Microsoft

Attention GAN for Fine-Grained Language-to-Image Generation

Pengchuan Zhang
Researcher
MSR AI

GTC 2018, San Jose, CA

Joint work with Tao Xu, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He

Language to Image Generation

"Generate a bird with **wings that are blue** and a **red belly**"

"Generate a bird with **wings that are black** and a **white belly**"

"Generate a bird with **wings that are red** and a **yellow belly**"

ARTIFICIAL IMAGINATION

Language-to-Image generation with GANs

this small bird has a pink breast and crown, and black primaries and secondaries.

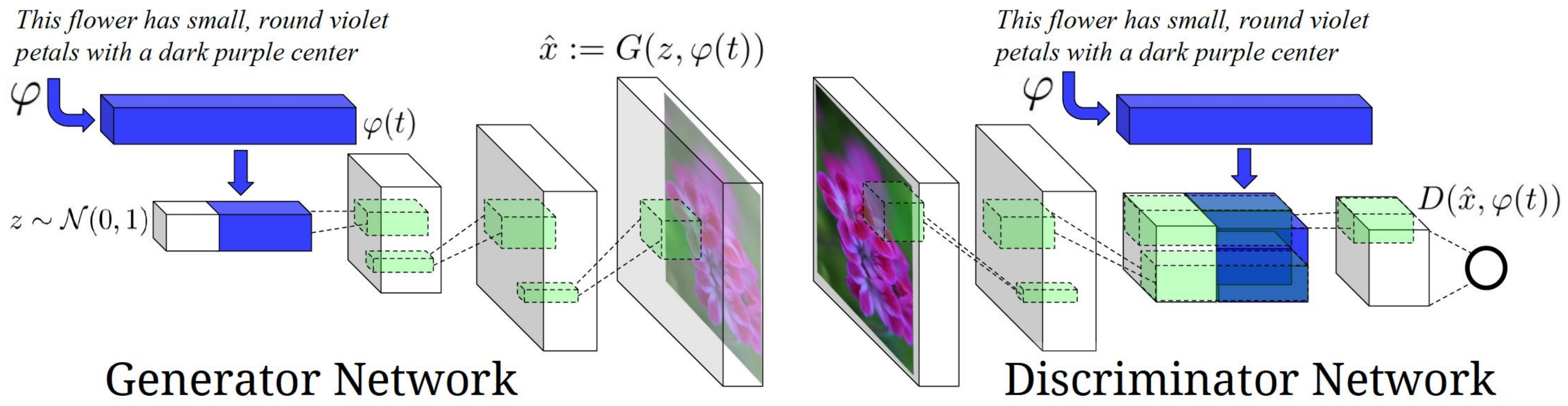


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

Objective function:

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

[Reed et al., Generative adversarial text-to-image synthesis, ICML, 2016]

Language-to-Image generation with GANs

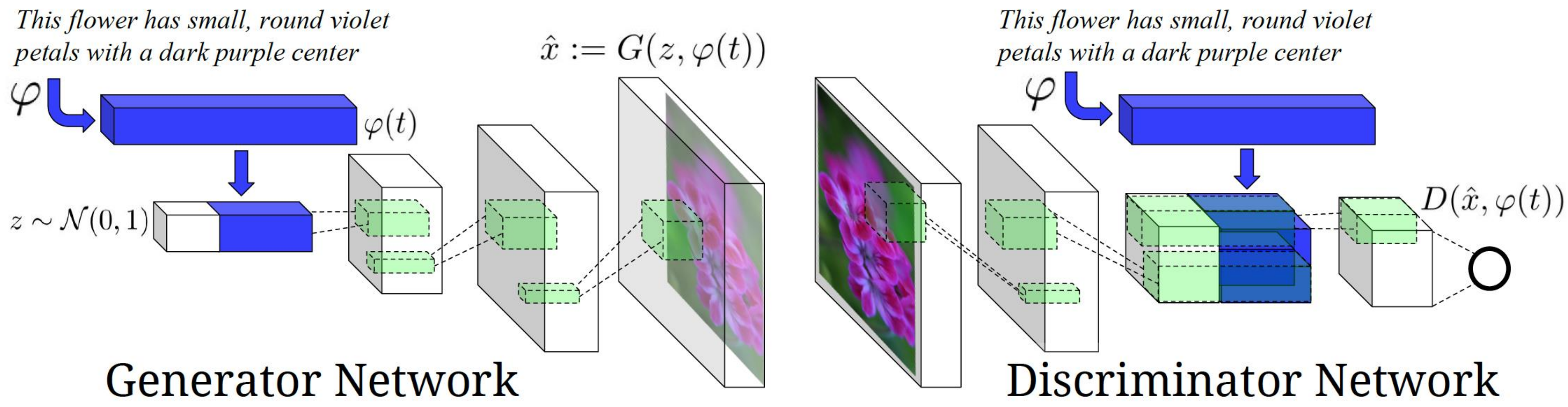


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

this small bird has a pink breast and crown, and black primaries and secondaries.

Objective function:

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

[Reed et al., Generative adversarial text-to-image synthesis, ICML, 2016]
[Xu et al., 'AttnGAN: Fine-grained text to image generation with Attentional GANs, CVPR 2018]

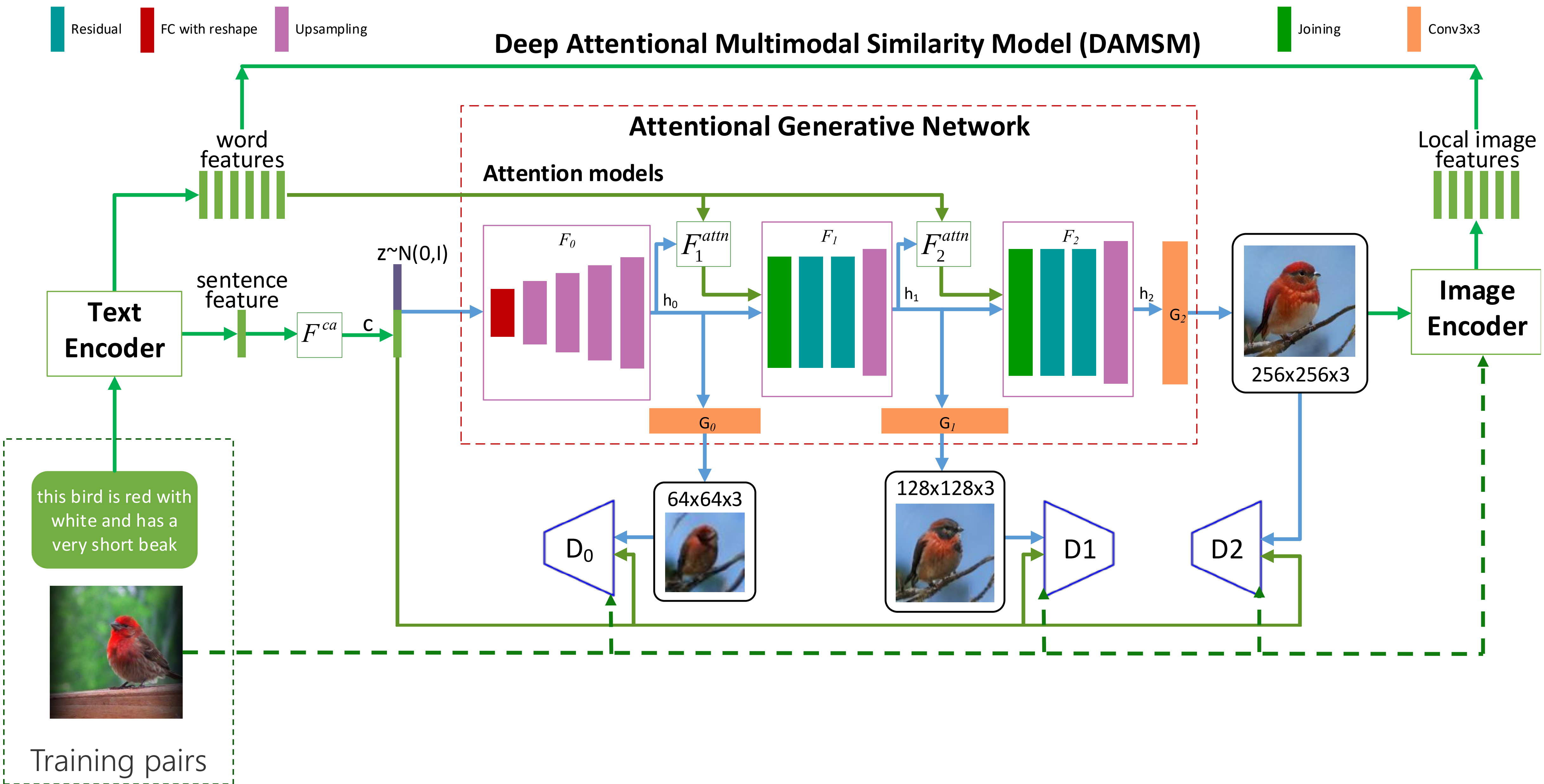
Attention Generative Adversarial Networks (AttnGANs)

Propose AttnGANs to improve image generation

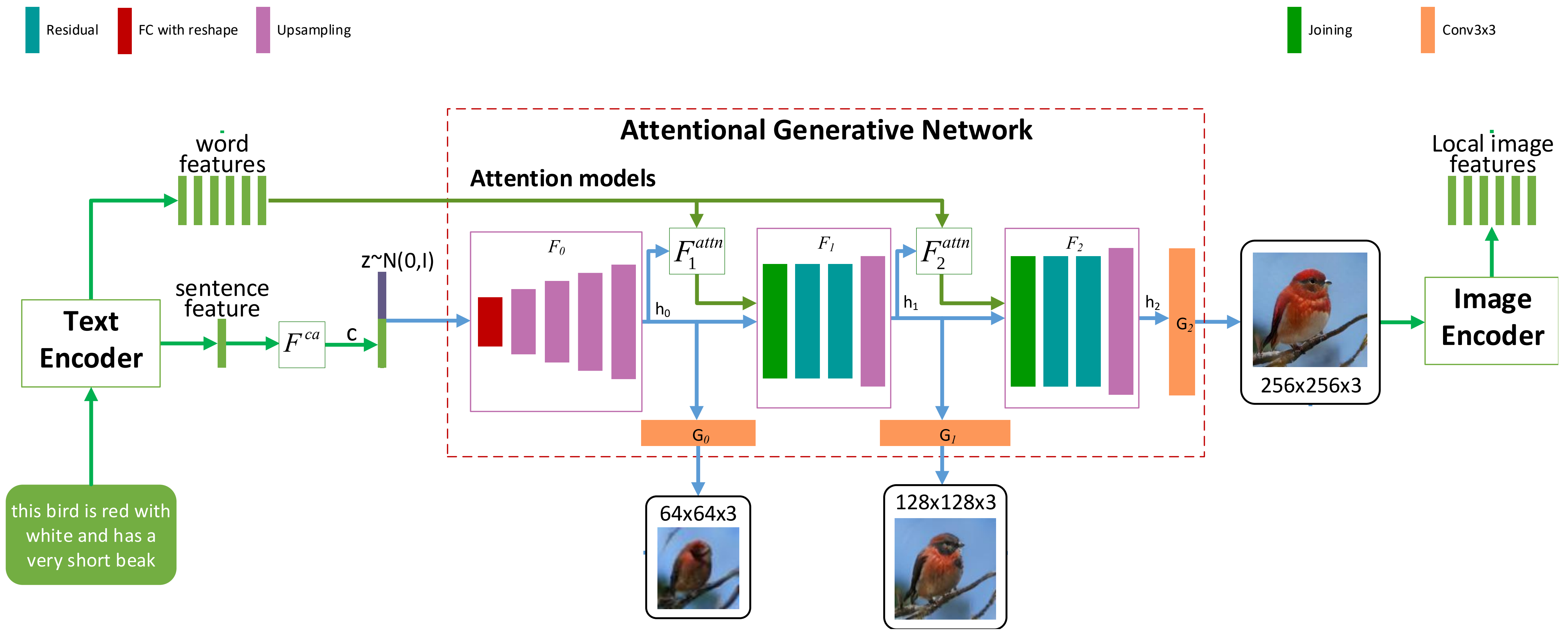
- Goals
  - Improve the quality of generated images
  - Improve the interpretability of GANs
  - Stabilize the training of GANs
- Two main contributions
  - Propose the generative networks with stacked attention to generate images from low-to-high resolutions at multiple stages.
  - Propose a deep attention multimodal similarity model to learn visually-discriminative word features in an semi-supervised manner.

[Xu, Zhang, Huang, Zhang, Gan, Huang, He, "AttnGAN: Fine-grained text to image generation with Attentional GANs", CVPR, 2018]

AttnGAN: Overview



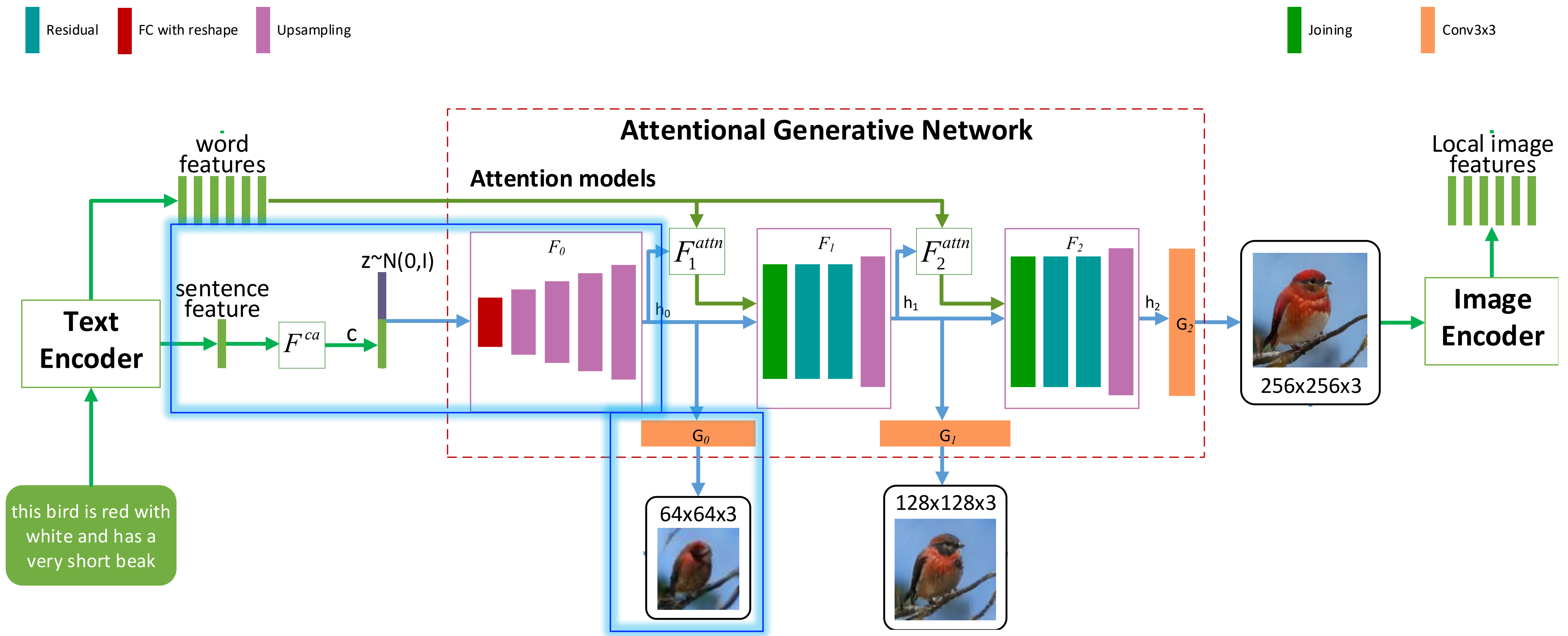
AttnGAN: Attentional Generative Network



Attentional Generative Network:

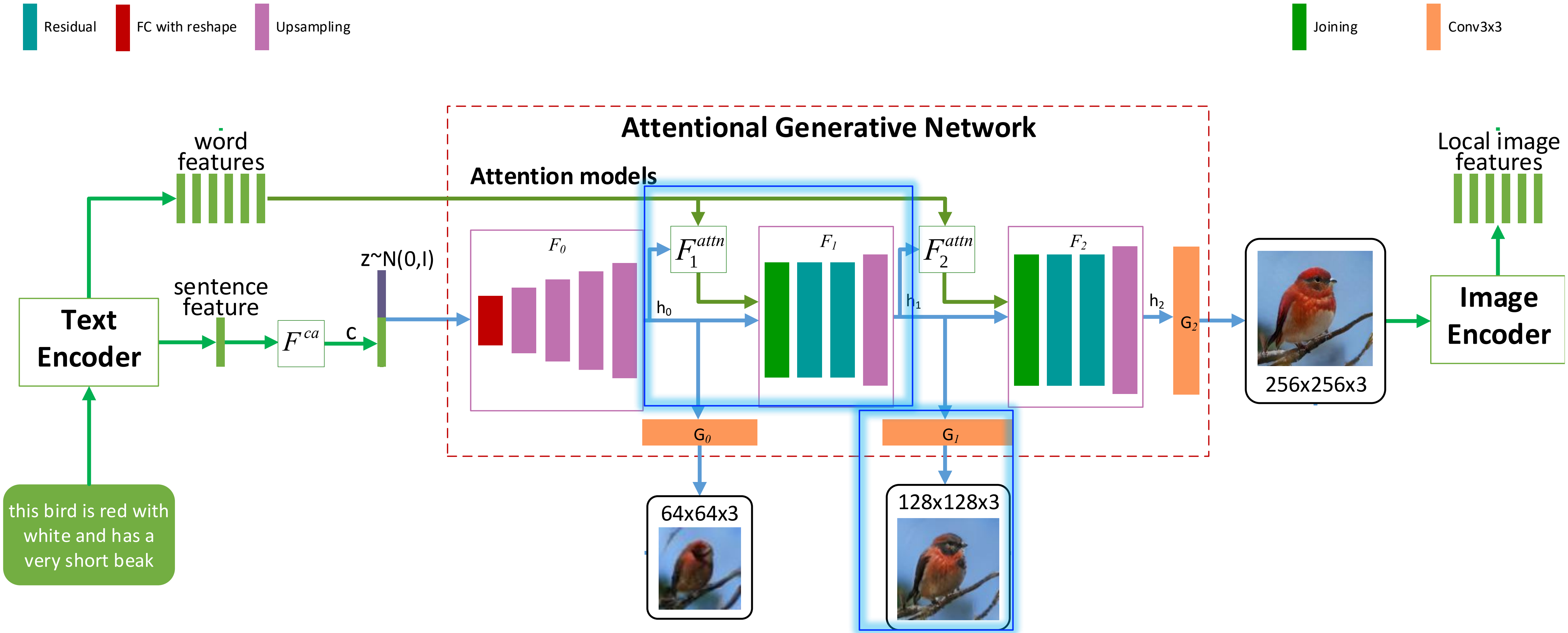
- Takes multi-level conditions (global-level sentence feature and fine-grained word features) as input.
- Generates images from low-to-high resolutions at multiple stages.

AttnGAN: Attentional Generative network



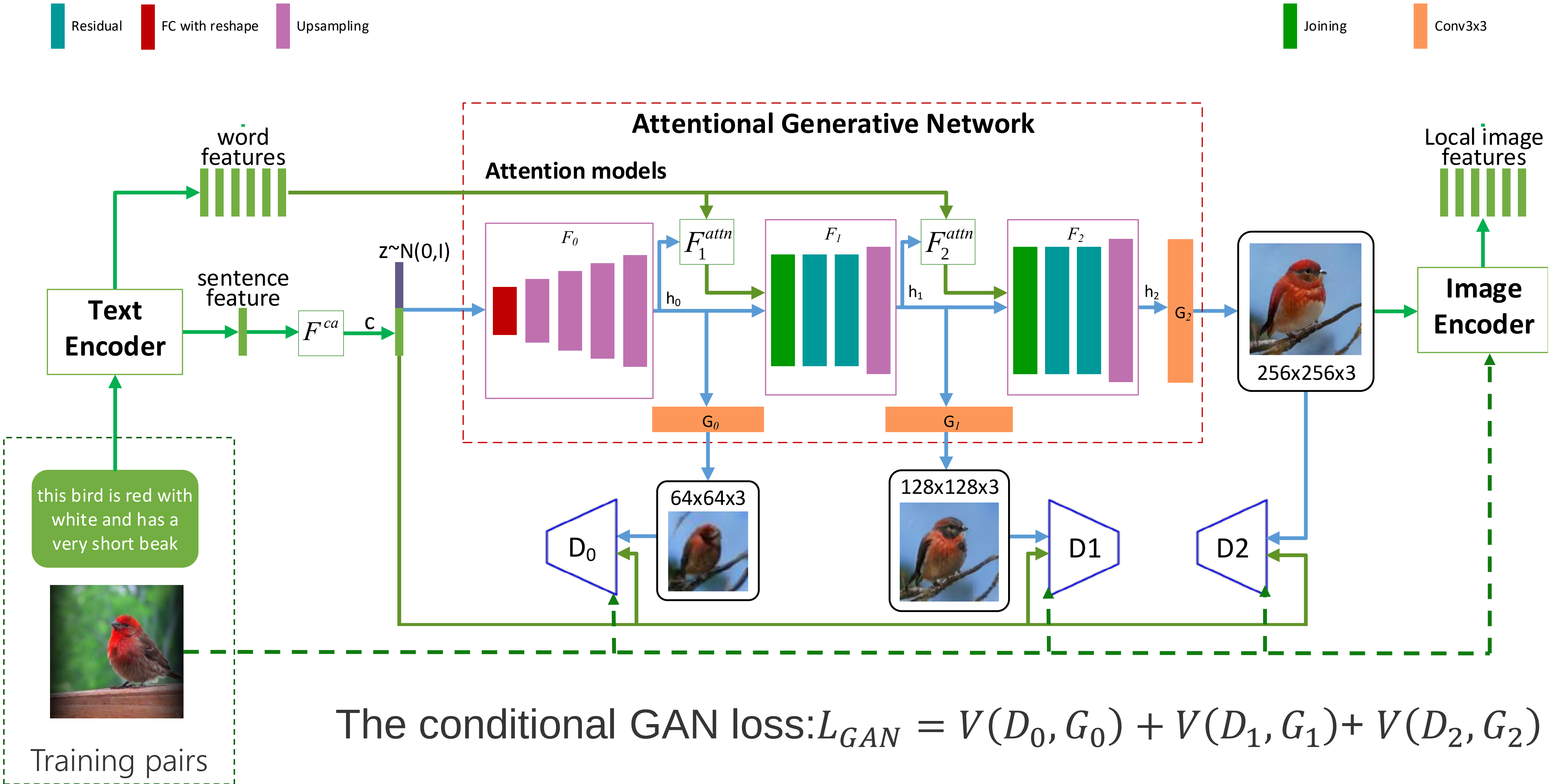
- In the first stage:
  - based on the sentence feature, the image with basic color and shape is generated by generator G_0;
  - hidden features h_0 are decoded from the sentence feature.

AttnGAN: Attentional Generative network



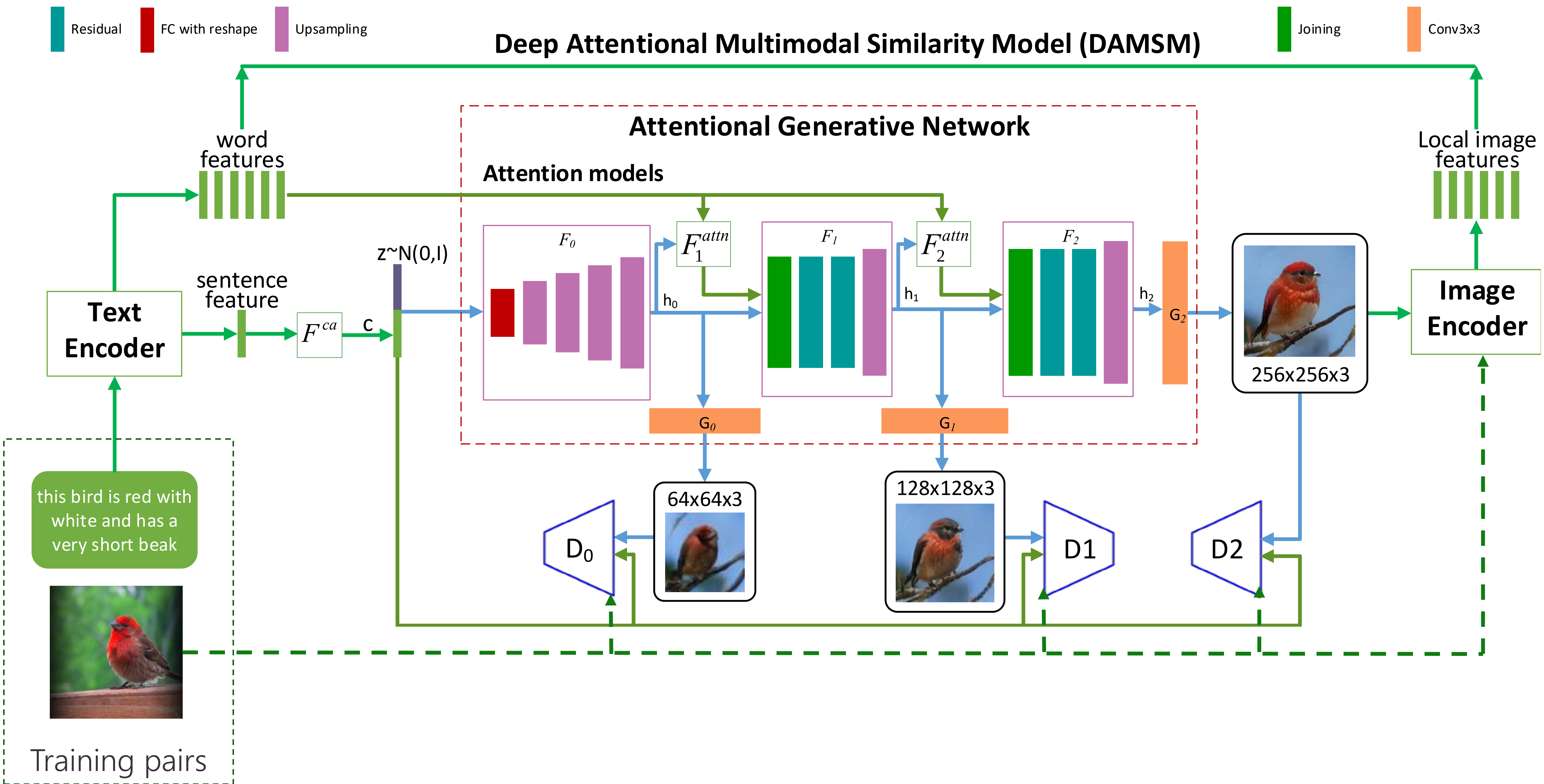
- In following stages, attention models are built.
  - For each region feature of previous generated image, compute its word-context vector.
  - Concatenate previous image region features (e.g., h_0) and word-context vectors to generate image with higher resolution.

AttnGAN: the conditional GAN loss



The conditional GAN loss: $L_{GAN} = V(D_0, G_0) + V(D_1, G_1) + V(D_2, G_2)$

AttnGAN: Overview



AttnGAN: DAMSM sub-network

Deep Attentional Multimodal Similarity Model (DAMSM)

word features

Local image features

Text Encoder

Image Encoder

- Text encoder (LSTM) extracts word features $e_1, e_2, \ldots, e_T$
- Image encoder (CNN) extracts image region features $v_1, v_2, \ldots, v_N$, where N = 288
- Attention mechanism: for the *i-th* word, compute its region-context vector c_i,

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_1 \bar{s}_{i,k})}$$

  - $\bar{s}_{i,j}$ is the dot product between features of the i-th word and the j-th image region;
- Compute the similarity score $R(c_i, e_i)$ between word and image from cosine similarity between e_i and c_i;
- Compute the similarity score between the sentence (D) and the image (Q) from the fine-grained word-region pair information.

$$R(Q, D) = \log\left(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i))\right)^{\frac{1}{\gamma_2}}$$

this bird is red with white and has a very short beak

Training pairs

AttnGAN: DAMSM sub-network

Deep Attentional Multimodal Similarity Model (DAMSM)

word features

Local image features

Text Encoder

Image Encoder

- The DAMSM loss: maximize the similarity score between the images and their corresponding text descriptions (ground truth), i.e.,

$$L_{DAMSM} = -\sum_{i=1}^{M} \log P(D_i|Q_i)$$

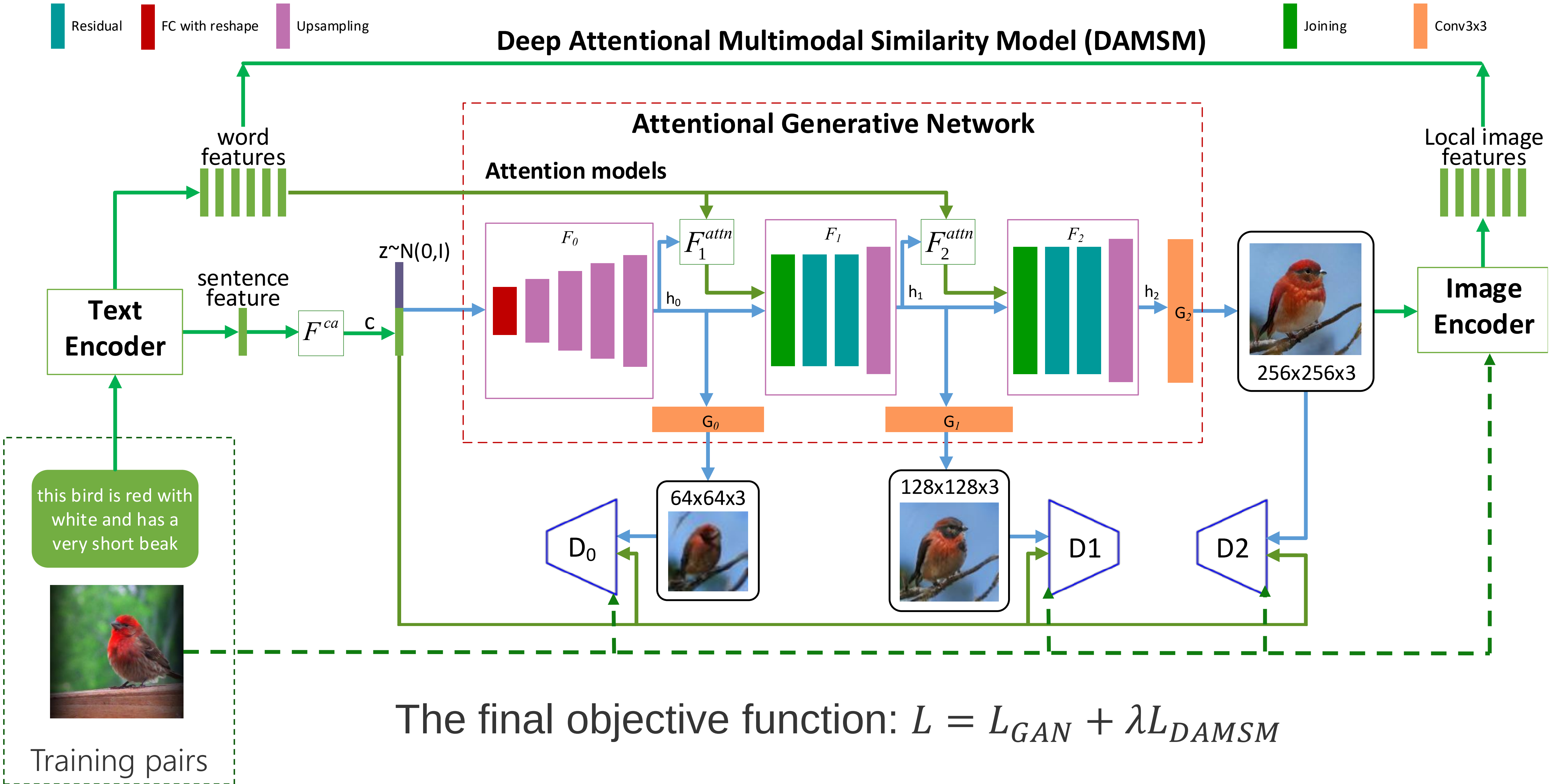
  - M is the number of training pairs.

- The DAMSM loss provides a fine-grained word-region matching loss for training the generator.

this bird is red with white and has a very short beak

Training pairs

AttnGAN: Overview

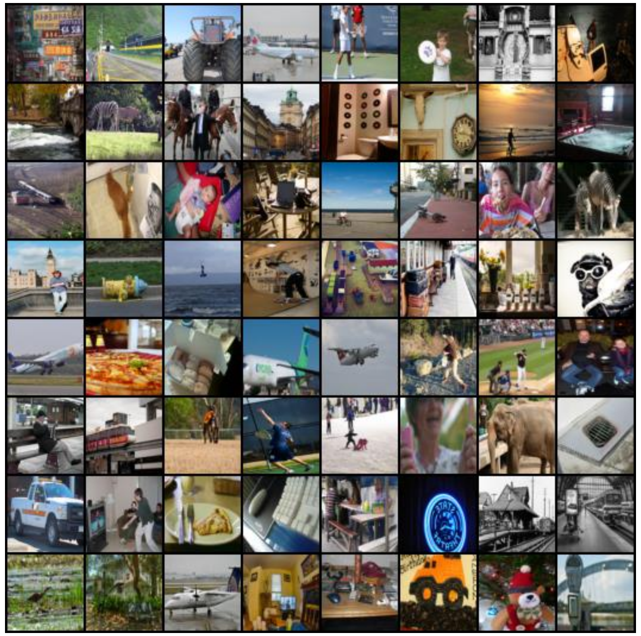


The final objective function: $L = L_{GAN} + \lambda L_{DAMSM}$

Experiments

- Datasets

<table>
<tr><th rowspan="2">Datasets</th><th colspan="2">CUB-2011</th><th colspan="2">MS-COCO</th></tr>
<tr><th>train</th><th>test</th><th>train</th><th>test</th></tr>
<tr><td># samples</td><td>8,855</td><td>2,933</td><td>80,000</td><td>40,000</td></tr>
<tr><td>caption/ image</td><td>10</td><td>10</td><td>5</td><td>5</td></tr>
</table>



- Evaluation metrics:
  - **Inception score** reflects the quality and diversity of the generated images.
  - **R-precision** reflects whether the generated images are well conditioned.

Comparison with previous methods

- On CUB dataset, our AttnGAN achieves 4.36 inception score, which significantly outperforms the previous best inception score of 3.82.
- On the COCO dataset, our AttnGAN boosts the best reported inception score from 9.58 to 25.89, a 170.25% improvement relatively.

Dataset	GAN-INT-CLS [1]	GAWWN [2]	StackGAN [3]	StackGAN-v2 [4]	PPGN [5]	Our AttnGAN
CUB	$2.88 \pm .04$	$3.62 \pm .07$	$3.70 \pm .04$	$3.82 \pm .06$	\	**$4.36 \pm .03$**
COCO	$7.88 \pm .07$	\	$8.45 \pm .03$	\	$9.58 \pm .21$	**$25.89 \pm .47$**

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016.
[2] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In NIPS, 2016.
[3] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017.
[4] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916, 2017.
[5] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In CVPR, 2017.

Quantitative analysis

- The DAMSM loss is important

Method	inception score	R-precision(%)
AttnGAN1, no DAMSM	$3.98 \pm .04$	10.37 ± 5.88
AttnGAN1, $\lambda = 0.1$	$4.19 \pm .06$	16.55 ± 4.83
AttnGAN1, $\lambda = 1$	$4.35 \pm .05$	34.96 ± 4.02
AttnGAN1, $\lambda = 5$	$4.35 \pm .04$	58.65 ± 5.41
AttnGAN1, $\lambda = 10$	$4.29 \pm .05$	63.87 ± 4.85
AttnGAN2, $\lambda = 5$	$\mathbf{4.36 \pm .03}$	$\mathbf{67.82 \pm 4.43}$

Higher inception score means better image quality and diversity.

Higher R-precision rate means better conditioned.

The inception score and the corresponding R-precision rate of AttnGAN models on CUB.
- "AttnGAN1" architecture has one attention model and generates images of 128x128 resolution;
- "AttnGAN2" architecture has two attention models and generates images of 256x256 resolution.

Quantitative analysis

- The DAMSM loss is important

Method	inception score	R-precision(%)
AttnGAN1, no DAMSM	$3.98 \pm .04$	10.37 ± 5.88
AttnGAN1, $\lambda = 0.1$	$4.19 \pm .06$	16.55 ± 4.83
AttnGAN1, $\lambda = 1$	$4.35 \pm .05$	34.96 ± 4.02
AttnGAN1, $\lambda = 5$	$4.35 \pm .04$	58.65 ± 5.41
AttnGAN1, $\lambda = 10$	$4.29 \pm .05$	63.87 ± 4.85
AttnGAN2, $\lambda = 5$	**$4.36 \pm .03$**	**67.82 ± 4.43**

Higher inception score means better image quality and diversity.

Higher R-precision rate means better conditioned.

The inception score and the corresponding R-precision rate of AttnGAN models on CUB.

- "AttnGAN1" architecture has one attention model and generates images of 128x128 resolution;
- "AttnGAN2" architecture has two attention models and generates images of 256x256 resolution.

Quantitative analysis

- Stacking more attention models helps

Method	inception score	R-precision(%)
AttnGAN1, no DAMSM	$3.98 \pm .04$	10.37 ± 5.88
AttnGAN1, $\lambda = 0.1$	$4.19 \pm .06$	16.55 ± 4.83
AttnGAN1, $\lambda = 1$	$4.35 \pm .05$	34.96 ± 4.02
AttnGAN1, $\lambda = 5$	$4.35 \pm .04$	58.65 ± 5.41
AttnGAN1, $\lambda = 10$	$4.29 \pm .05$	63.87 ± 4.85
AttnGAN2, $\lambda = 5$	**$4.36 \pm .03$**	**67.82 ± 4.43**

Higher inception score means better image quality and diversity.

Higher R-precision rate means better conditioned.

The inception score and the corresponding R-precision rate of AttnGAN models on CUB.
- "AttnGAN1" architecture has one attention model and generates images of 128x128 resolution;
- "AttnGAN2" architecture has two attention models and generates images of 256x256 resolution.

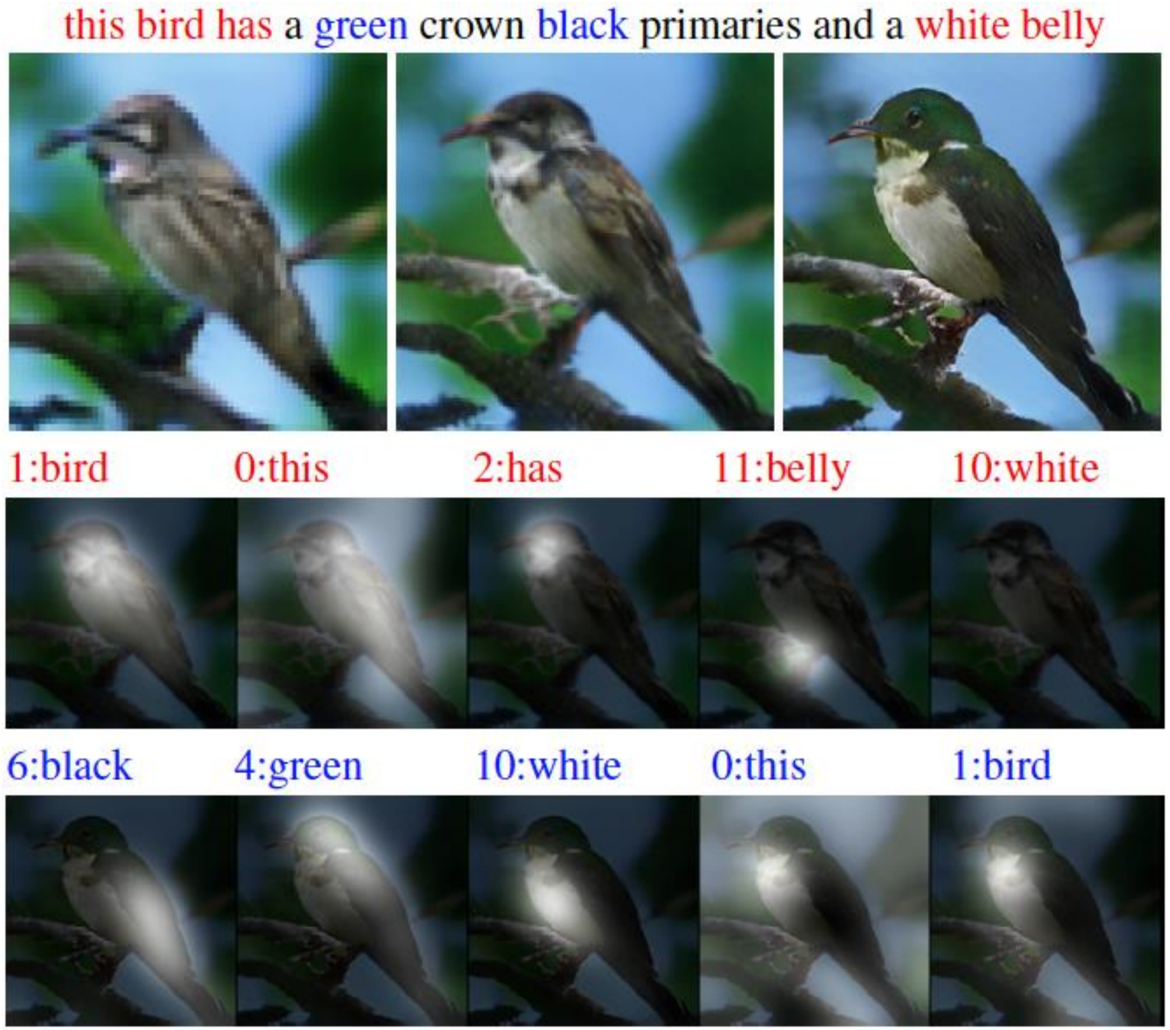
Examples – CUB attention maps

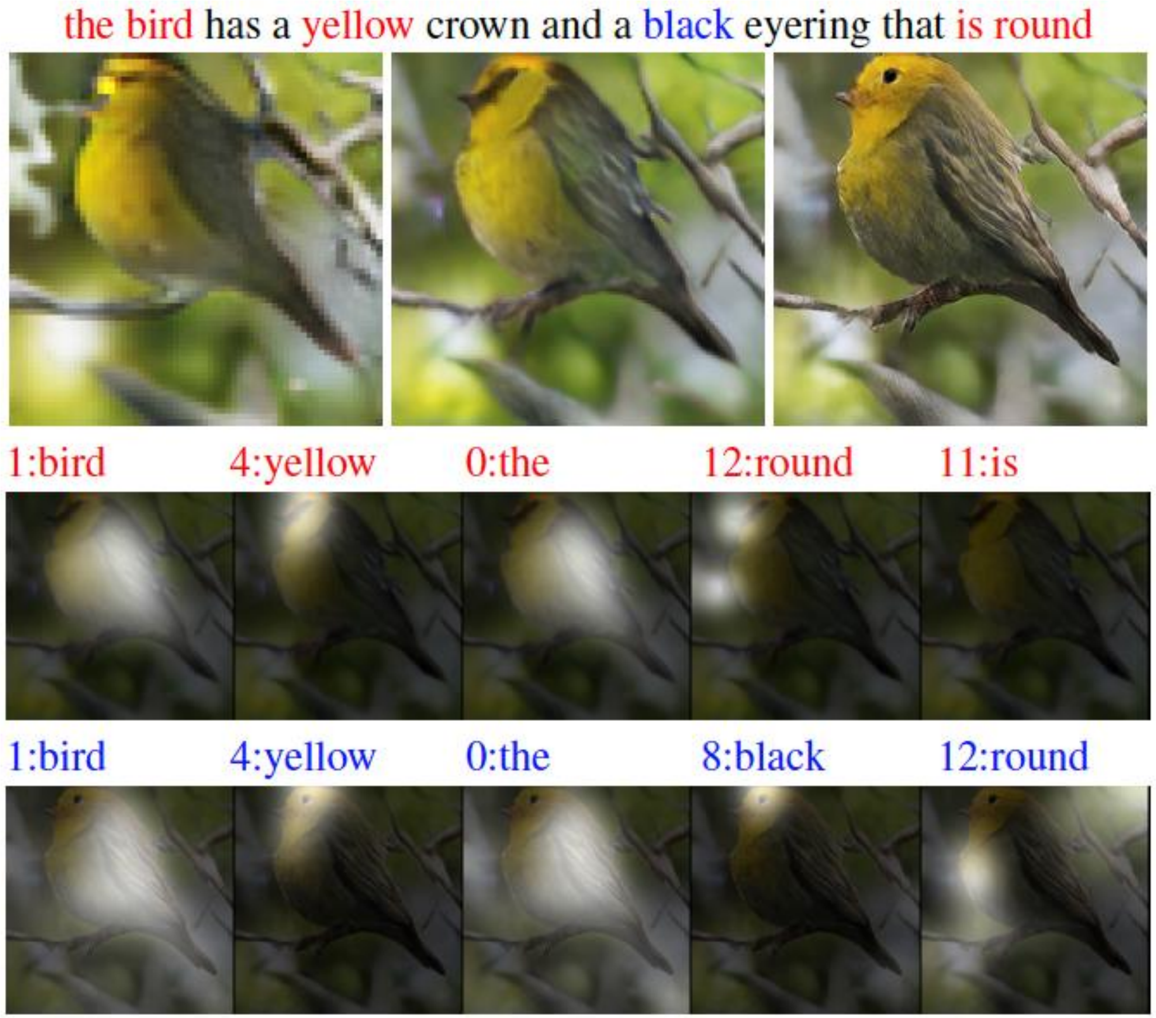
this bird is red with white and has a very short beak.

Challenges – COCO attention maps

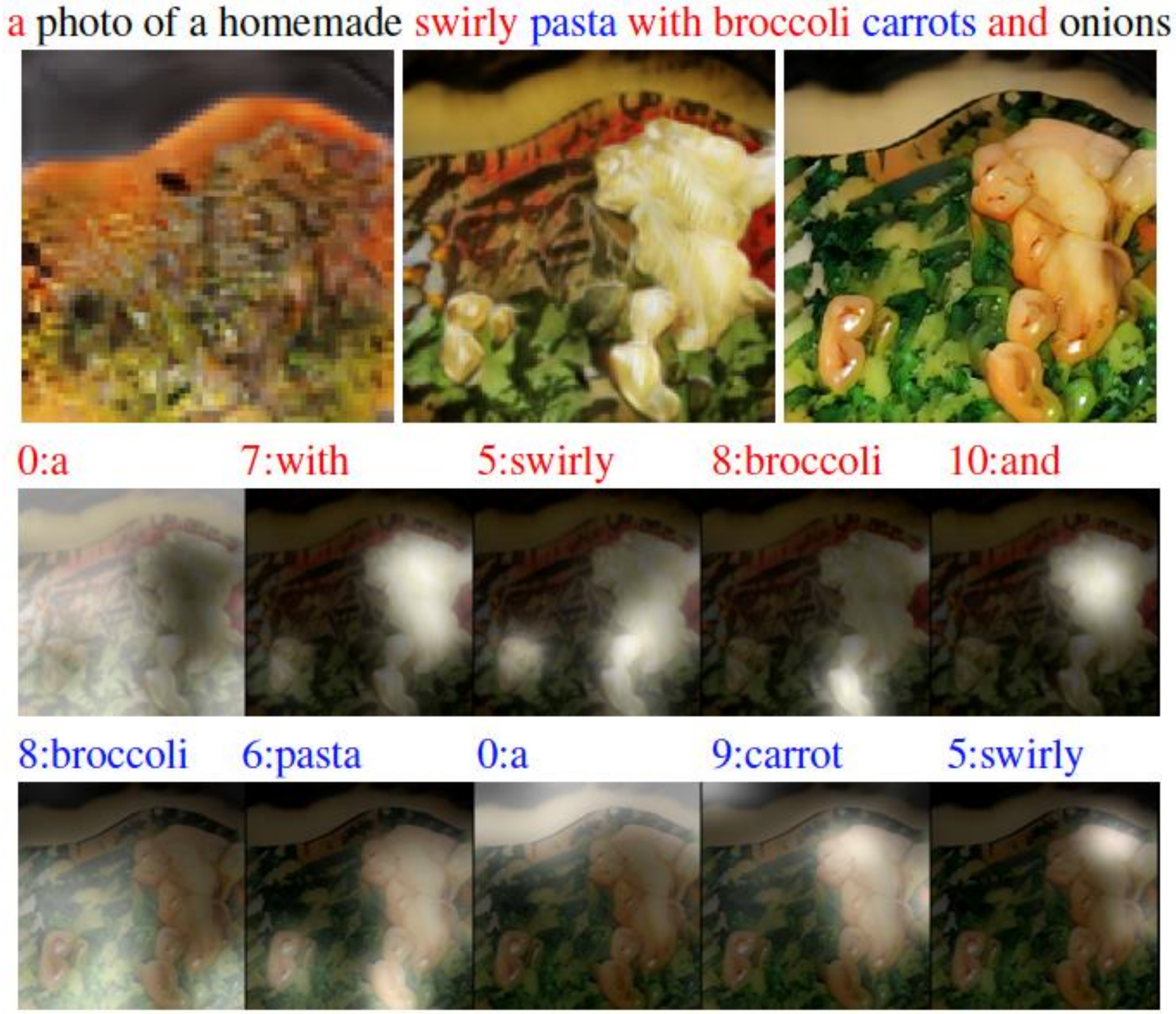
A fruit stand display with bananas and kiwi.

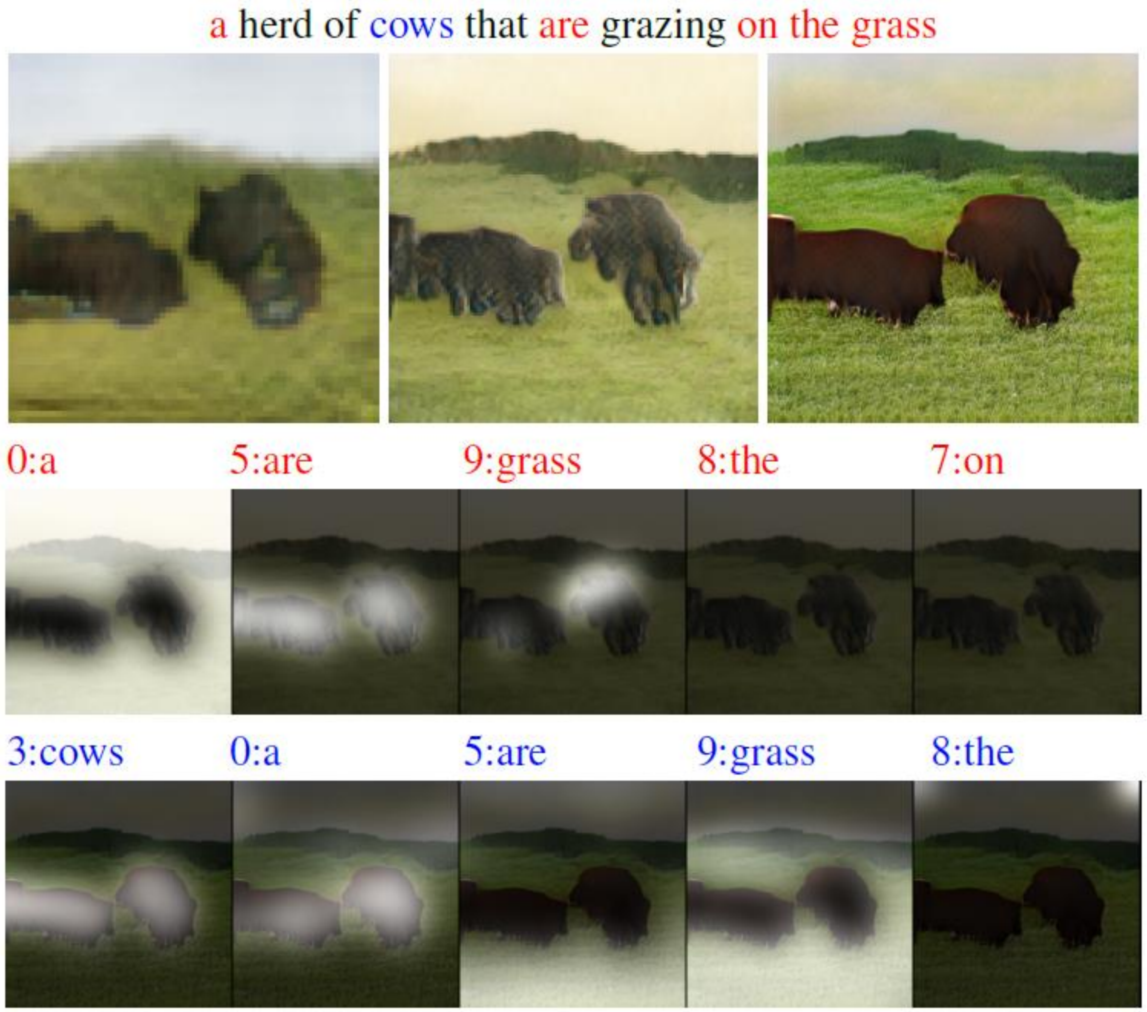
Examples – CUB attention maps





Challenges – COCO attention maps





Qualitative analysis - generalization ability

- Change some most attended words in the text descriptions

this bird has wings that are black and has a white belly

this bird has wings that are red and has a yellow belly

this bird has wings that are blue and has a red belly

Qualitative analysis - generalization ability

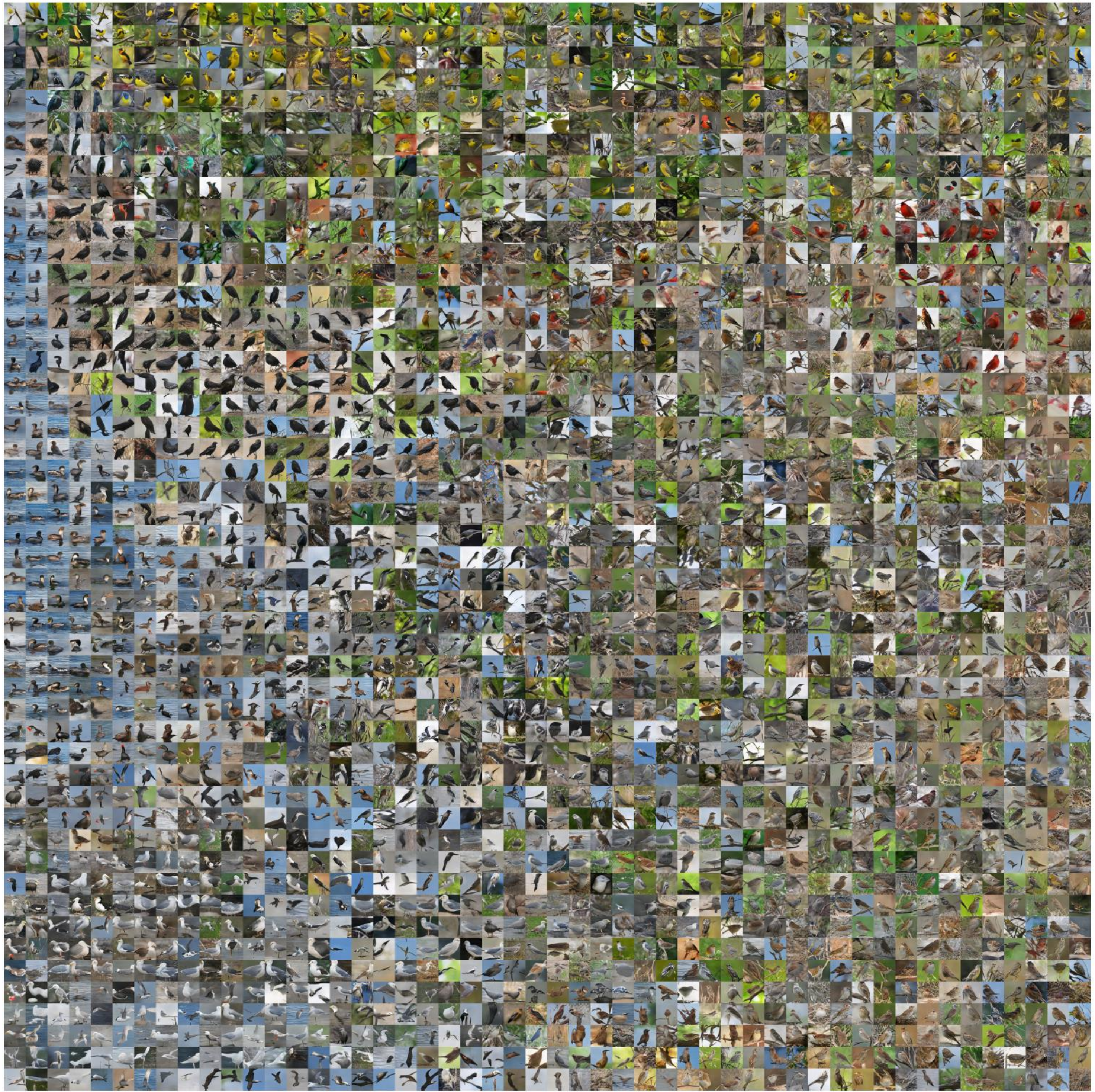
- Images generated from descriptions of novel scenarios

Qualitative analysis - generalization ability

- Novel images (failure cases) generated by AttnGAN on the CUB test set

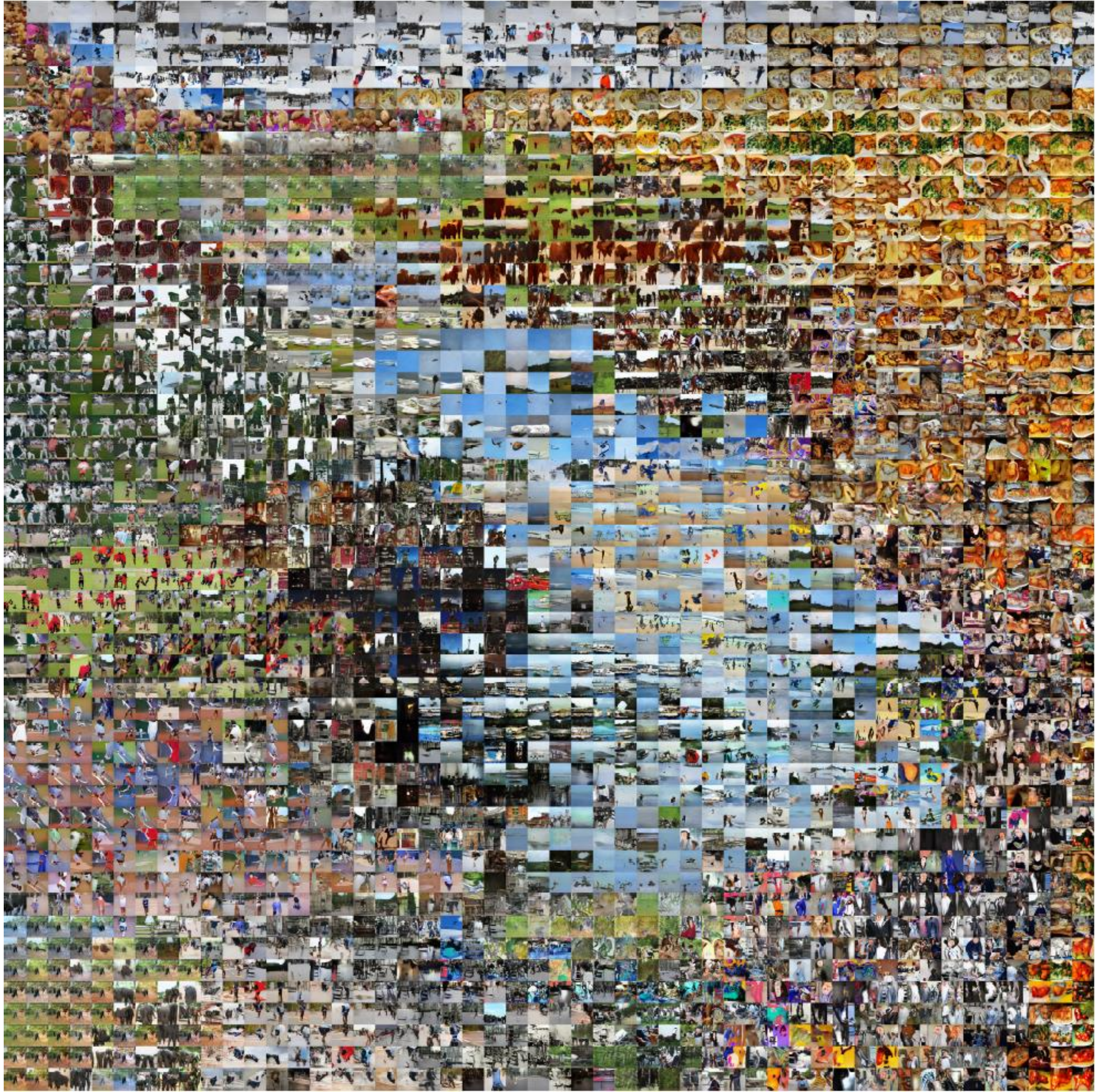
Utilizing t-SNE to embed a large number of images generated by the AttnGAN

CUB-2011



Utilizing t-SNE to embed a large number of images generated by the AttnGAN

MS-COCO



DrawingBot vs CaptionBot

A herd of cows that are grazing on the grass.

An old clock next to a light post in front of a steeple.

The girl is surfing a small wave in the water.

A stop sign flying in the sky.

A red bus is floating on a lake.

What Microsoft CaptionBot sees... https://www.captionbot.ai/

I think it's a herd of cattle grazing on a lush green field.

I think it's a clock tower in the middle of the street.

I think it's a young girl riding a wave on a surfboard in the water.

I think it's a red and white sign.

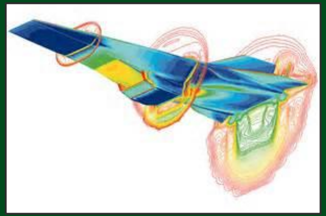
I think it's a boat that is sitting on a bus.

Summary

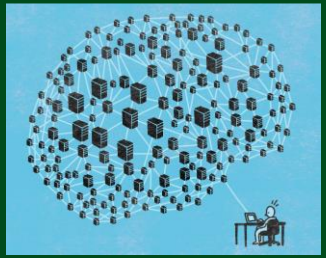
- An Attentional Generative Adversarial Network (AttnGAN) is proposed for fine-grained language-to-image generation.
- Our AttnGAN significantly outperforms previous state-of-the-art GAN models.
- AttnGAN is more stable to train, and has better interpretability.
- AttnGAN code: https://github.com/taoxugit/AttnGAN
- DrawingBot demo: under construction
- CaptionBot demo: https://www.captionbot.ai/

Widest selection of GPUs in the cloud

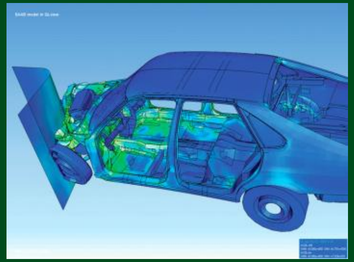
High Performance Computing



AI & Machine Learning

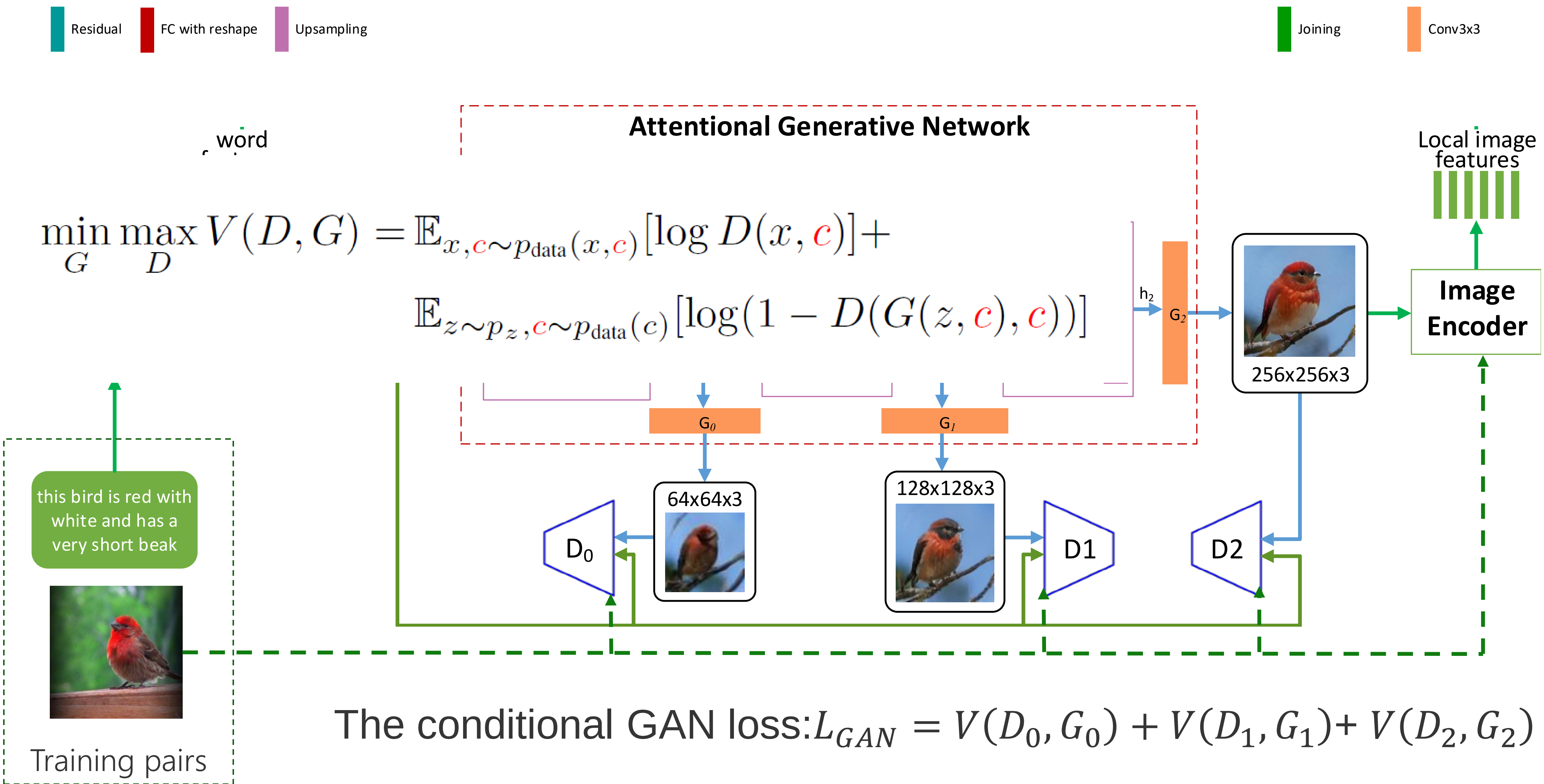


Remote Visualization



Learn more at booth #603

AttnGAN: the conditional GAN loss



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

The conditional GAN loss: $L_{GAN} = V(D_0, G_0) + V(D_1, G_1) + V(D_2, G_2)$