StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks

Zhang, Xu, Li, Zhang, Wang, Huang & Metaxas

Presented by Michael O. Vertolli

Oct. 15, 2018

Outline

Motivation

- Latent variable models approximate complex distributions indirectly via notions of divergence.
- The challenge of real world distributions is they are *ugly*.
- Assumptions of divergence are often reflected in model errors.

Basic Notation and Task

- Let...
  - x be a real image from the data distribution p_{data}.
  - z be a noise vector sampled from the prior p_z.
  - G be a differentiable *generator* network, which samples from the decoding distribution $G(z) \sim p(x|z)$.
- Then we want to minimize...
  - The divergence $\mathcal{D}_x$ between p_{data} and $p(x|z)$.

Prototypical Example: Variational Autoencoders

- VAE indirectly model p_{data} by representing p_z as a family of Gaussians parameterized by mean and variance.
  - Include an encoding network Q that samples from $Q(x) \sim q(z|x)$.
- Introduce a new divergence $\mathcal{D}_z$ between p_z and $q(z|x)$.
- Simplify to regularized ($\mathcal{D}_z$) reconstruction loss ($\mathcal{D}_x$).

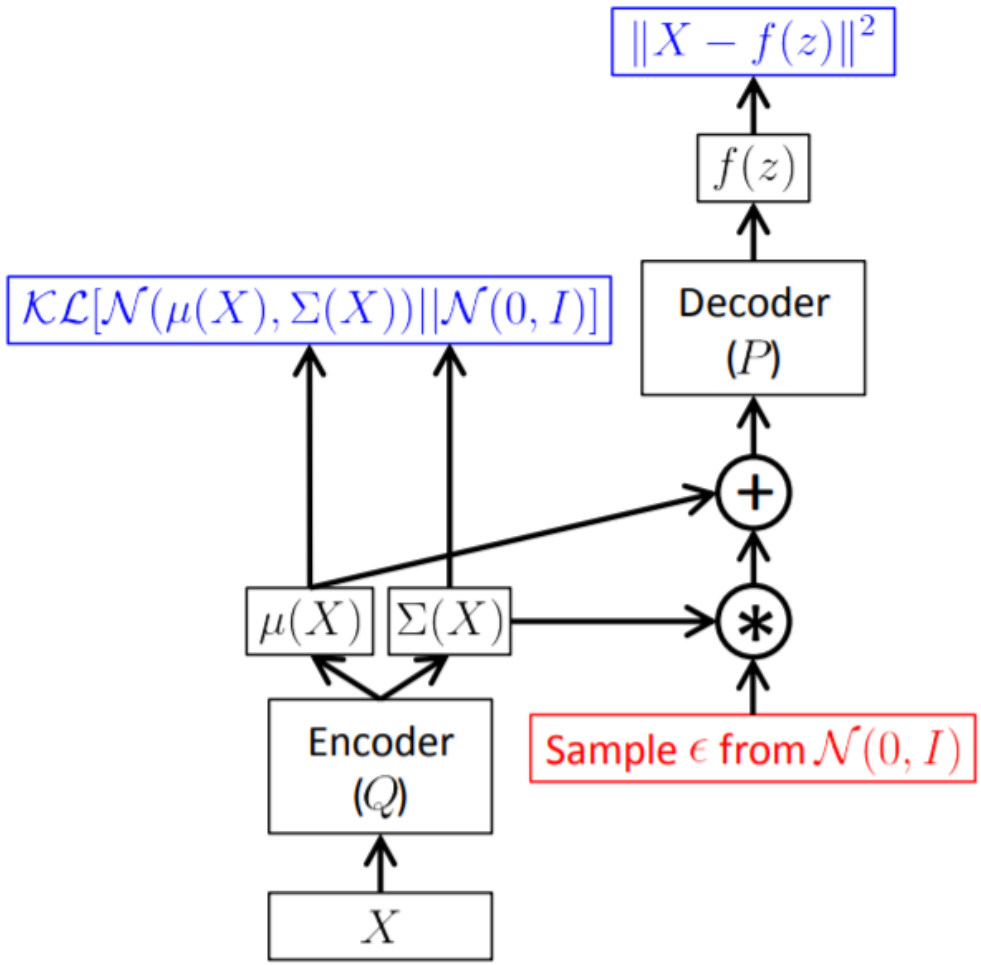


Figure 1: VAE diagram from Doersch (2016)

Figure 2: Blurry VAE output from Dosovitskiy & Brox (2016)

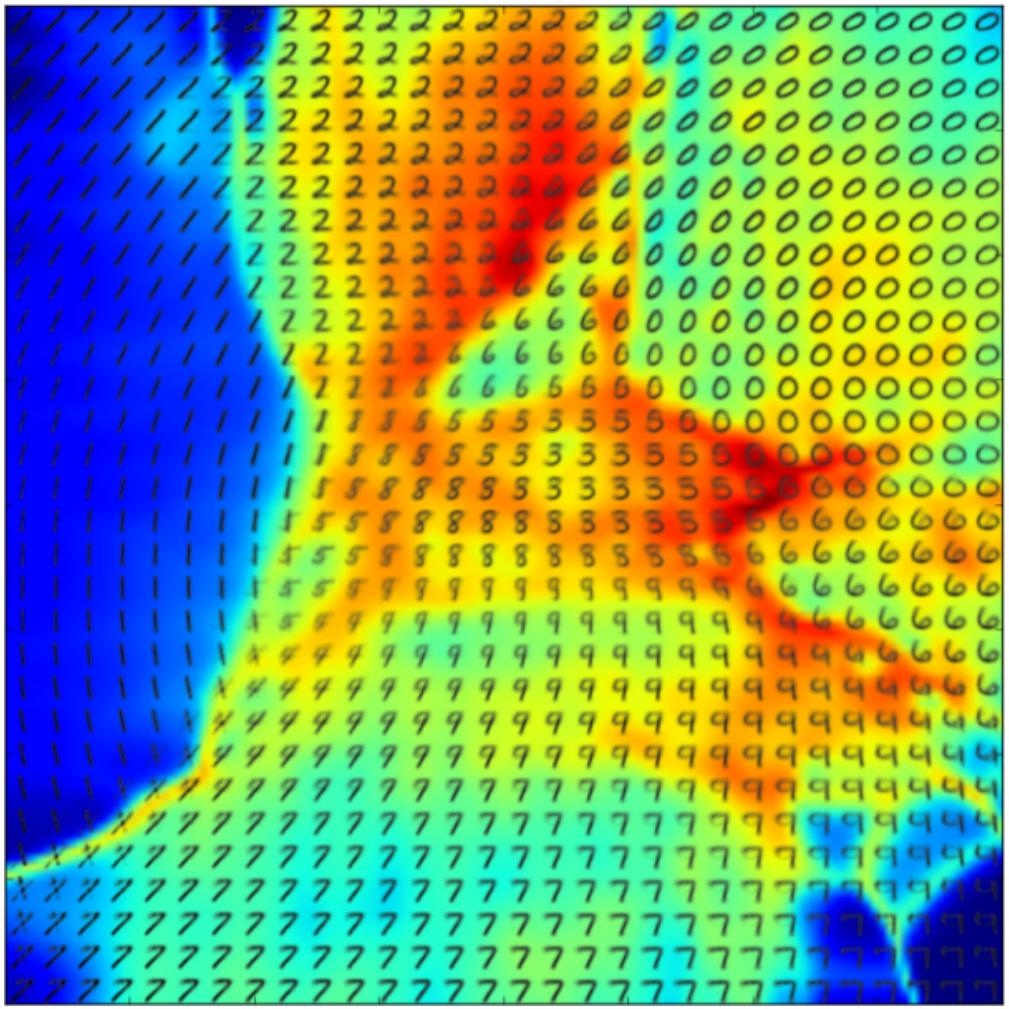


Figure 3: Output variance plotted on the latent space of VAE from Mescheder, et al. (2017)

Adversarial Autoencoders

- Adversarial AE generalize VAEs by mapping to arbitrary p_z.
  - Include a *discriminator* network D that computes the probability that z came from p_z.
- Use adversarial regularization for $\mathcal{D}_z$.

Adversarial Loss for Q

$$\min_Q \max_D V(D,Q) = \mathbb{E}_{z\sim p_z}[\log D(z)] + \mathbb{E}_{x\sim p_{data}}[\log (1 - D(Q(x)))]$$

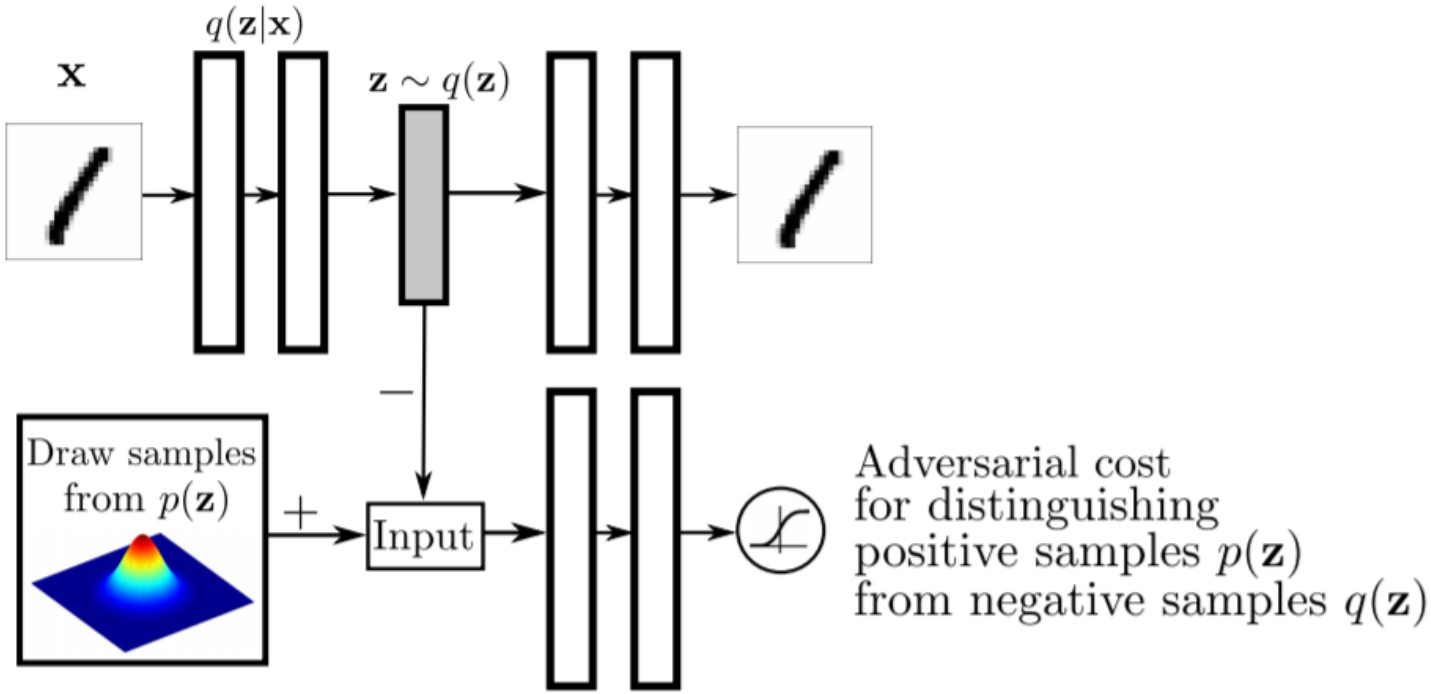


Figure 4: Adversarial AE diagram from Makhzani, et al. (2016)

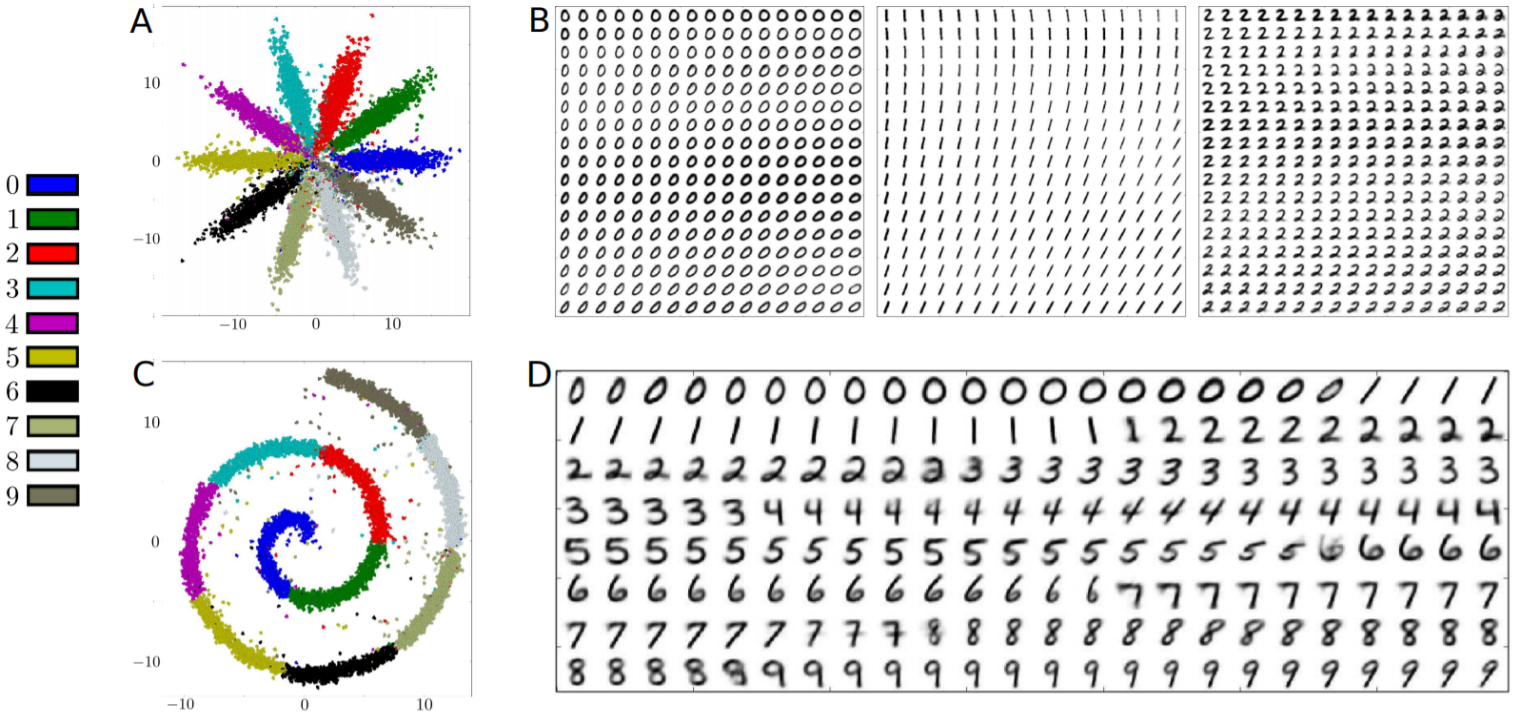


Figure 5: Adversarial AE latent representations from Makhzani, et al. (2016)

Generative Adversarial Networks

- Generative Adversarial Networks (GANs) are Adversarial Autoencoders that model $\mathcal{D}_x$ directly.
  - D now computes the probability that x came from p_{data}.
- The divergence $\mathcal{D}_z$ from p_z is now implicit as z is sampled directly.

Adversarial Loss for G

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_d}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log (1 - D(G(z)))]$$

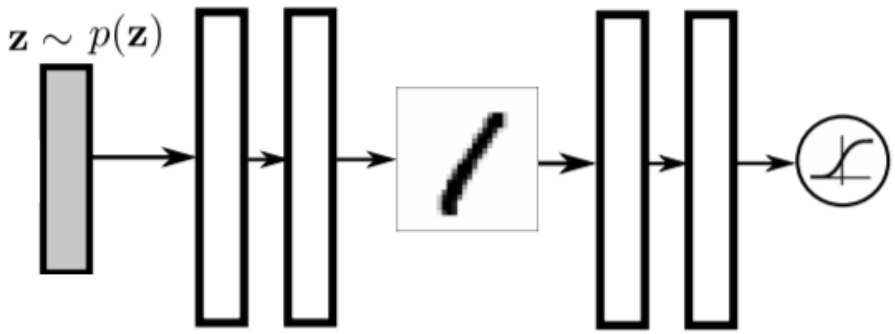


Figure 6: GAN diagram

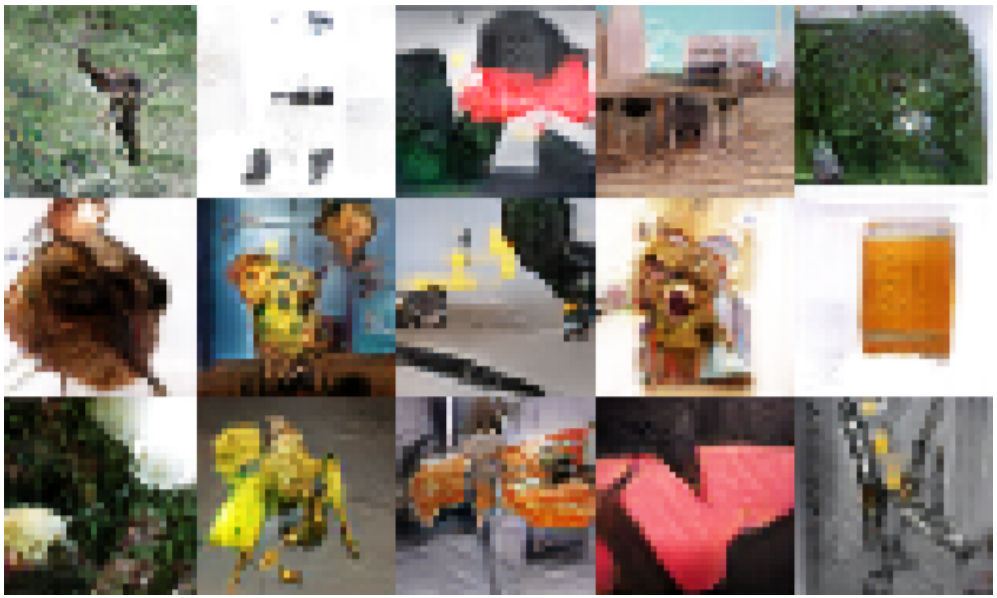


Figure 7: DCGAN output from Radford et al., (2015)

Recent Developments

- Conditional GANs
- LAPGAN
- GAN-INT-CLS
- GAWWN
- StackGAN-v1
- Progressive GAN

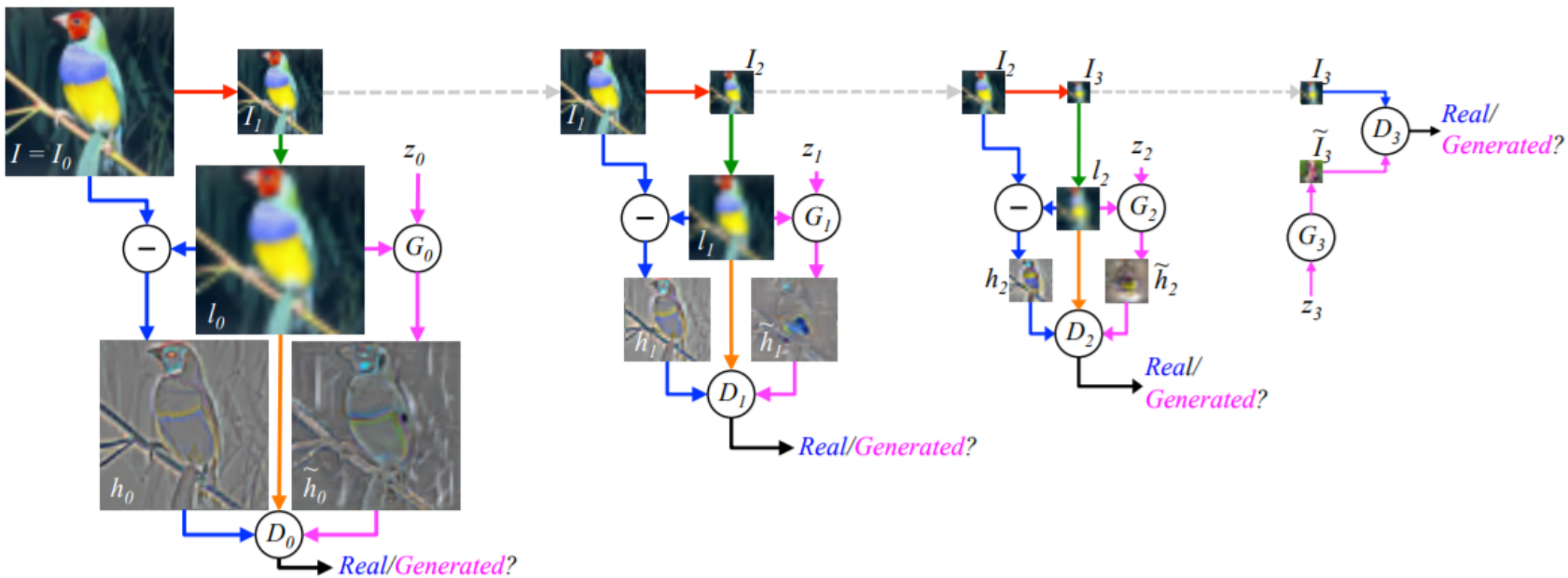


Figure 8: LAPGAN training procedure from Denton et al., (2015)

Recent Developments

- Conditional GANs
- LAPGAN
- GAN-INT-CLS
- GAWWN
- StackGAN-v1
- Progressive GAN

StackGAN-V1

- StackGAN-V1 breaks task into three sub-tasks.
  - Conditionally generate small images from text.
  - Conditionally generate big images from small images plus text.
  - Learn a VAE-like "augmentation" of text with G.

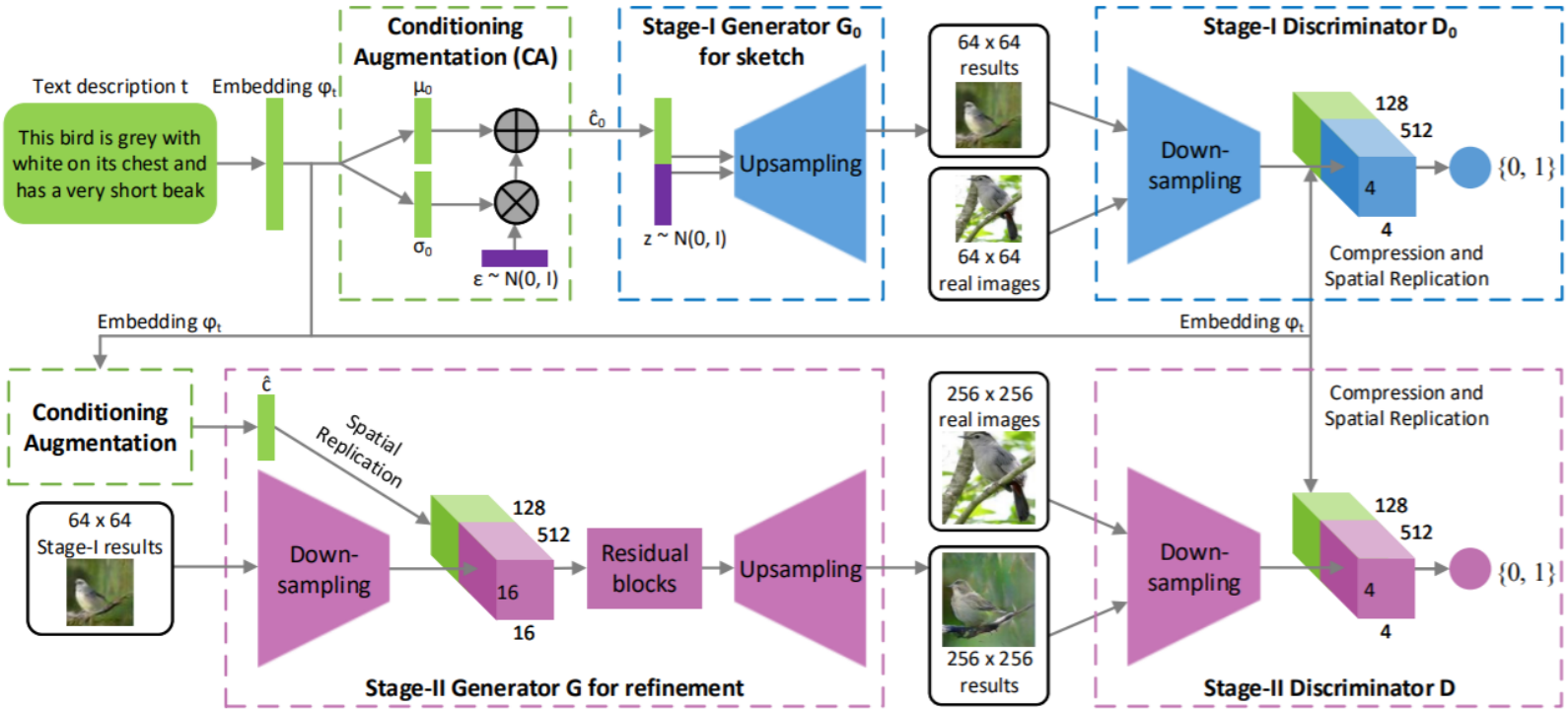


Figure 9: StackGAN-V1 network diagram

StackGAN-V2

- Adds multiple G_i-D_{s_i} pairs.
- Trains all G_i jointly.
  - $\mathcal{D}_x = \sum_{i=1}^{m} \mathcal{D}_{s_i}$, where $s_i = G_i(h_i)$ and $h_i = F_i(h_{i-1}, z)$ for $i > 0$ and $h_0 = F(z)$
- Conditioning occurs at multiple levels.
- Adds a "color-consistency" regularization for the generator.

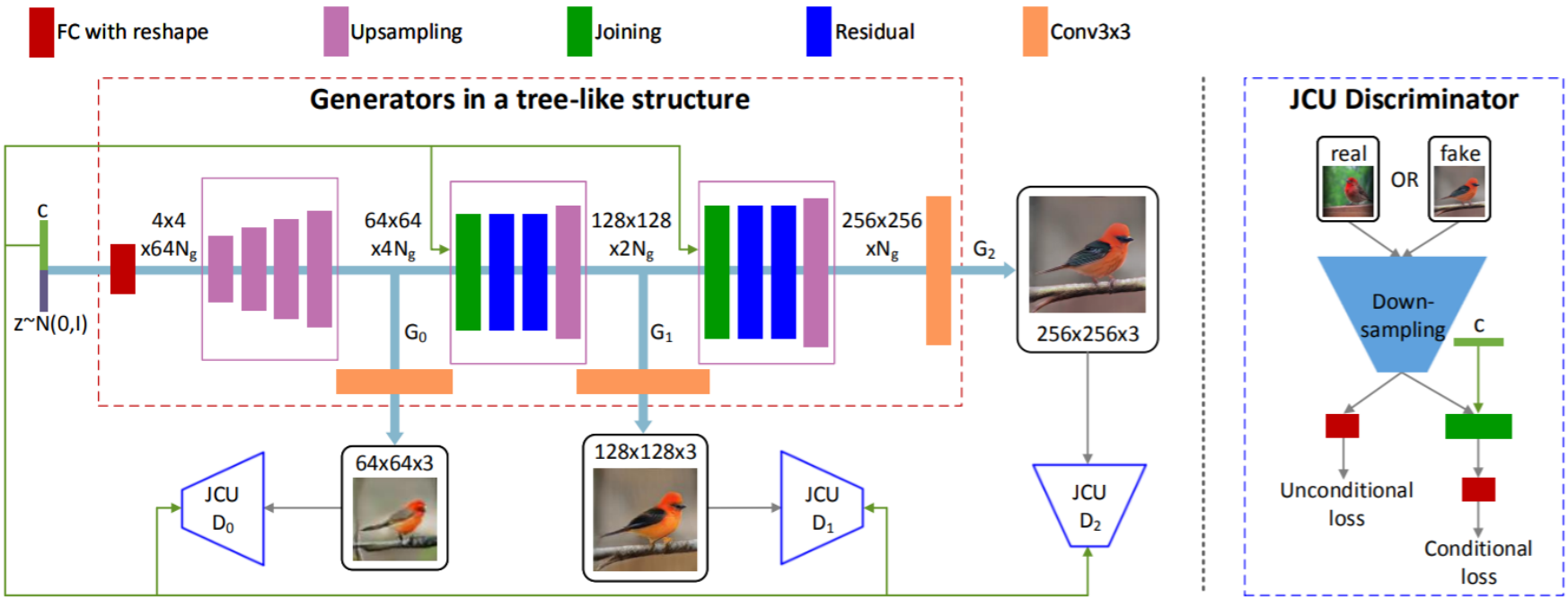


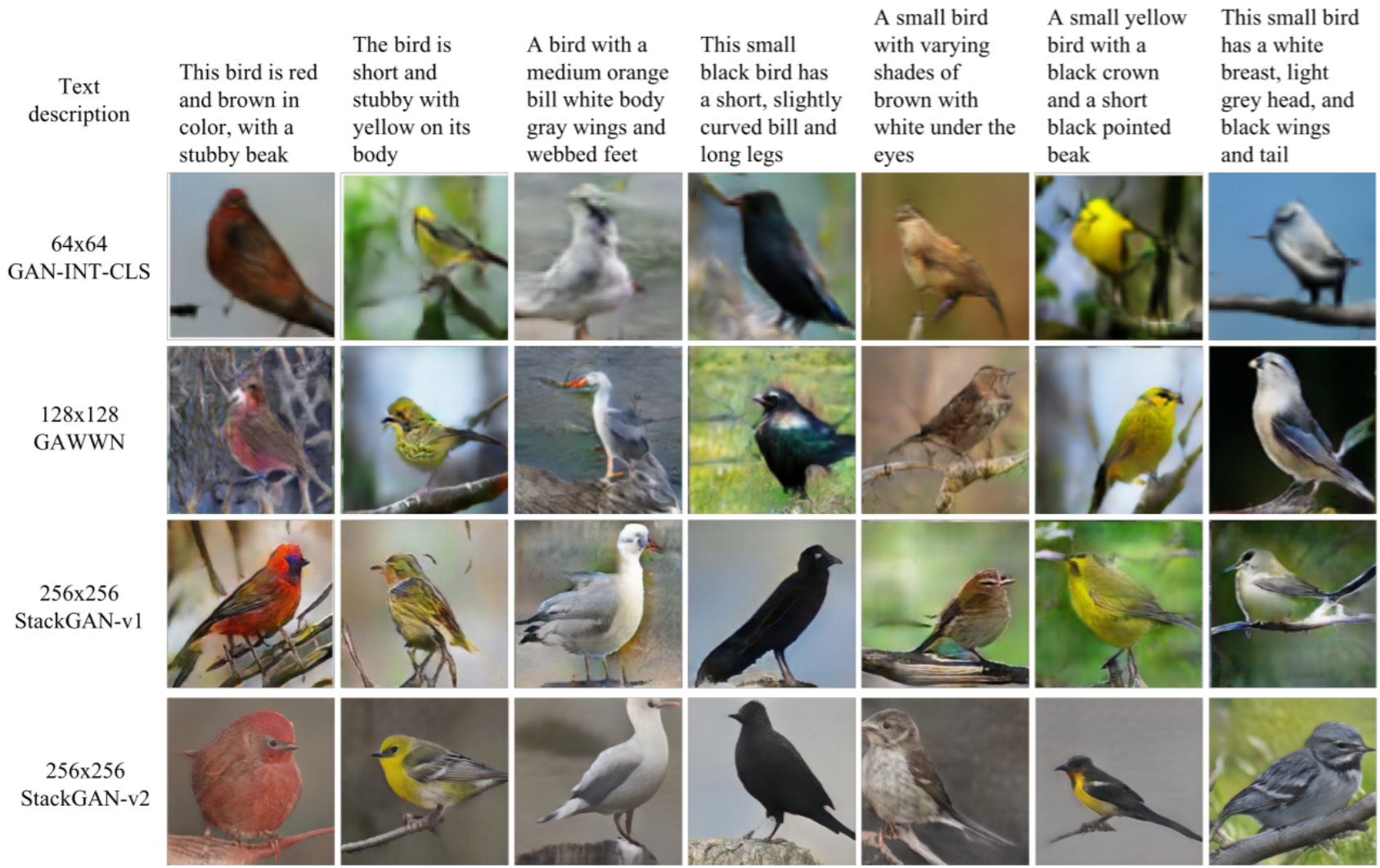
Figure 10: StackGAN-V2 network diagram

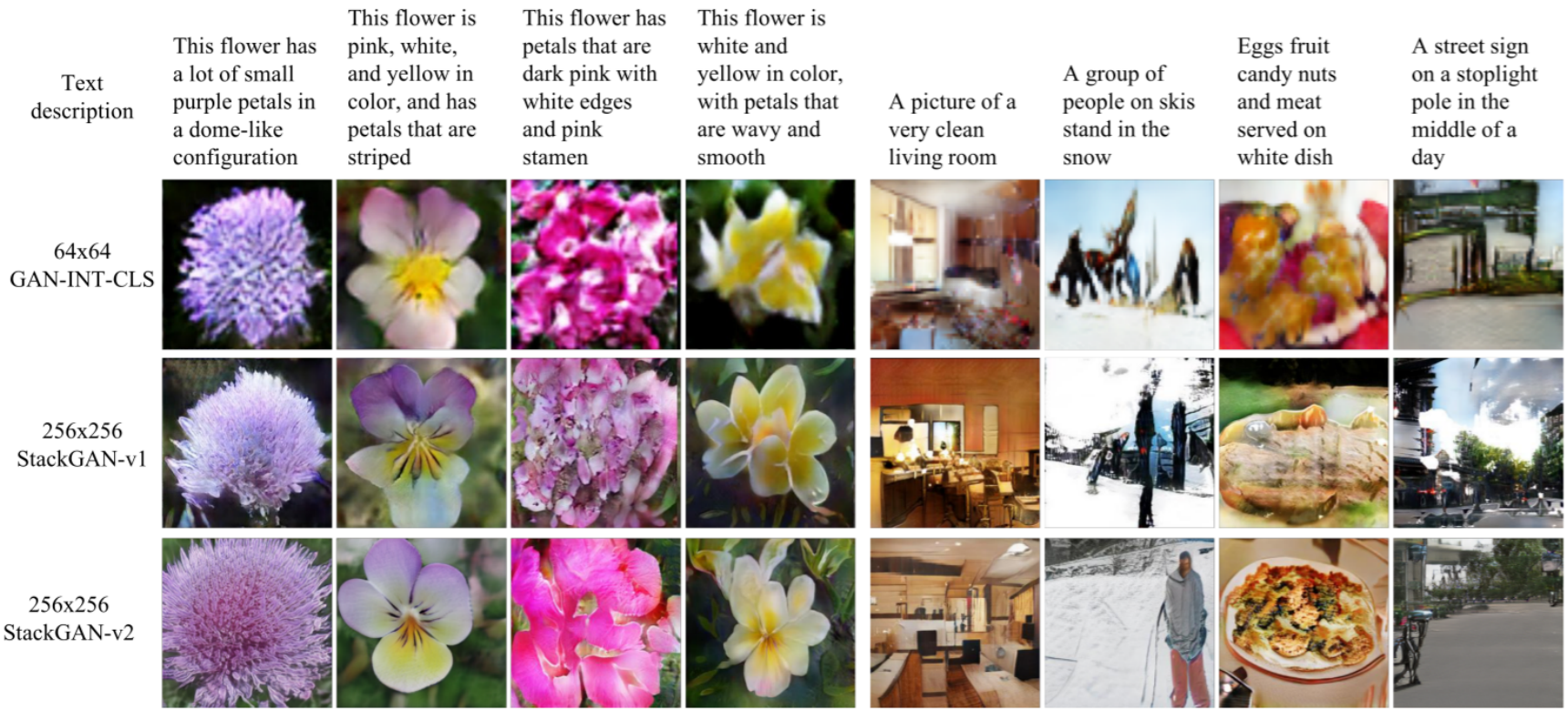
Experiments

- Datasets
  - CUB, Oxford-102, and COCO for text-to-image.
  - LSUN(bedroom, church) and ImageNet(dog, cat) for unconditional.
- Evaluations
  - Inception Score: compares the divergence between $p(h|G(z))$ and marginal $\int p(h|G(z))dz$.
  - Fréchet Distance: compares the Wasserstein-2 distance between $p(h|x)$ and $p(h|G(z))$.
  - Scaling adjustments for IS and FID.
  - Raters ranked results.
  - MS-SSIM comparison of output images.
  - 2D t-SNE projection.

Results of Text-to-Image Comparisons

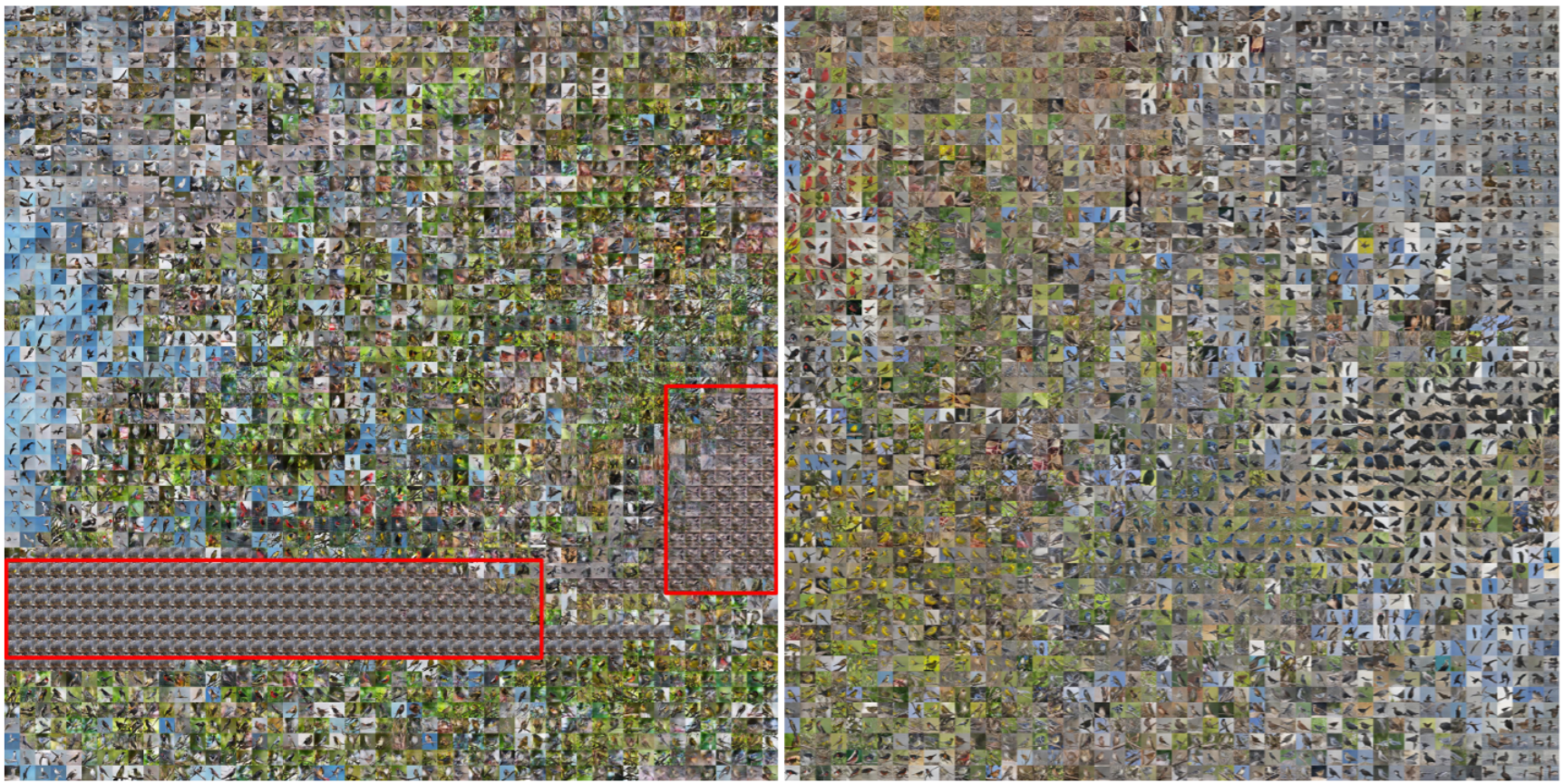
- StackGAN-V1 outperforms on everything.
- StackGAN-V2 outperforms V1 on FID and ratings for all except COCO.
  - Mixed results for IS.
  - Differences are small.
- StackGAN-V1 produces moderate/severe mode collapse for all except Oxford-102.

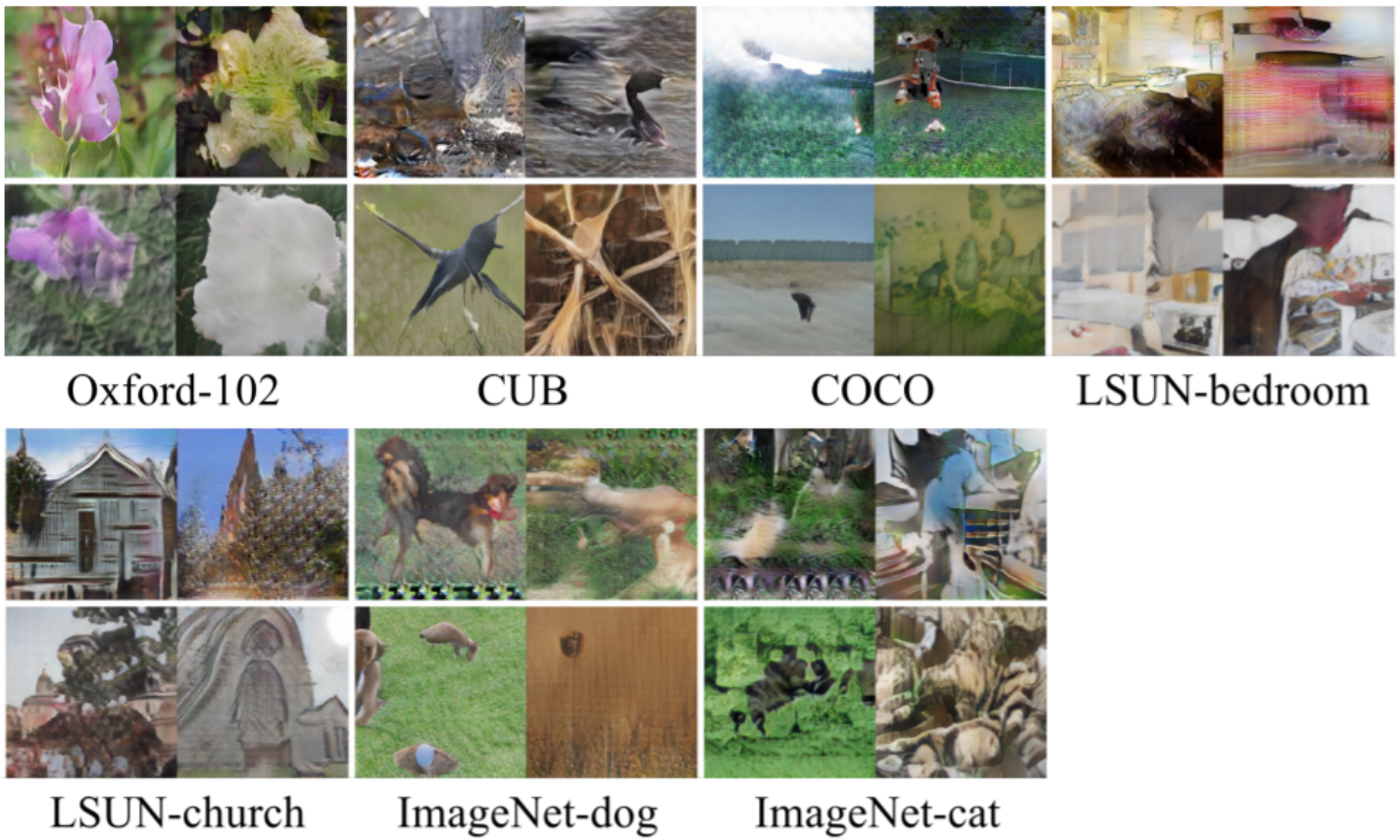

Text description
This bird is red and brown in color, with a stubby beak
The bird is short and stubby with yellow on its body
A bird with a medium orange bill white body gray wings and webbed feet
This small black bird has a short, slightly curved bill and long legs
A small bird with varying shades of brown with white under the eyes
A small yellow bird with a black crown and a short black pointed beak
This small bird has a white breast, light grey head, and black wings and tail
64x64 GAN-INT-CLS
128x128 GAWWN
256x256 StackGAN-v1
256x256 StackGAN-v2


Text description
This flower has a lot of small purple petals in a dome-like configuration
This flower is pink, white, and yellow in color, and has petals that are striped
This flower has petals that are dark pink with white edges and pink stamen
This flower is white and yellow in color, with petals that are wavy and smooth
A picture of a very clean living room
A group of people on skis stand in the snow
Eggs fruit candy nuts and meat served on white dish
A street sign on a stoplight pole in the middle of a day
64x64 GAN-INT-CLS
256x256 StackGAN-v1
256x256 StackGAN-v2

Results of Text-to-Image Comparisons

- StackGAN-V1 outperforms on everything.
- StackGAN-V2 outperforms V1 on FID and ratings for all except COCO.
  - Mixed results for IS.
  - Differences are small.
- StackGAN-V1 produces moderate/severe mode collapse for all except Oxford-102.




Oxford-102
CUB
COCO
LSUN-bedroom
LSUN-church
ImageNet-dog
ImageNet-cat

Other Results

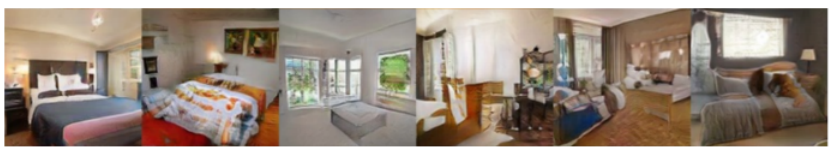
- StackGAN-V2 qualitatively outperforms other GANs on LSUN bedrooms.
- Outperforms DCGAN IS on ImageNet Dog dataset.
- Branches and conditional augmentation improve the IS score of StackGAN-V2.

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

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

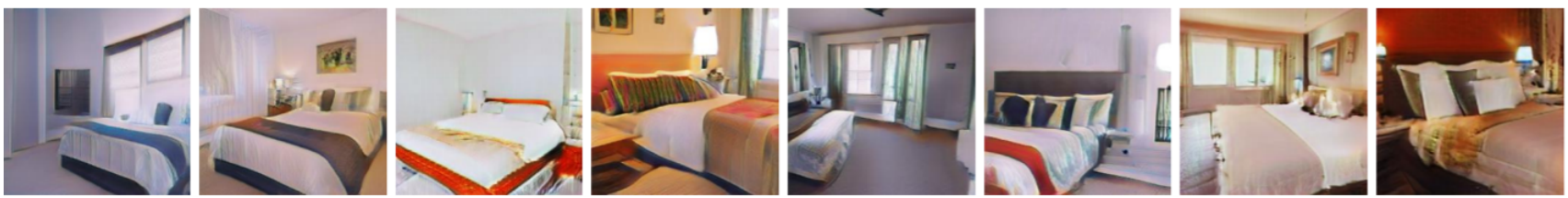
64×64 samples by EBGAN-PT (Reported in [56])

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



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

256×256 samples by our StackGAN-v1

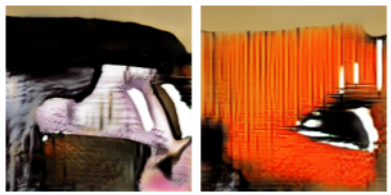


256×256 samples by our StackGAN-v2

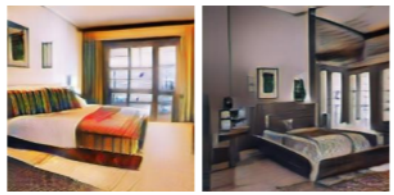
Other Results

- StackGAN-V2 qualitatively outperforms other GANs on LSUN bedrooms.
- Outperforms DCGAN IS on ImageNet Dog dataset.
- Branches and conditional augmentation improve the IS score of StackGAN-V2.


(a) StackGAN-v2-all256


(b) StackGAN-v2-G_3


(c) StackGAN-v2-3G_3


(d) StackGAN-v2

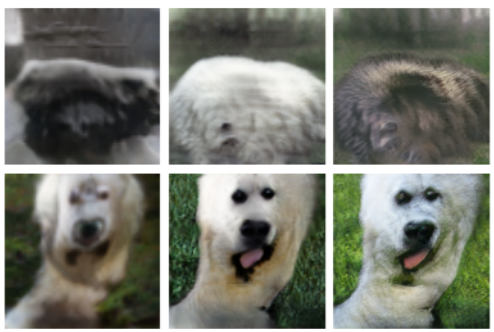
This black and white and grey bird has a black bandit marking around it's eyes


(e) StackGAN-v2-all256


(f) StackGAN-v2-G_3


(g) StackGAN-v2-no-JCU


(h) StackGAN-v2


ImageNet dog [39]


ImageNet cat [39]


LSUN church [54]

Discussion Points

- There are critical dependencies across image scales.
  - E.g., frequency, conditional, color-consistency.
- There is a need to examine compositional divergence criteria.
  - StackGAN-V2 D_x defined in terms of the contribution each embedding (s_i) makes to the partial divergences of each subsequent network ($\mathcal{D}_{s_j}$ s.t. $j \geq i$).
- Refinement-based latent variable modeling.

Further Reading

Amari, S. I., & Cichocki, A. (2010). Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences, 58*(1), 183-195.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine, 34*(4), 18-42.

Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. *In Advances in Neural Information Processing Systems*, (pp. 658-666).

McInnes, L., & Healy, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mescheder, L., Nowozin, S., & Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *International Conference on Machine Learning (ICML)* (pp. 2391-2400). PMLR.

Wakin, M. B., Donoho, D. L., Choi, H., & Baraniuk, R. G. (2005). The multiscale structure of non-differentiable image manifolds. In *Wavelets XI* 5914, (p. 59141B). International Society for Optics and Photonics.

Image-Text Embeddings

- Reed et al. (2016) created the Deep Symmetric Structured Joint Embedding (DS-SJE) using text-image pairs.
- Used a variety of network architectures based around character or word-level CNNs and LSTMs
- The objective is based off of a surrogate of the 0-1 loss, which is roughly the hinge loss with the hinge replaced by the 0-1 function.

Surrogate 0-1 Loss

$$\mathcal{L}_v(v_n, t_n, y_n) = \max_{y \in \mathcal{Y}}(0, \Delta(y_n, y) + \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v_n, t) - F(v_n, t_n)]$$
$$\mathcal{L}_t(v_n, t_n, y_n) = \max_{y \in \mathcal{Y}}(0, \Delta(y_n, y) + \mathbb{E}_{v \sim \mathcal{V}(y)}[F(v, t_n) - F(v_n, t_n)]$$

Equations for StackGAN-V2

Joint Generator Loss

$\mathcal{L}_G = \sum_{i=1}^{m} -\mathbb{E}_{s_i \sim p_{G_i}}[\log D_i(s_i)]$ $(-\mathbb{E}_{s_i \sim p_{G_i}}[\log D_i(s_i, c)])$
where $s_i = G_i(h_i)$ and $h_i = F_i(h_{i-1}, z)$ for $i > 0$ and $h_0 = F(z)$

Joint Discriminator Loss

$\mathcal{L}_{D_i} = -\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i)] - \mathbb{E}_{s_i \sim p_{G_i}}[\log (1 - D_i(s_i))]$
$(-\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i, c)] - \mathbb{E}_{s_i \sim p_{G_i}}[\log (1 - D_i(s_i, c))])$

Color-consistency Regularization

$\mathcal{L}_{C_i} = \frac{1}{n} \sum_{j=1}^{n} (\lambda_1 ||\mu_{s_i^j} - \mu_{s_{i-1}^j}||_2^2 + \lambda_2 ||\Sigma_{s_i^j} - \Sigma_{s_{i-1}^j}||_F^2)$