URL Coursework 2

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

Bruno Sánchez Gómez

May 31, 2025

Table of Contents

Introduction

Area of Research of the Paper

Generative Adversarial Networks (GANs) for Realistic Image Synthesis

- High-resolution images (256 $\times$ 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 $\rightarrow$ Poor gradients

Contributions by StackGAN++

1. **Conditioning Augmentation (CA):** Improve sample diversity by augmenting the image-text pairs.
2. **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

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:

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

StackGAN-v1

Core Idea

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

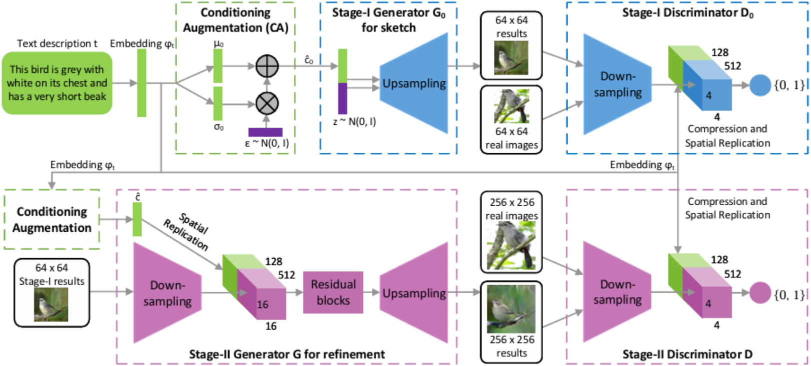

Figure: StackGAN-v1 Architecture (Source: [1])

StackGAN-v2

Core Idea

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

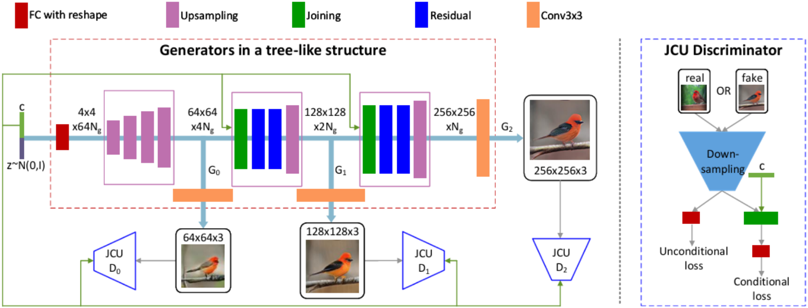

Figure: StackGAN-v2 Architecture and JCU Discriminators (Source: [1])

Table of Contents

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)

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.

- **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.

CUB Dataset (Birds):

Oxford-102 (Flowers) & COCO:

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

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

Table of Contents

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) $\rightarrow$ 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) $\rightarrow$ 3.70 (text at both stages).
  - Stage-II benefits from re-processing text.

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 $\rightarrow$ 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 $\rightarrow$ 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 $\rightarrow$ 9.02 without it.
  - Not critical for text-to-image due to strong text conditioning.

Table of Contents

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.
- **General Framework (StackGAN-v2):**
  - Applicable to both conditional (text-to-image, class-conditional) and unconditional image generation tasks.
- **Ability to Correct Defects:** The multi-stage approach allows later stages to refine and correct errors or omissions from earlier stages.

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."
- **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 lower peak memory.
- **Dependence on Text Embedding Quality:**
  - The quality of generated images is highly dependent on the quality of the input text embeddings from the pre-trained text encoder.
- **Subtle Mode Collapses:** While large-scale mode collapses are

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.

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.