

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 Conclusions

Area of Research of the Paper

Generative Adversarial Networks (GANs) for Realistic Image Synthesis

- High-resolution images (256×256 pixels)
- Two tasks:
 - *Unconditional Image Generation*
 - *Text-to-Image Synthesis* (Conditional Image Generation)

Limitations of Prior Work

- **GAN Training Instability:**

- Sensitive to hyperparameters
- Can suffer from non-convergence

- **Mode Collapse:**

- Limited variety of generated samples
- Fail to capture full diversity of the training data

- **Limited to low-resolution images:**

- Training GANs for high-resolution images is especially difficult and unstable.
- Low overlap between model and data distributions → Poor gradients

Contributions by StackGAN++

- ➊ **Conditioning Augmentation (CA):** Improve sample diversity by augmenting the image-text pairs.
- ➋ **StackGAN:** Two GAN frameworks for conditional and unconditional image synthesis with high resolution (256×256):
 - **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 Conclusions

Conditioning Augmentation (CA)

Core Idea

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

- The latent space for text embeddings, ϕ_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

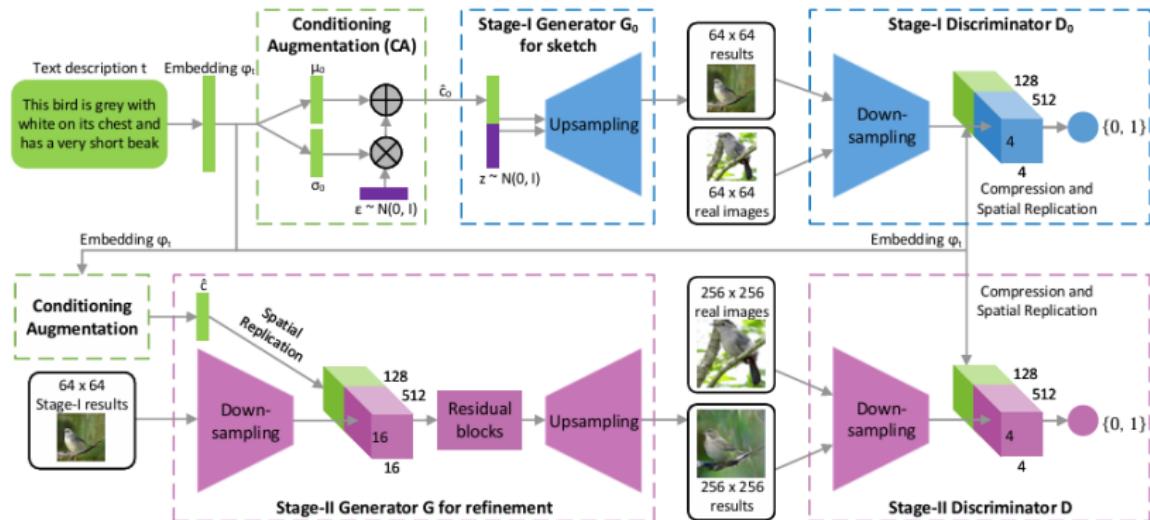


Figure: StackGAN-v1 Architecture (Source: [1])

Core Idea

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

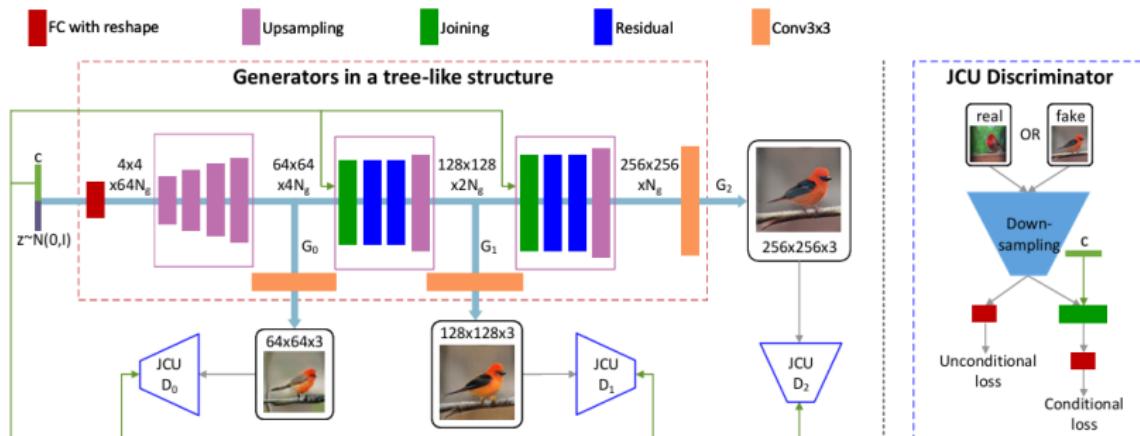


Figure: StackGAN-v2 Architecture and JCU Discriminators (Source: [1])

Table of Contents

1 Introduction

2 StackGAN++ Methodology

3 Results

4 Conclusions

Experimental Setup

Datasets	<i>Unconditional</i>	LSUN (Bedroom, Church) ImageNet (Dog, Cat subsets)
	<i>Conditional</i>	CUB-200-2011 (Birds) Oxford-102 (Flowers) MS COCO (Challenging general scenes)

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

Competing Methods	<i>Unconditional</i>	DCGAN, WGAN, EBGAN-PT, LSGAN, WGAN-GP
	<i>Conditional</i>	GAN-INT-CLS, GAWWN

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	51.89	79.55	55.28	60.62	74.05
FID* ↓	68.79	53.51	35.11	79.55	43.02	60.62	33.88
IS ↑	2.88 ± .04	3.62 ± .07	3.70 ± .04	2.66 ± .03	3.20 ± .01	7.88 ± .07	8.45 ± .03
IS* ↑	2.88 ± .04	3.10 ± .03	3.02 ± .03	2.66 ± .03	2.73 ± .03	7.88 ± .07	8.35 ± .11
HR ↓	2.76 ± .01	1.95 ± .02	1.29 ± .02	1.84 ± .02	1.16 ± .02	1.82 ± .03	1.18 ± .03

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 64×64 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	74.05	91.94	57.20	89.21	58.73
	StackGAN-v2	15.30	48.68	81.59	35.61	25.36	44.54	28.59
IS ↑	StackGAN-v1	3.70 ± .04	3.20 ± .01	8.45 ± .03	3.59 ± .05	2.87 ± .05	8.84 ± .08	4.77 ± .06
	StackGAN-v2	4.04 ± .05	3.26 ± .01	8.30 ± .10	3.02 ± .04	2.38 ± .03	9.55 ± .11	4.23 ± .05
HR ↓	StackGAN-v1	1.81 ± .02	1.70 ± .03	1.45 ± .04	1.95 ± .01	1.86 ± .02	1.90 ± .01	1.88 ± .02
	StackGAN-v2	1.19 ± .02	1.30 ± .03	1.55 ± .05	1.05 ± .01	1.14 ± .02	1.10 ± .01	1.12 ± .02

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])



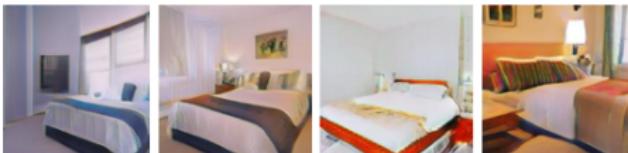
112×112 samples by LSGAN (Reported in [26])



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

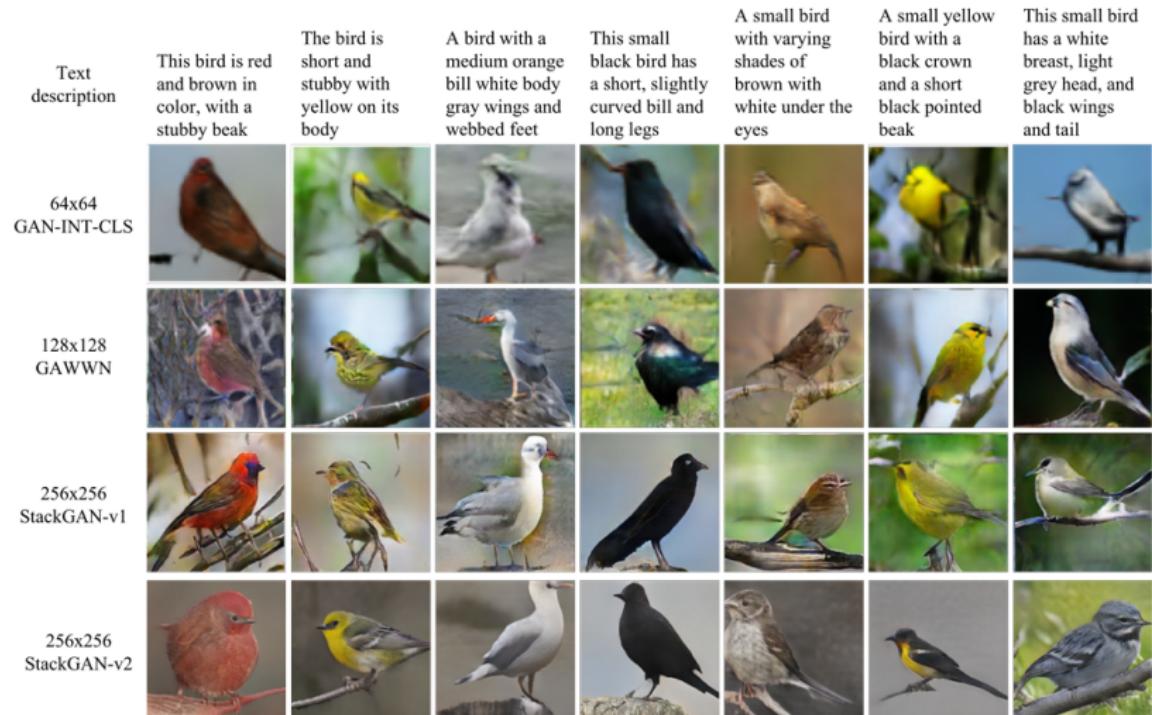
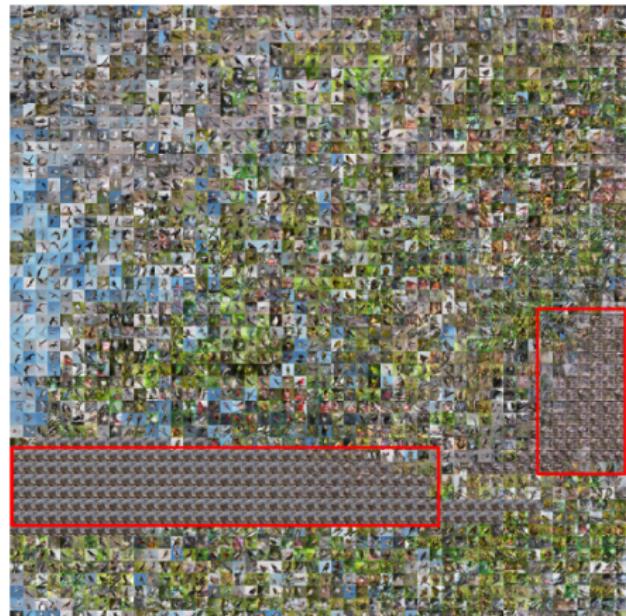


Figure: Comparison of generated samples with text descriptions from CUB
(Source: [1])

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 Conclusions

Conclusions

- **Stacked/Multi-Stage GANs** are highly effective
- **Conditioning Augmentation** significantly improves sample diversity and training stability
- **StackGAN-v1** succeeds in generating high-resolution images with photo-realistic details
- **StackGAN-v2** improves robustness by jointly approximating:
 - 1 Multi-scale image distributions
 - 2 Conditional and unconditional image distributions
- **Quantitative and Qualitative Results** demonstrate superior performance over prior SOTA methods
- **Ablation Studies** validate the effectiveness of each component

Cons

- 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.