A Survey of Text-Guided Image Generation

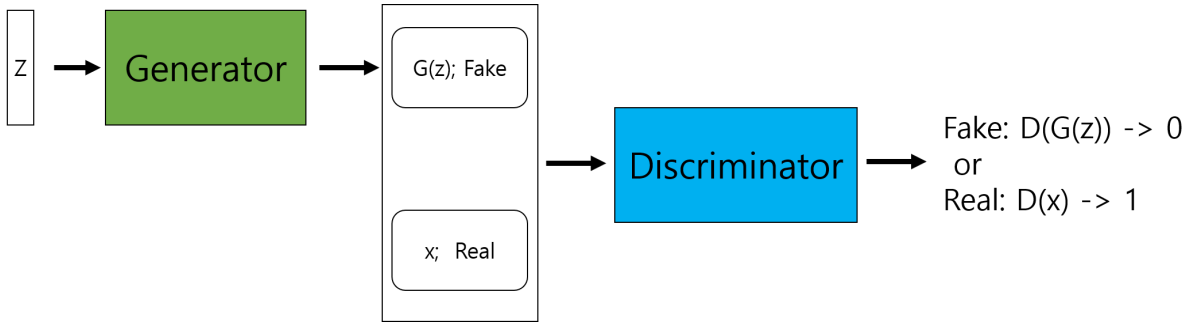
: from cGAN to ManiGAN

Woosung Choi

Contents

Preliminaries - GAN

- Generative Adversarial Network
  - Framework

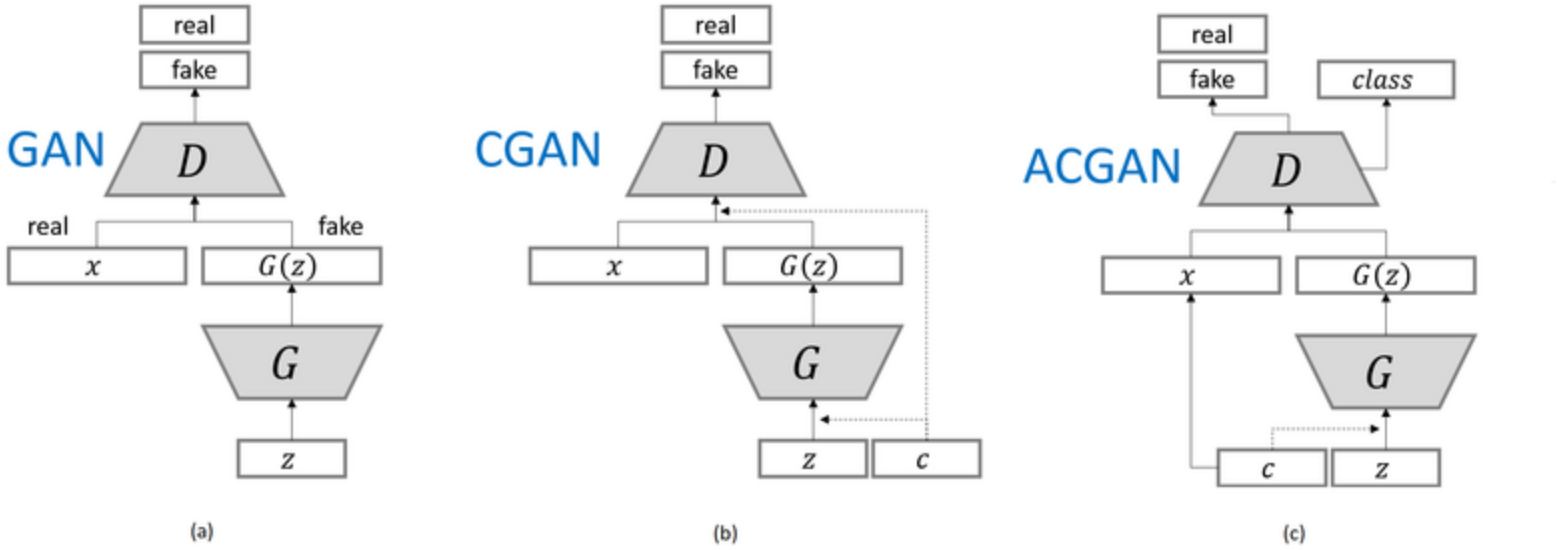


  - Loss Function

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Preliminaries - cGAN

- Conditional Generative Adversarial Network



- Mino, Ajkel, and Gerasimos Spanakis. "LoGAN: Generating logos with a generative adversarial neural network conditioned on color." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.

Text-to-Image Generation

- **[ICML 2016]** Reed, Scott, et al. "Generative Adversarial Text to Image Synthesis." International Conference on Machine Learning. 2016.
- **[StackGAN]** Zhang, Han, et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." ICCV. 2017.
- **[AttnGAN]** Xu, Tao, et al. "Attngan: Fine-grained text to image generation with attentional generative adversarial networks." CVPR. 2018.
- **[ControlGAN]** Li, Bowen, et al. "Controllable text-to-image generation." Advances in Neural Information Processing Systems. 2019.

Text-to-Image Generation: ICML 2016

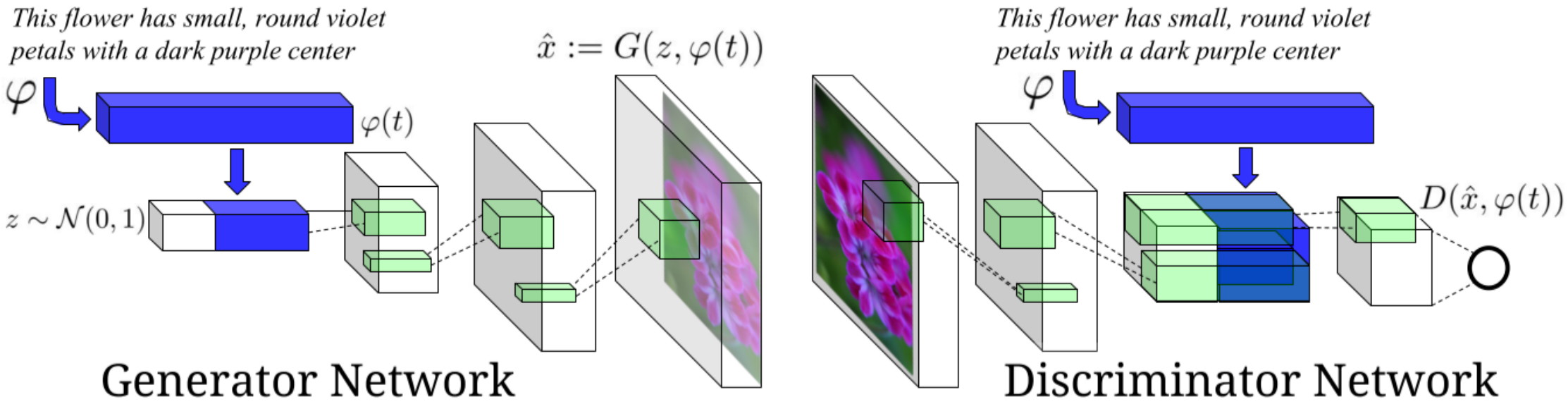


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

- conditioning by concatenatiion
  - φ: text encoder
  - $\varphi(t)$: embedding of the text description t

Text-to-Image Generation: StackGAN

- Motivation
  - [ICML 2016] can generate images that are highly related to the text, but it is very difficult to train GAN to generate ***high-resolution*** images from text
  - Simply adding more upsampling layers? : Empirically have failed
- Stacked Generative Adversarial Networks
  - State-I-GAN: sketches primitive shape and basic colors, ... (**coarse-grained**)
  - State-II-GAN: corrects defects, complete details (**fine-grained**)
- ***Conditioning Augmentation***
  - to stabilize conditional GAN training, and also improves the diversity of the generated samples

Text-to-Image Generation: StackGAN - Overview

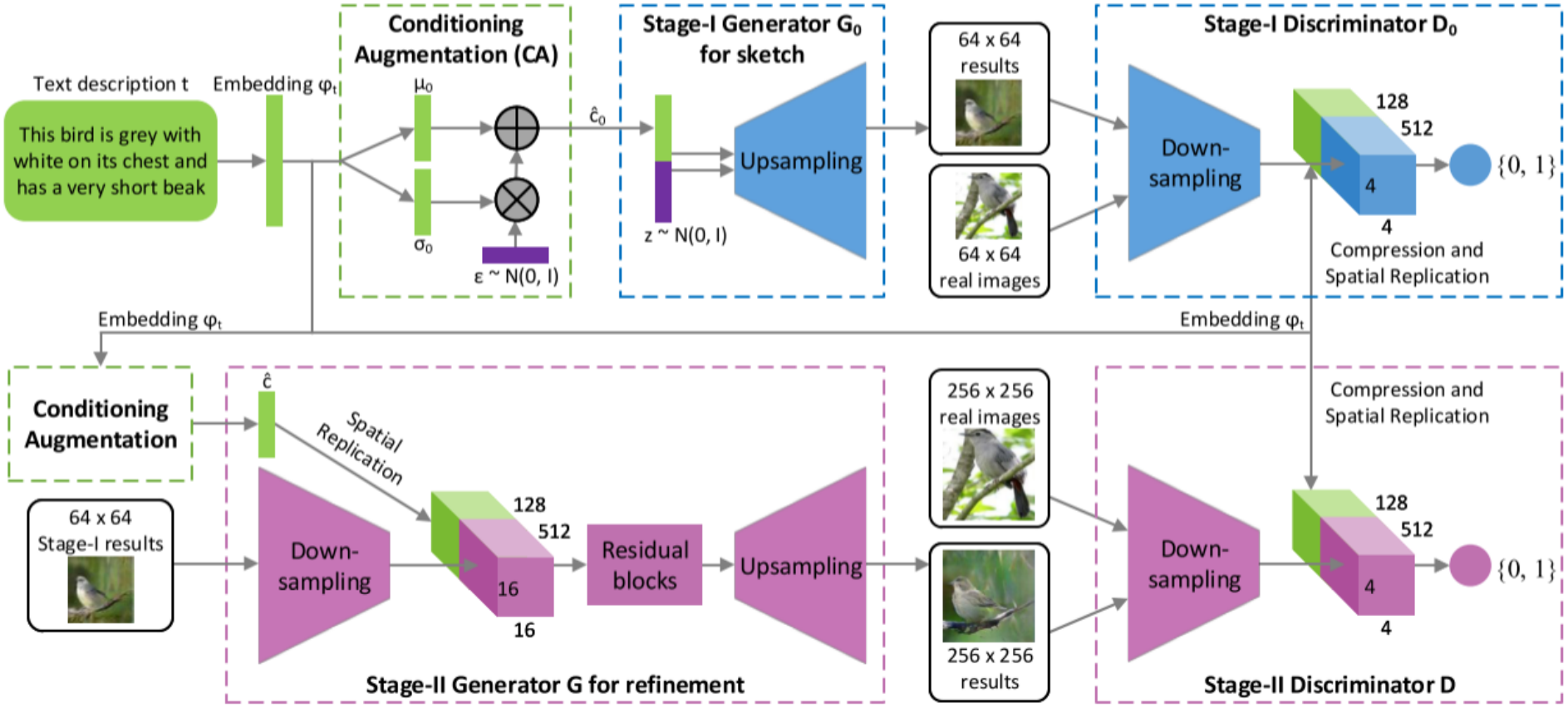


Figure 2. The architecture of the proposed StackGAN. The Stage-I generator draws a low-resolution image by sketching rough shape and basic colors of the object from the given text and painting the background from a random noise vector. Conditioned on Stage-I results, the Stage-II generator corrects defects and adds compelling details into Stage-I results, yielding a more realistic high-resolution image.

Text-to-Image Generation: StackGAN - CA

Conditioning Augmentation

- the text embedding is nonlinearly transformed to generate conditioning latent variables in [ICML 2016]
- However, latent space for the text embedding is usually high dimensional
- With limited amount of data, it usually causes discontinuity in the latent data manifold, which is not desirable
- To avoid overfitting, we add the regularization term to the objective function

$$D_{KL}(\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t)) \,||\, \mathcal{N}(0, I))$$

Text-to-Image Generation: StackGAN - Ablation Study for CA

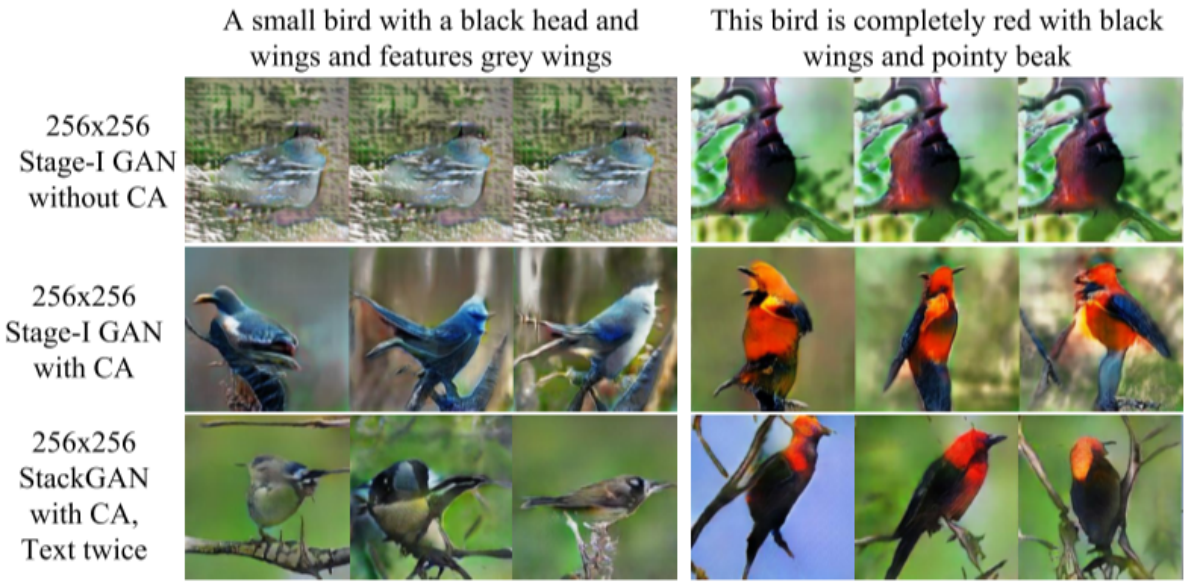


Figure 7. Conditioning Augmentation (CA) helps stabilize the training of conditional GAN and improves the diversity of the generated samples. (Row 1) without CA, Stage-I GAN fails to generate plausible 256×256 samples. Although different noise vector z is used for each column, the generated samples collapse to be the same for each input text description. (Row 2-3) with CA but fixing the noise vectors z, methods are still able to generate birds with different poses and viewpoints.

Text-to-Image Generation: AttnGAN

Motivation

- Conditioning GAN only on the global sentence vector lacks important fine-grained information at the word level and prevents the generation of high-quality images
- This problem becomes even more severe when generating complex scenes

AttnGAN

- To address this issue, AttnGAN allows attention-driven, multi-stage refinement for fine-grained text-to-image generation

Text-to-Image Generation: AttnGAN - Overview

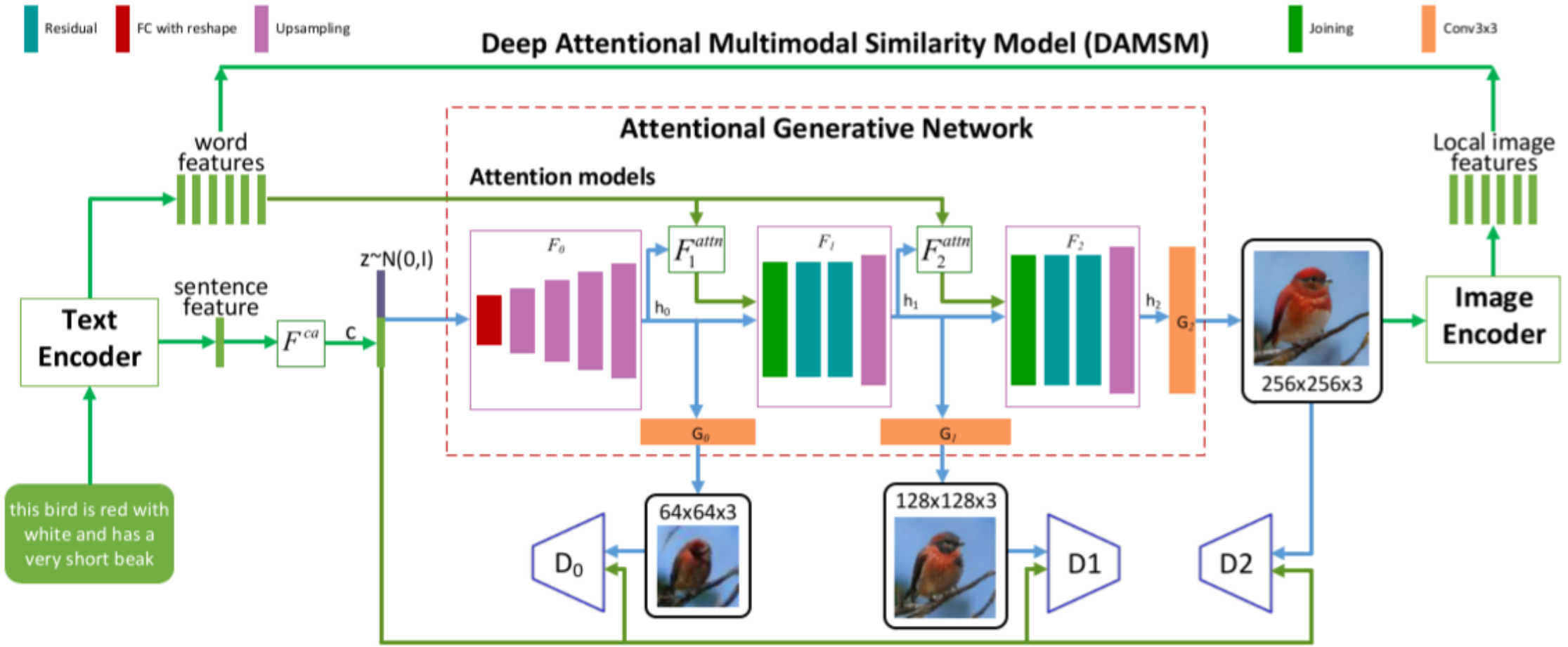


Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

Text-to-Image Generation: AttnGAN vs StackGAN

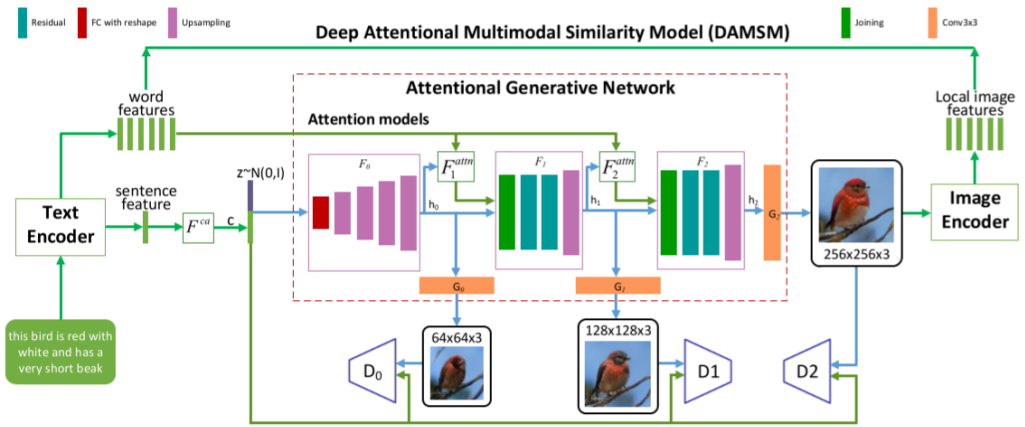


Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

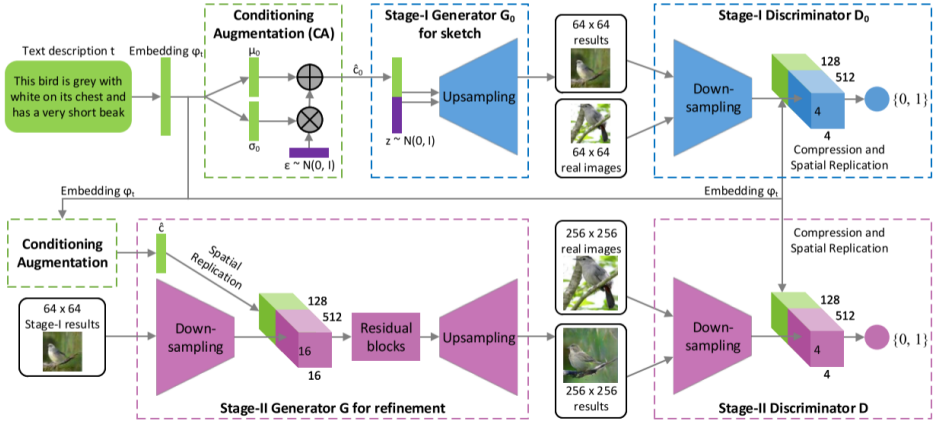


Figure 2. The architecture of the proposed StackGAN. The Stage-I generator draws a low-resolution image by sketching rough shape and basic colors of the object from the given text and painting the background from a random noise vector. Conditioned on Stage-I results, the Stage-II generator corrects defects and adds compelling details into Stage-I results, yielding a more realistic high-resolution image.

Text-to-Image Generation: AttnGan - Attention

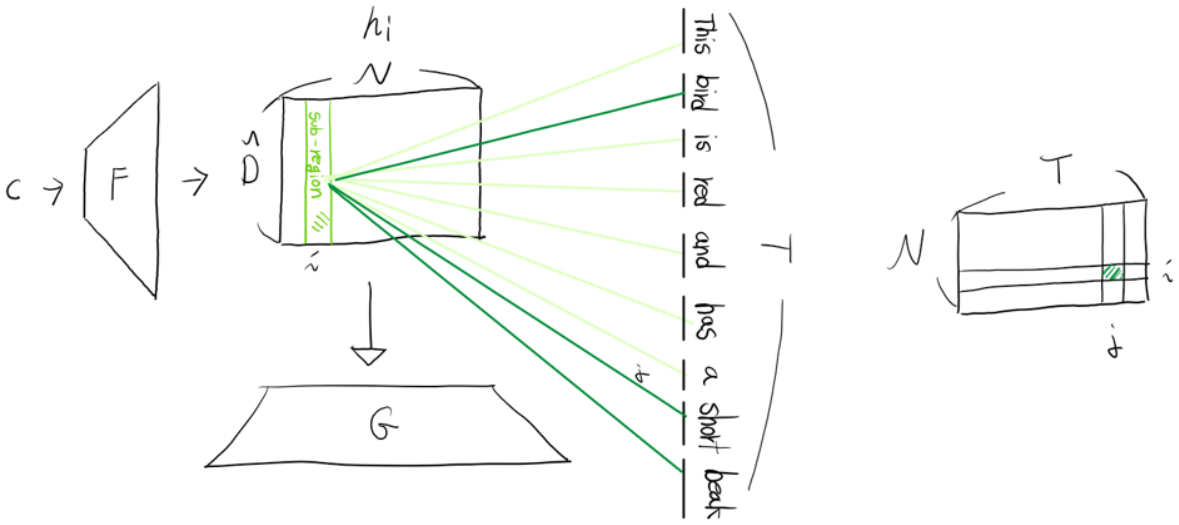


Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0, $G1$ and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Here, images of G_0 and G_1 are bilinearly upsampled to have the same size as that of G_2 for better visualization.

Advanced

- Controllable Text-to-Image Generation
    - **[ControlGAN]** Li, Bowen, et al. "Controllable text-to-image generation." Advances in Neural Information Processing Systems. 2019.
- Text-Guided Image Manipulation
    - **[ManiGAN]** Li, Bowen, et al. "Manigan: Text-guided image manipulation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- Bert-like System
    - **[X-LXMERT]** Cho, Jaemin, Jiasen Lu, D. Schwenk, Hannaneh Hajishirzi and Aniruddha Kembhavi. “X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers.”, EMNLP 2020

Advanced 1: ControlGAN

- Input: a sentence S
- Output: a realistic image I' that semantically aligns with S
- Constraints: we want to make this generation process ***controllable***
  - if S is modified to be S_m,
  - the synthetic result $\tilde{I}'$ should semantically match S_m while preserving irrelevant content existing in I'

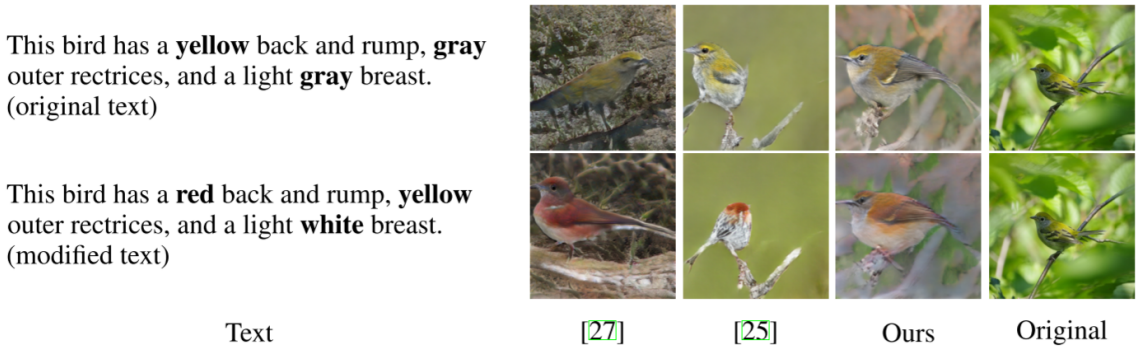


Figure 1: Examples of modifying synthetic images using a natural language description. The current state of the art methods generate realistic images, but fail to generate plausible images when we slightly change the text. In contrast, our method allows parts of the image to be manipulated in correspondence to the modified text description while preserving other unrelated content.

Advanced 1: ManiGAN

- Input: an input image I and a text description S'
- Output: a realistic image I' that is semantically aligned with S
- Constraints: preserving text-irrelevant contents existing in I

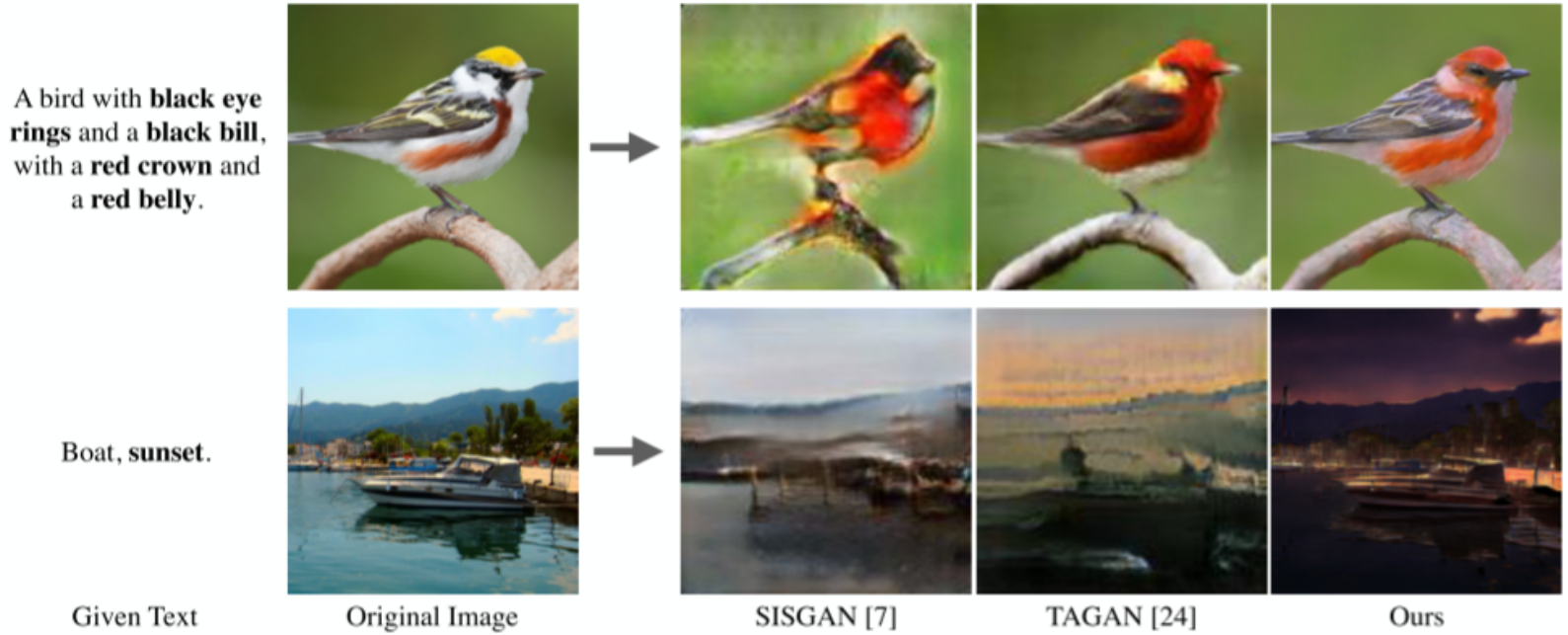


Figure 1: Given an original image that needs to be edited and a text provided by a user describing desired attributes, the goal is to edit parts of the image according to the given text while preserving text-irrelevant contents. Current state-of-the-art methods only generate low-quality images, and fail to do manipulation on COCO. In contrast, our method allows the original image to be manipulated accurately to match the given description, and also reconstructs text-irrelevant contents.