URL Coursework 2

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

Bruno Sánchez Gómez

May 14, 2025

Table of Contents

1. Introduction
2. StackGAN++ Methodology
3. Results
4. Ablation Studies
5. Analysis

Area of Research of the Paper

Generative Adversarial Networks (GANs) for Realistic Image Synthesis

This paper focuses on generating high-quality, high-resolution images using GANs. Specifically, it addresses:

- **Text-to-Image Synthesis:** Generating photo-realistic images from textual descriptions (e.g., "a red bird with a short beak").
- **Conditional Image Generation:** Creating images based on various conditions, including text or class labels.
- **Unconditional Image Generation:** Synthesizing diverse images from random noise, learning the underlying data distribution.

The core approach involves **Stacked/Multi-Stage GAN architectures** to progressively refine images from low to high resolution (e.g., up to 256x256 pixels).

Problem Addressed

- **Generating High-Resolution Images:**
  - Training GANs for high-resolution (e.g., 256x256) images is notoriously difficult and unstable.
  - High-dimensional image spaces make it hard for model and data distributions to overlap, leading to poor gradients.
- **GAN Training Instability:**
  - GANs are sensitive to hyperparameters and can suffer from non-convergence.
- **Mode Collapse:**
  - Generators often produce a limited variety of samples, failing to capture the full diversity of the training data.
- **Limitations of Prior Work:**
  - Most previous methods were limited to low-resolution images.
  - Achieving higher resolutions often required strong supervision beyond text (e.g., object part locations).
  - Super-resolution techniques could only add minor details and couldn't fix major defects in low-resolution inputs.

Contributions: StackGAN & StackGAN++

The paper introduces two main frameworks:

StackGAN-v1:

- A two-stage GAN for text-to-image synthesis generating 256x256 photo-realistic images.
  - Stage-I: Low-resolution sketch (64x64).
  - Stage-II: High-resolution refinement (256x256), correcting defects.
- **Conditioning Augmentation (CA):** A novel technique to stabilize conditional GAN training and improve sample diversity by creating smoother conditioning manifolds.

StackGAN-v2:

- An advanced multi-stage GAN for both conditional and unconditional generation.
- **Tree-like Structure:** Multiple generators and discriminators for different image scales.
- **Joint Approximation of Multiple Distributions:** Stabilizes training by modeling related distributions at different scales.
- **Color-Consistency Regularization:** Ensures coherence across scales.

Table of Contents

1. Introduction

2. StackGAN++ Methodology

3. Results

4. Ablation Studies

5. Analysis

StackGAN-v1: Two-Stage Text-to-Image Synthesis

Core Idea: Decompose text-to-image generation into a sketch-refinement process.

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

- **Stage-I GAN:**
  - Input: Text description ‘t‘ + noise ‘z‘.
  - Uses **Conditioning Augmentation (CA)** on text embedding ‘ϕ_t‘ to get ‘ĉ $_0$‘. CA samples ‘ĉ $_0$‘ from ‘N($\mu_0(\phi_t)$, $\Sigma_0(\phi_t)$)‘, adding KL divergence regularization.
  - Generator ‘G $_0$‘: Produces a low-resolution image (64x64) focusing on rough shapes and colors.
  - Discriminator ‘D $_0$‘: Distinguishes real image-text pairs from fake ones.
- **Stage-II GAN:**
  - Input: Stage-I image + text ‘t‘ (again via CA to get ‘ĉ‘).
  - Generator ‘G‘: An encoder-decoder with residual blocks. Upsamples Stage-I result to high-resolution (256x256), correcting defects and adding details.

StackGAN-v2: Multi-Stage General Image Synthesis

Core Idea: A more general, end-to-end multi-stage framework with a tree-like structure.

Figure: StackGAN-v2 Architecture for Conditional Synthesis (Source: [1])

- **Tree-like Structure:**
  - Input: Noise ‘z‘ (unconditional) or ‘(z, c)‘ (conditional, ‘c‘ is e.g., text embedding).
  - Multiple generators (‘G_0, G_1, G_z‘) produce images at increasing scales (e.g., 64x64, 128x128, 256x256).
  - Each ‘G $_i$‘ has a corresponding discriminator ‘D $_i$‘.
- **Joint Multi-Distribution Approximation:**
  - Generators are jointly trained to approximate image distributions at multiple scales.
  - For conditional tasks, discriminators ‘D $_i$‘ have both unconditional (real vs fake image) and conditional (image-condition match vs mismatch) loss terms.
- **Color-Consistency Regularization:**

Table of Contents

1. Introduction
2. StackGAN++ Methodology
3. Results
4. Ablation Studies
5. Analysis

Experimental Setup

Datasets:

- *Text-to-Image (Conditional):*
  - CUB-200-2011 (Birds)
  - Oxford-102 (Flowers)
  - MS COCO (Challenging general scenes)
- *Unconditional Generation:*
  - LSUN (Bedroom, Church)
  - ImageNet (Dog, Cat subsets)

Figure: Statistics of Datasets (Source: [1])

Evaluation Metrics:

- **Inception Score (IS):** Measures image quality and diversity. Higher is better.
- **Fréchet Inception Distance (FID):** Measures similarity between generated and real image distributions. Lower is better.
- **Human Rank (HR):** User studies to assess perceptual quality and text-image alignment. Lower is better.
- **t-SNE Visualizations:** To check for mode collapse and sample diversity.

Quantitative Results: Text-to-Image (StackGAN-v1)

StackGAN-v1 significantly outperforms prior text-to-image models.

Figure: Comparison of StackGAN-v1 with GAN-INT-CLS and GAWWN (Source: [1])

- **Higher Resolution:** StackGAN-v1 generates 256x256 images.
- **Improved IS:** e.g., on CUB, StackGAN-v1 (3.70) vs. GAN-INT-CLS (2.88).
- **Drastically Lower FID*:** FID* (on 64x64 resized images) shows better distribution matching. e.g., on CUB, StackGAN-v1 (35.11) vs. GAN-INT-CLS (68.79).
- **Better Human Rank (HR):** Indicates more realistic and text-relevant images.

Qualitative Results: Text-to-Image

CUB Dataset (Birds):

Figure: StackGANs vs. GAWWN vs. GAN-INT-CLS on CUB (Source: [1])

Oxford-102 (Flowers) & COCO:

Figure: StackGANs vs. GAN-INT-CLS on Oxford-102 and COCO (Source: [1])

StackGAN-v1 and v2 produce much more detailed and realistic images compared to GAN-INT-CLS (64x64) and GAWWN (128x128, often blurry without part annotations).

StackGAN-v2 vs. StackGAN-v1 & Unconditional SOTA

Figure: Comparison of StackGAN-v1 and StackGAN-v2 (Source: [1])

- **StackGAN-v2 often improves FID over StackGAN-v1**, e.g., CUB FID: 15.30 (v2) vs 51.89 (v1).
- **StackGAN-v2 IS generally higher or competitive.**
- **Less Mode Collapse in StackGAN-v2:**

Figure: t-SNE: StackGAN-v1 (a) has collapsed modes, StackGAN-v2 (b) does not. (Source: [1])

- **Unconditional Generation:** StackGAN-v2 outperforms SOTA like DCGAN, WGAN-GP in quality and resolution (256x256).

Figure: Unconditional generation on LSUN Bedroom by various GANs and StackGANs. (Source: [1])

Table of Contents

1 Introduction

2 StackGAN++ Methodology

3 Results

4 Ablation Studies

5 Analysis

Ablation Studies: StackGAN-v1 Components

Testing importance of StackGAN-v1 design choices on CUB dataset (Table 4 in paper).

- **Necessity of Stacked Structure:**
  - Stage-I GAN direct 256x256 output: Poor IS (3.02) vs. StackGAN-v1 (3.70). Visually much worse (Fig. 11 in paper).
- **Effect of Conditioning Augmentation (CA):**
  - Stage-I GAN (64x64) IS: 2.66 (no CA) → 2.95 (with CA).
  - Without CA, 256x256 Stage-I GAN collapses (Fig. 11 in paper). CA stabilizes and improves diversity.
- **Inputting Text at Both Stages ("Text twice"):**
  - StackGAN-v1 256x256 IS: 3.45 (text only Stage-I) → 3.70 (text at both stages).
  - Stage-II benefits from re-processing text.

Figure: Stage-I (rough sketch) vs. Stage-II (refined details) in StackGAN-v1. (Source: [1])

Ablation Studies: StackGAN-v2 Components

Testing importance of StackGAN-v2 design choices on CUB (Table 5) and other datasets.

- **Multi-Scale/Multi-Stage Architecture:**
    - ‘StackGAN-v2-G3‘ (only final 256x256 generator): IS drops from 4.04 → 3.49.
    - ‘StackGAN-v2-all256‘ (all generators output 256x256): IS drops to 2.89.
    - Visuals (Fig. 14 in paper) show severe mode collapse or poor quality for these baselines.
- **Joint Conditional/Unconditional (JCU) Discriminators:**
    - ‘StackGAN-v2-no-JCU‘ (conventional conditional D): IS drops from 4.04 → 3.77.
- **Color-Consistency Regularization:**
    - Qualitatively (Fig. 15 in paper): Improves color consistency across scales for unconditional generation.
    - Quantitatively (ImageNet Dog): IS drops from 9.55 → 9.02 without it.
    - Not critical for text-to-image due to strong text conditioning.

Figure: (Bottom row) Visual comparison of StackGAN-v2 ablations on CUB.

Table of Contents

1. Introduction
2. StackGAN++ Methodology
3. Results
4. Ablation Studies
5. Analysis

Pros (Improvements over Competing Methods)

- **Achieves Higher Resolution (256x256) with Photo-Realism:**
  - StackGAN-v1 was pioneering in generating 256x256 images from text, a significant leap from previous 64x64 or 128x128 results.
- **Superior Image Quality and Diversity:**
  - Consistently better IS, FID, and human preference scores compared to prior text-to-image methods (GAN-INT-CLS, GAWWN).
  - StackGAN-v2 further improves stability and quality (especially FID) over StackGAN-v1 and SOTA unconditional GANs.
- **More Stable GAN Training:**
  - Conditioning Augmentation (CA) in StackGAN-v1 stabilizes conditional GANs.
  - StackGAN-v2's joint multi-distribution approximation and tree structure lead to more stable training and reduced mode collapse.

Figure: StackGAN-v2 (b) shows less mode collapse than v1 (a). (Source

- **General Framework (StackGAN-v2):**
  - Applicable to both conditional (text-to-image, class-conditional) and

Cons (Limitations of the Proposed Method)

- **Failure Cases Still Exist:**
  - While significantly improved, the methods can still produce imperfect images (e.g., blurry parts, unnatural shapes, minor artifacts), especially for complex text or scenes. StackGAN-v2 failures are generally "milder."

    Figure: Examples of failure cases from StackGAN-v1 (top) and StackGAN (bottom). (Source: [1])

- **Convergence on Complex Datasets (StackGAN-v2):**
  - StackGAN-v2's end-to-end joint training can be harder to converge on highly complex datasets (like COCO) compared to StackGAN-v1's simpler stage-by-stage optimization.
  - StackGAN-v1 sometimes yields slightly more appealing images on COCO by human rank, despite v2's better stability.
- **Computational Cost:**
  - Training multiple generators and discriminators in StackGAN-v2 is computationally intensive. StackGAN-v1, while two-stage, might have

Conclusions

- **Stacked/Multi-Stage GANs are Highly Effective:**
  - Decomposing high-resolution image synthesis into progressive, manageable sub-problems (low-to-high resolution) is a key strategy for success.
- **StackGAN-v1 Advanced Text-to-Image Synthesis:**
  - First to achieve 256x256 photo-realistic images from text, with Conditioning Augmentation (CA) improving stability and diversity.
- **StackGAN-v2 Offers Generality, Stability, and Quality:**
  - Its tree-like architecture, joint multi-distribution approximation, and color-consistency regularization lead to more stable training, reduced mode collapse, and often higher quality for both conditional and unconditional tasks.
- **Significant Progress in Realistic Image Generation:**
  - The paper demonstrates a substantial leap in GANs' capability to generate detailed, high-resolution images.
- **Future Directions:** Despite progress, achieving perfect realism, coherence for all inputs, and efficient training for extremely complex scenarios remain open challenges.

Q&A

Thank you for your attention!

Any questions?

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.