelling
- Art generation

Traditional way to solve this problem was relying mainly on word to image correlation analysis combined with supervised methods to find the best alignment of the visual content matching the text<sup>[2]</sup>. The main limitation in this solution is that models lack the ability to generate new image content; they can only change the characteristics of the given images. With the advances in Deep Learning, new methods have been introduced, particularly using generative adversarial networks (GANs).

## Challenges

There are many challenges in tackling this problem, they can be summarized as following<sup>[1]</sup>:

- Learning a text feature representation that captures the important visual details (Natural Language Representation)
- Using these features to synthesize a pseudo-real image (Image Synthesis)
- The distribution of images conditioned on a text description is highly multimodal, in the sense that there are many images that correctly illustrates the same description

---

## Datasets

There are three main datasets that are commonly used in literature. In this section we summarize them.

### CUB-200

Image dataset with photos of 200 bird species. There are two versions of the dataset. One contains 6,033 images, and the other contains 11,788 images. Both are associated with text descriptions of the images. The dataset is available at:

<http://www.vision.caltech.edu/visipedia/CUB-200-2011.html>

### Oxford-102

It contains 102 categories of flowers with 40-258 images each (with total 8,189 images) along with their text descriptions. The dataset is available at:

<https://www.robots.ox.ac.uk/~vgg/data/flowers/102/>

### MS-COCO

The two previous datasets contain a single object which makes them simple. More complex dataset is MS-COCO. It contains 328k images with 91 different object types. The dataset is available at: <http://cocodataset.org/>

---

## Evaluation Metrics

It is difficult to evaluate the performance of generative models. However, there are some widely-used evaluation metrics used in literature.

### Inception Score (IS)<sup>[3]</sup>

IS uses two criteria in measuring GAN performance which are the quality of generated images and their diversity. Inception network is used to classify the generated images and predict  $P(y|x)$  — where  $y$  is the label and  $x$  is the generated data. This reflects the quality of the images, and to measure diversity the data distribution for  $y$  should be uniform (i.e. high entropy) where  $P(y)$  is the marginal probability. To combine these two criteria, KL-divergence is computed between them and IS is then computed using the following equation

$$\text{IS}(G) = \exp \left( \mathbb{E}_{\mathbf{x} \sim p_g} D_{KL} \left( \underset{\text{higher is better}}{p(y|\mathbf{x})} \parallel \underset{\text{lower is better}}{p(y)} \right) \right)$$

Higher IS values are better.

### Fréchet Inception Distance (FID)<sup>[3]</sup>

FID uses Inception network to extract features from an intermediate layer. Then we model the data distribution for these features using a multivariate Gaussian distribution with mean  $\mu$  and covariance  $\Sigma$ . The following equation is used to compute The FID between the real images  $x$  and generated images  $g$

$$\text{FID}(x, g) = \|\mu_x - \mu_g\|_2^2 + \text{Tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}),$$

Lower FID values mean better image quality and diversity.

---

## Semantic Object Accuracy (SOA)<sup>[4]</sup>

Introduced as most evaluation metrics only judge image quality but not the conformity between the image and its caption. It specifically evaluates images given an image caption using a pre-trained object detector to evaluate if a generated image contains objects that are specifically mentioned in the image caption, e.g. whether an image generated from “a car driving down the street” contains a car. Calculations are complex and can be found in [\[4\]](#).

---

## Literature Review

In this section we review some selected research work focusing on their approach, architecture, used datasets, and results. Comparative analysis between them is shown at the end of this section.

### DC-GAN (Reed et al., 2016)<sup>[1]</sup>

**Approach:** Train a deep convolutional generative adversarial network (DC-GAN) conditioned on text features encoded by a hybrid character-level convolutional recurrent neural network.

**Architecture:**

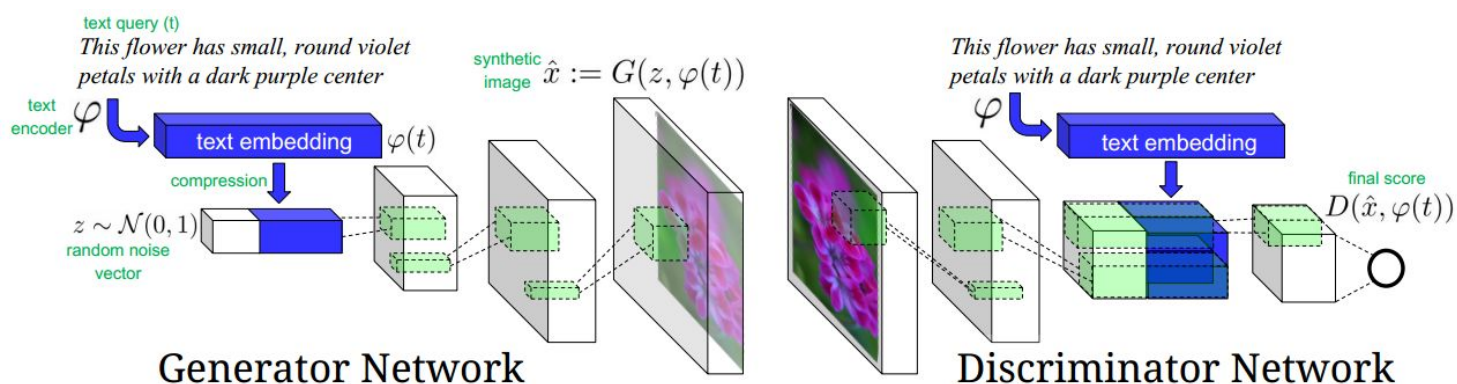


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding  $\varphi(t)$  is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

**Dataset:** They trained the model on CUB and Oxford-102 datasets. The training image size was set to  $64 \times 64 \times 3$ . For text features, they first pre-train a deep convolutional recurrent text encoder on structured joint embedding of text captions with 1,024-dimensional GoogLeNet image embeddings. They used MS-COCO to show the generalization capability of the approach on a general set of images that contain multiple objects and variable backgrounds.

**Results:** They didn't provide quantitative measures, but by human inspection, the performance on single object datasets was better than multi-objects dataset (i.e. MS-COCO). This is the case for most models we discuss.



## StackGAN (Zhang et al., 2017)<sup>[5]</sup>

**Approach:** Decomposing the problem into two more manageable sub-problems with stacked Generative Adversarial Networks (StackGAN). A low resolution image is generated using Stage-I GAN. It sketches the primitive shape and basic colors of the object. By conditioning on the generated image and the text again, Stage-II GAN then learns to capture text information that are omitted by Stage-I GAN and draws more details for the object yielding high resolution image.

### Architecture:

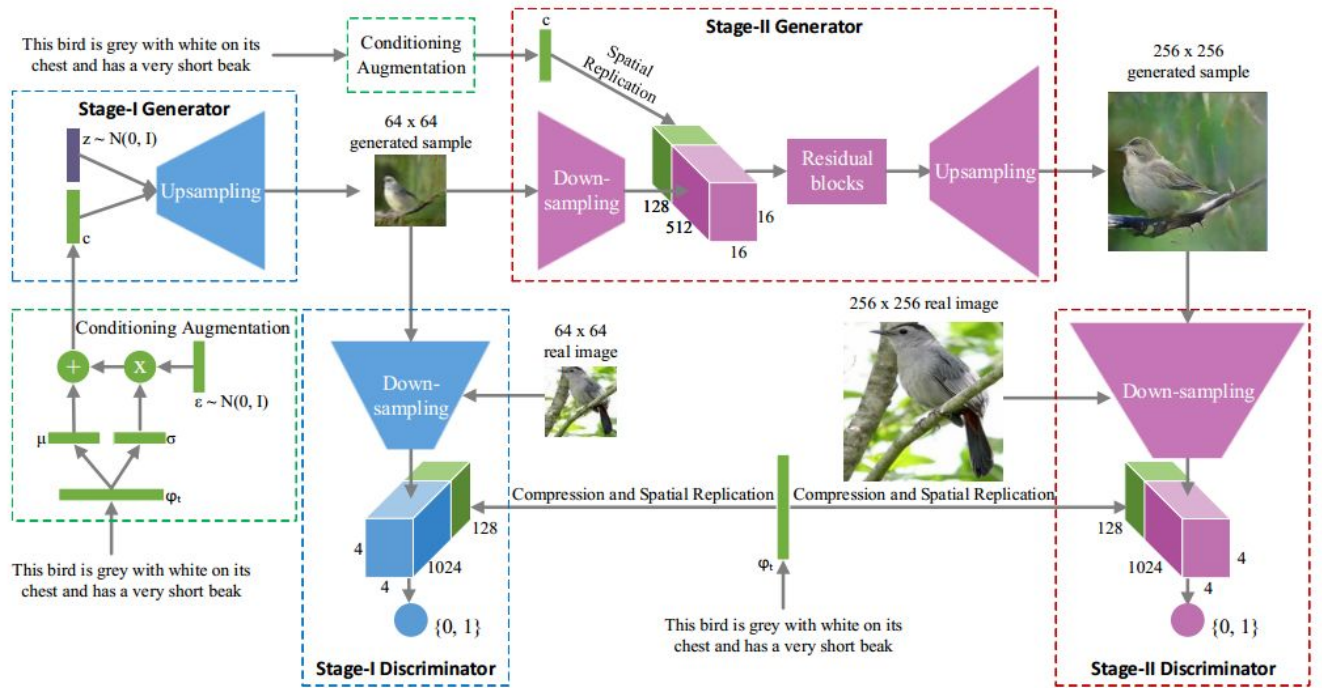


Figure 2. The architecture of the proposed StackGAN. The Stage-I generator draws a low resolution image by sketching rough shape and basic colors of the object from the given text and painting the background from a random noise vector. The Stage-II generator generates a high resolution image with photo-realistic details by conditioning on both the Stage-I result and the text again.

**Dataset:** Using CUB and Oxford-102 datasets, 256x256 images are produced.

**Results:** It achieves 28.47% improvement in terms of inception score on CUB dataset, and 20.30% improvement on Oxford-102. In some cases, the model fails if Stage-I GAN fails to generate plausible shapes or colors of the objects.

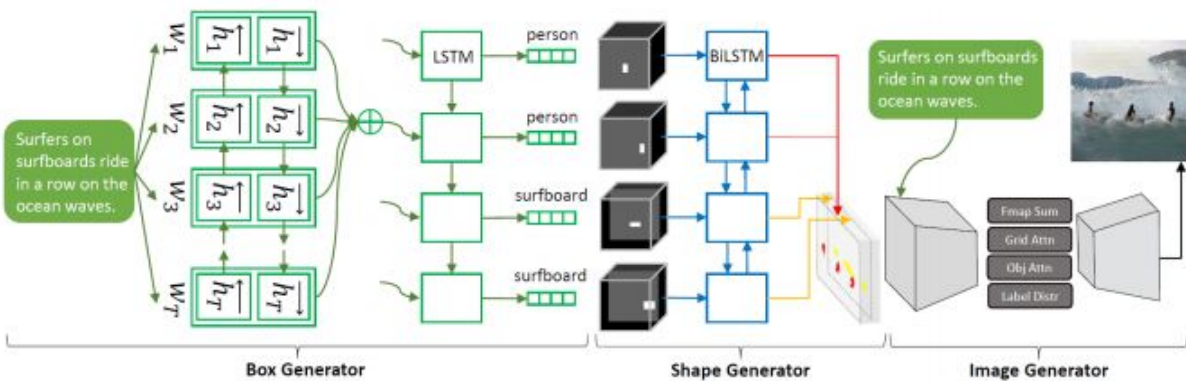


## Obj-GAN (Li et al., 2019)<sup>[6]</sup>

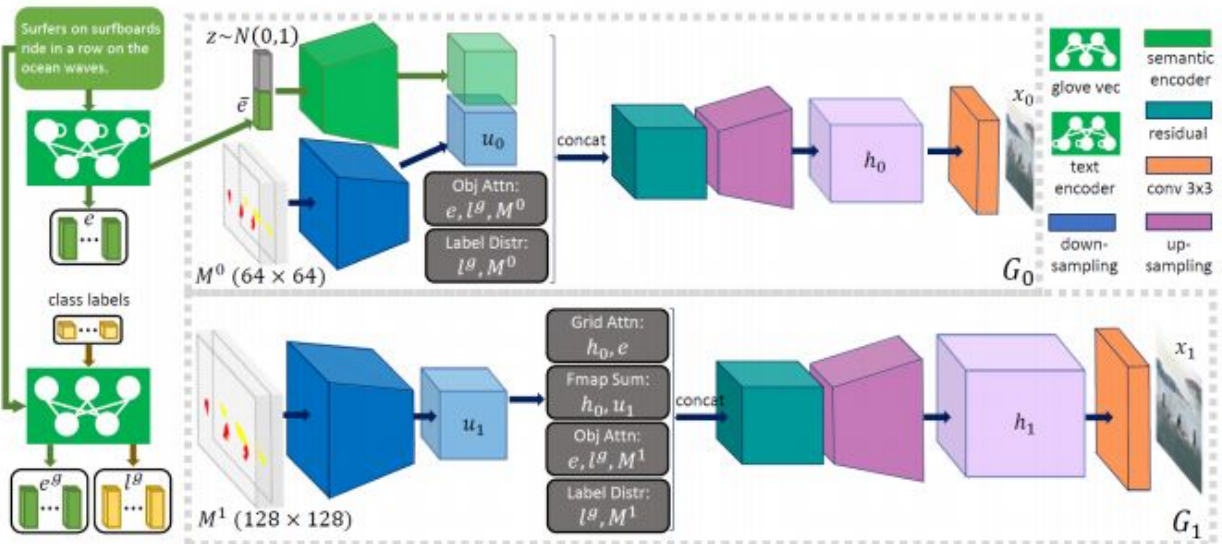
**Approach:** Proposing Obj-GAN that allows object-centered text-to-image synthesis for complex scenes. A generator is proposed to synthesize objects to the most relevant words. A R-CNN discriminator is proposed to determine if synthesized object matches the image caption.

### Architecture:

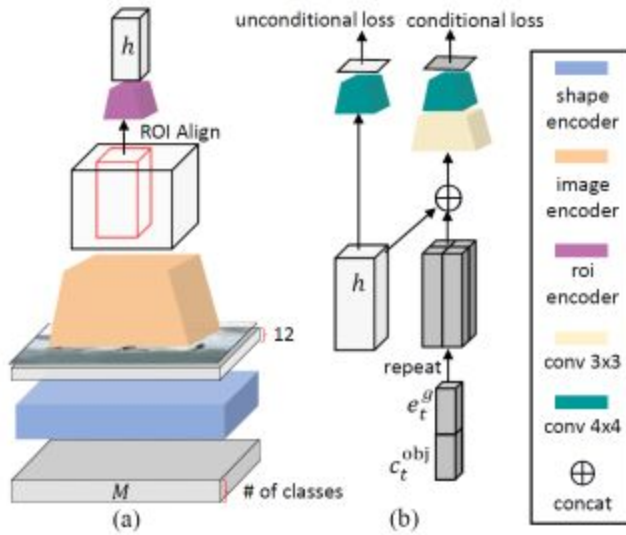
Obj-GAN:



Generator:



Discriminator:



**Dataset:** They used the MS-COCO dataset for evaluation, they used the official 2014 train (over 80K images) and validation (over 40K images) splits for training and test stages, respectively.

**Results:** They found that IS can be saturated, even over-fitted, while FID is a more robust measure and aligns better with human qualitative evaluation.

## Comparative Analysis

Comparison<sup>[2]</sup> between the discussed models and the state-of-the-art one is shown in the following table. (dash indicates that data not found)

Model	Datasets & Metrics					
	CUB-200		Oxford-102		MS-COCO	
	IS	FID	IS	FID	IS	FID
DC-GAN	2.88	68.79	2.66	79.55	7.88	60.82
StackGAN	3.74	51.89	3.21	55.28	8.6	74.05
Obj-GAN	-	-	-	-	30.11	20.75
DM-GAN	4.82	16.09	-	-	31.06	32.64

We noted that DM-GAN<sup>[2]</sup> is the state-of-the-art model, followed by Obj-GAN.

---

## Our Contribution

**Model to build on:** DC-GAN. We choose it as it's architecture is not complex taking into consideration the lack of hardware resources and training time. As used in literature, we will evaluate our contribution using IS, and FID metrics discussed in the Evaluation Metrics section.

### Proposed contribution:

As we still need to have a strong background in the architecture of GANs, we don't guarantee that our contribution is feasible, however

- We will **tune DC-GAN hyper-parameters** trying to increase its performance.
- We will **try to modify the architecture** to get more realistic output images for either birds (CUB dataset) or flowers (Oxford-102 dataset).
- As shown in the literature review section, DC-GAN produces images that contain only one object. We will **explore the capability of generating 2 specific objects in one image (containing birds and flowers)**. We can filter the COCO dataset to get images with flowers and birds in the same text description and use it to train our model allowing it to learn the relation between 2 objects in a single image, as we not expecting a good output if we train the model using both birds (CUB dataset) and flowers (Oxford-102 dataset) merged together in one dataset.

## Graduation Project Problem Statement

- Scaling Verifiable E-voting systems using public Blockchain
- DeepFake, Using StyleGan2 we transfer source voice and image to target video

---

## References

1. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016b). Generative adversarial text to image synthesis. Proceedings of the International Conference on Machine Learning (ICML). Available: <https://arxiv.org/pdf/1605.05396.pdf>
2. Agnese, J., Herrera, J., Tao, H., Zhu, X.: A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis. (2019). Available: <https://arxiv.org/pdf/1910.09399.pdf>
3. GAN — How to measure GAN performance? - Jonathan Hui. Available: [https://medium.com/@jonathan\\_hui/gan-how-to-measure-gan-performance-64b988c47732](https://medium.com/@jonathan_hui/gan-how-to-measure-gan-performance-64b988c47732)
4. Hinz, Tobias & Heinrich, Stefan & Wermter, Stefan. (2019). Semantic Object Accuracy for Generative Text-to-Image Synthesis. Available: <https://arxiv.org/pdf/1910.13321v1.pdf>
5. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, Z., Huang, X., and Metaxas, D. (2017b). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In IEEE International Conference on Computer Vision (ICCV), Venice, pages 5908–5916. Available: <https://arxiv.org/pdf/1612.03242v2.pdf>
6. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., and Gao, J. (2019a). Object-driven text-to-image synthesis via adversarial training. Available: <https://arxiv.org/pdf/1902.10740.pdf>
7. Zhu, M., Pan, P., Chen, W., and Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5802–5810. Available: <https://arxiv.org/pdf/1904.01310.pdf>