

# URL Coursework 2

StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks [1]

Bruno Sánchez Gómez

June 1, 2025

# Table of Contents

- 1 Introduction
- 2 StackGAN++ Methodology
- 3 Results
- 4 Limitations & Ablation Studies
- 5 Conclusions

## Area of Research of the Paper

### **Generative Adversarial Networks** (GANs) for Image Generation

- **StackGAN++** focuses on:
  - High-resolution images ( $256 \times 256$  pixels)
  - Two tasks:
    - *Unconditional Image Generation*
    - *Text-to-Image Synthesis* (Conditional Image Generation)

## Area of Research of the Paper

### **Generative Adversarial Networks (GANs) for Image Generation**

- **StackGAN++** focuses on:
  - High-resolution images ( $256 \times 256$  pixels)
  - Two tasks:
    - *Unconditional Image Generation*
    - *Text-to-Image Synthesis* (Conditional Image Generation)
- **Limitations of prior work with GANs:**
  - Training instability
  - Mode collapse
  - Difficulty in generating high-resolution images

# Contributions by StackGAN++

- **Conditioning Augmentation (CA):** Improve sample diversity by augmenting the image-text pairs.
- **StackGAN:** Two GAN frameworks for conditional and unconditional image synthesis with high resolution ( $256 \times 256$ ):
  - **StackGAN-v1:** Two-stage GAN
  - **StackGAN-v2:** Multi-stage GAN with a tree-like structure

# Table of Contents

- 1 Introduction
- 2 StackGAN++ Methodology
- 3 Results
- 4 Limitations & Ablation Studies
- 5 Conclusions

# Conditioning Augmentation (CA)

## Core Idea

Augment the text conditioning to improve sample diversity and stabilize GAN training

- The latent space for text embeddings,  $\phi_t$ , is high-dimensional
- Limited data causes discontinuity in the latent data manifold
- CA samples a new embedding  $\hat{c}$  from a Gaussian distribution:

$$\hat{c} = N(\mu_\theta(\phi_t), \Sigma_\theta(\phi_t))$$

- To further enforce smoothness over the conditioning manifold and avoid overfitting, a regularization term is added to the loss:

$$\mathcal{L}_{KL} = D_{KL}(N(\mu_\theta(\phi_t), \Sigma_\theta(\phi_t)) \parallel N(0, I))$$

## Core Idea

## Decompose text-to-image generation into a sketch-refinement process

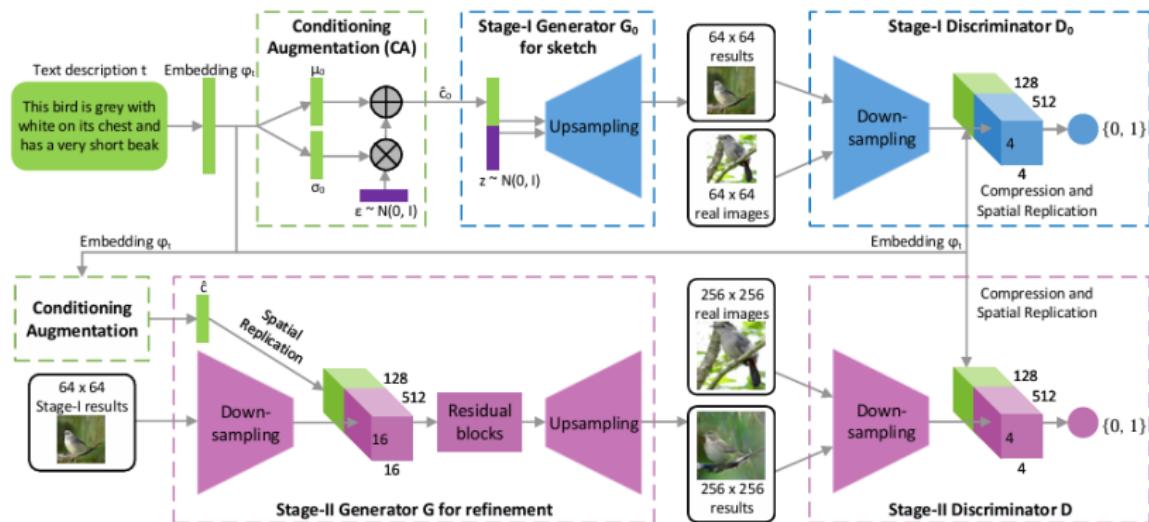


Figure: StackGAN-v1 Architecture (Source: [1])

## Core Idea

A more general, end-to-end multi-stage framework with a tree-like structure

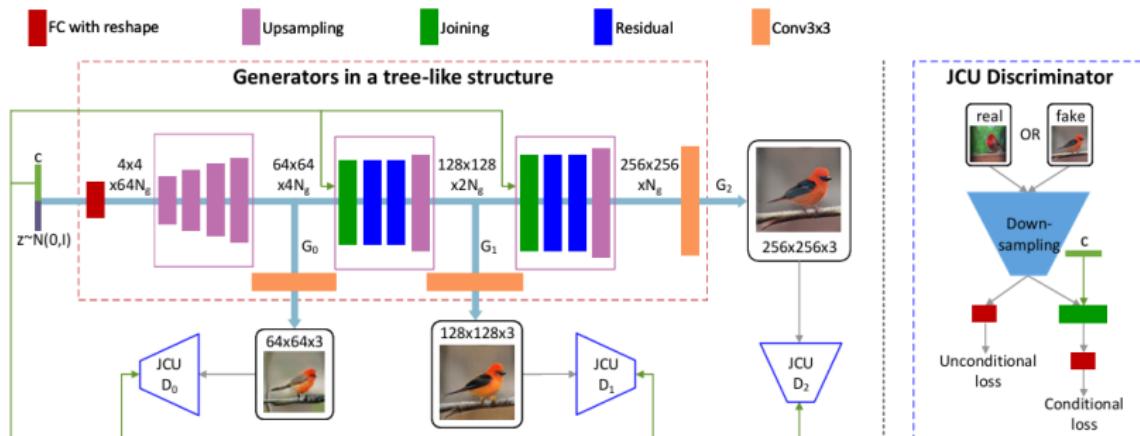


Figure: StackGAN-v2 Architecture and JCU Discriminators (Source: [1])

# Table of Contents

- 1 Introduction
- 2 StackGAN++ Methodology
- 3 Results
- 4 Limitations & Ablation Studies
- 5 Conclusions

# Experimental Setup

	<b>Unconditional</b>	<b>Conditional</b>
<b>Datasets</b>	LSUN ImageNet	CUB-200-2011 Oxford-102 MS COCO
<b>Competing Methods</b>	DCGAN WGAN EBGAN-PT LSGAN WGAN-GP	GAN-INT-CLS GAWWN

<b>Evaluation Metrics</b>	<i>Inception Score (IS) <math>\uparrow</math></i>
	<i>Fréchet Inception Distance (FID) <math>\downarrow</math></i>
	<i>Human Rank (HR) <math>\downarrow</math></i>

# Quantitative Results

Metric	CUB			Oxford		COCO	
	GAN-INT-CLS	GAWWN	Our StackGAN-v1	GAN-INT-CLS	Our StackGAN-v1	GAN-INT-CLS	Our StackGAN-v1
FID ↓	68.79	67.22	<b>51.89</b>	79.55	<b>55.28</b>	<b>60.62</b>	74.05
FID* ↓	68.79	53.51	<b>35.11</b>	79.55	<b>43.02</b>	60.62	<b>33.88</b>
IS ↑	2.88 ± .04	3.62 ± .07	<b>3.70 ± .04</b>	2.66 ± .03	<b>3.20 ± .01</b>	7.88 ± .07	<b>8.45 ± .03</b>
IS* ↑	2.88 ± .04	<b>3.10 ± .03</b>	3.02 ± .03	2.66 ± .03	<b>2.73 ± .03</b>	7.88 ± .07	<b>8.35 ± .11</b>
HR ↓	2.76 ± .01	1.95 ± .02	<b>1.29 ± .02</b>	1.84 ± .02	<b>1.16 ± .02</b>	1.82 ± .03	<b>1.18 ± .03</b>

TABLE 2: Inception scores (IS), fréchet inception distance (FID) and average human ranks (HR) of GAN-INT-CLS [35], GAWWN [33] and our StackGAN-v1 on CUB, Oxford-102, and COCO. (\* means that images are re-sized to  $64 \times 64$  before computing IS\* and FID\*)

Dataset		CUB	Oxford-102	COCO	LSUN-bedroom	LSUN-church	ImageNet-dog	ImageNet-cat
FID ↓	StackGAN-v1	51.89	55.28	<b>74.05</b>	91.94	57.20	89.21	58.73
	StackGAN-v2	<b>15.30</b>	<b>48.68</b>	81.59	<b>35.61</b>	<b>25.36</b>	<b>44.54</b>	<b>28.59</b>
IS ↑	StackGAN-v1	3.70 ± .04	3.20 ± .01	<b>8.45 ± .03</b>	<b>3.59 ± .05</b>	<b>2.87 ± .05</b>	8.84 ± .08	<b>4.77 ± .06</b>
	StackGAN-v2	<b>4.04 ± .05</b>	<b>3.26 ± .01</b>	8.30 ± .10	3.02 ± .04	2.38 ± .03	<b>9.55 ± .11</b>	4.23 ± .05
HR ↓	StackGAN-v1	1.81 ± .02	1.70 ± .03	<b>1.45 ± .04</b>	1.95 ± .01	1.86 ± .02	1.90 ± .01	1.88 ± .02
	StackGAN-v2	<b>1.19 ± .02</b>	<b>1.30 ± .03</b>	1.55 ± .05	<b>1.05 ± .01</b>	<b>1.14 ± .02</b>	<b>1.10 ± .01</b>	<b>1.12 ± .02</b>

TABLE 3: Comparison of StackGAN-v1 and StackGAN-v2 on different datasets by inception scores (IS), fréchet inception distance (FID) and average human ranks (HR).

Figure: Tables of quantitative results (Source: [1])

# Qualitative Results: Unconditional Image Generation



64×64 samples by DCGAN (Reported in [32])



64×64 samples by WGAN (Reported in [3])



64×64 samples by EBGAN-PT (Reported in [56])



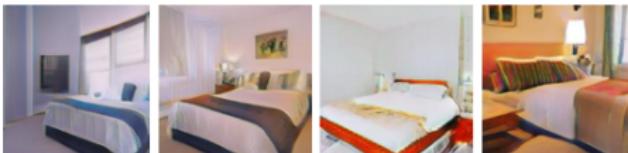
112×112 samples by LSGAN (Reported in [26])



128×128 samples by WGAN-GP (Reported in [13])



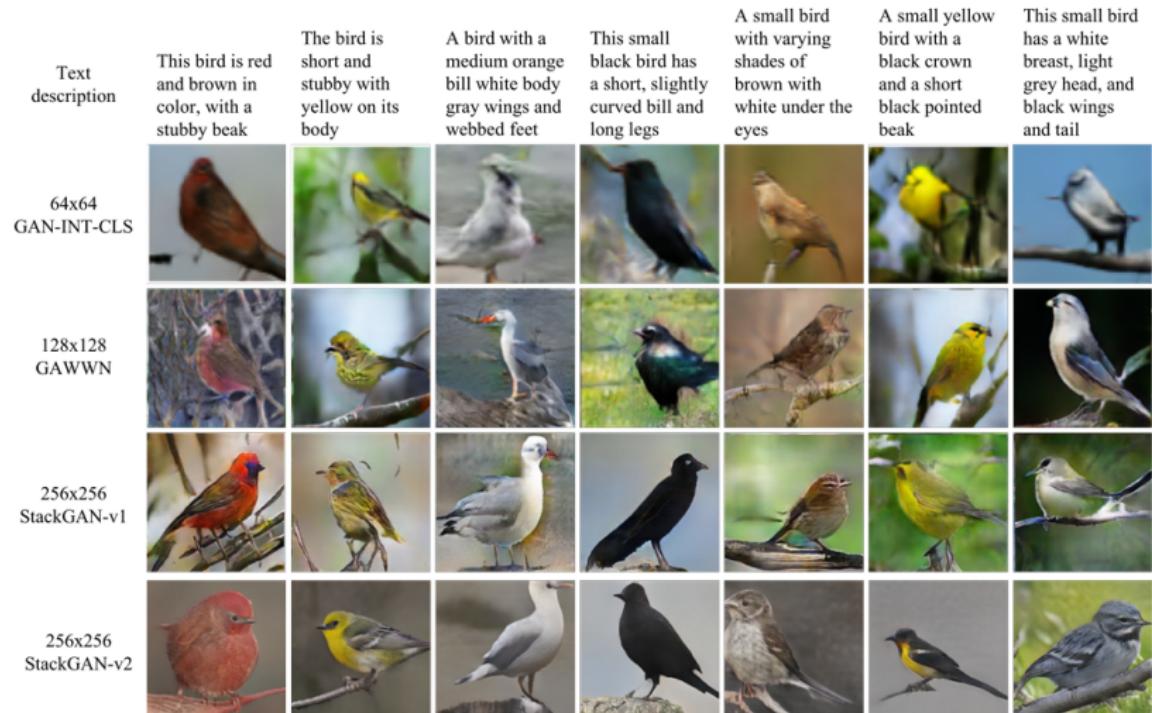
256×256 samples by our StackGAN-v1



256×256 samples by our StackGAN-v2

Figure: Comparison of generated samples from LSUN Bedroom (Source: [1])

# Qualitative Results: Text-to-Image



**Figure:** Comparison of generated samples with text descriptions from CUB  
(Source: [1])

# Table of Contents

- 1 Introduction
- 2 StackGAN++ Methodology
- 3 Results
- 4 Limitations & Ablation Studies
- 5 Conclusions

# Limitations: StackGAN-v1 Mode Collapse



(a) StackGAN-v1 has two collapsed modes (in red rectangles). (b) StackGAN-v2 contains no collapsed nonsensical mode.

Fig. 5: Utilizing t-SNE to embed images generated by our StackGAN-v1 and StackGAN-v2 on the CUB test set.

**Figure:** StackGAN-v1 suffers from mode collapse (Source: [1])

# Limitations: Failure Cases



Fig. 9: Examples of failure cases of StackGAN-v1 (top) and StackGAN-v2 (bottom) on different datasets.

**Figure:** Failure cases of both StackGAN-v1 and StackGAN-v2 (Source: [1])

# Ablation Studies

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-v1	yes	no	3.13 ± .03
	no	yes	3.20 ± .03
	yes	yes	3.35 ± .02
256×256 StackGAN-v1	yes	no	3.45 ± .02
	no	yes	3.31 ± .03
	yes	yes	3.70 ± .04

Figure: Component analysis of StackGAN-v1 (Source: [1])

Model	branch $G_1$	branch $G_2$	branch $G_3$	JCU	inception score
StackGAN-v2	64×64	128×128	256×256	yes	4.04 ± .05
StackGAN-v2-no-JCU	64×64	128×128	256×256	no	3.77 ± .04
StackGAN-v2- $G_3$	removed	removed	256×256	yes	3.49 ± .04
StackGAN-v2-3 $G_3$	removed	removed	three 256×256	yes	3.22 ± .02
StackGAN-v2-all256	256×256	256×256	256×256	yes	2.89 ± .02

Figure: Component analysis of StackGAN-v2 (Source: [1])

# Table of Contents

- 1 Introduction
- 2 StackGAN++ Methodology
- 3 Results
- 4 Limitations & Ablation Studies
- 5 Conclusions

# Conclusions

- ➊ **Conditioning Augmentation** significantly improves sample diversity and training stability
- ➋ **StackGAN-v1** succeeds in generating high-resolution images with photo-realistic details
- ➌ **StackGAN-v2** improves robustness by jointly approximating:
  - Multi-scale image distributions
  - Conditional and unconditional image distributions
- ➍ **Quantitative and Qualitative Results** demonstrate superior performance over prior SOTA methods
- ➎ **Ablation Studies** validate the effectiveness of each component

# Personal Comments

## Cons

- **JCU Discriminators** could have also been used in **StackGAN-v1**
- The authors **did not include StackGAN-v2 in the quantitative analysis** against SOTA methods
- The improvement in **image quality** from the qualitative results is **not very significant**

## Pros

- It maintains the **same level of quality** at **higher resolutions**
- The idea of progressive refinement as a way to tackle high-resolution image synthesis is **well-motivated**, **intuitive**, and **empirically validated**
- The paper was published in 2018, and since then, there have been **many advancements** in the field of GANs (e.g. *StyleGAN* [2])

# References I

-  Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas.  
Stackgan++: Realistic image synthesis with stacked generative adversarial networks, 2018.
-  Tero Karras, Samuli Laine, and Timo Aila.  
A style-based generator architecture for generative adversarial networks.  
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.