the flower has petals that are bright pinkish purple with white stigma

this white and yellow flower have thin white petals and a round yellow stamen

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

this magnificent fellow is almost all black with a red crest, and white cheek patch.

Generative Adversarial Text to Image Synthesis [*]

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran
Bernt Schiele, Honglak Lee

REEDSCOT[1], AKATA[2], XCYAN[1], LLAJAN[1]
SCHIELE[2],HONGLAK[1]

[1] University of Michigan, Ann Arbor, MI, USA (UMICH.EDU)
[2] Max Planck Institute for Informatics, Saarbrücken, Germany (MPI-INF.MPG.DE)

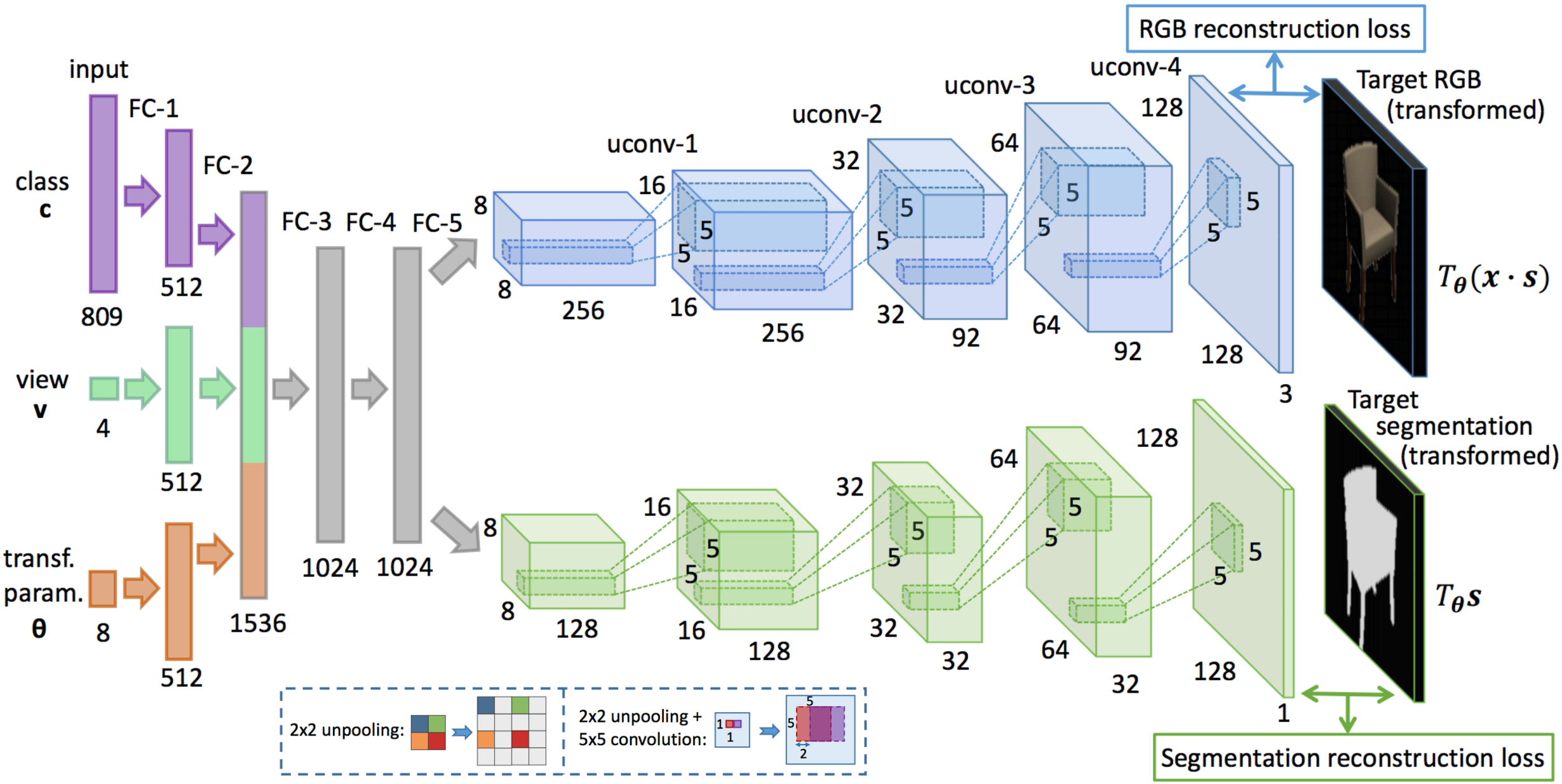
Presented By: Ezgi Ekiz

[*] Reed, Scott, et al. "Generative adversarial text to image synthesis." Proceedings of The 33rd International Conference on Machine Learning. Vol. 3. 2016.

Outline

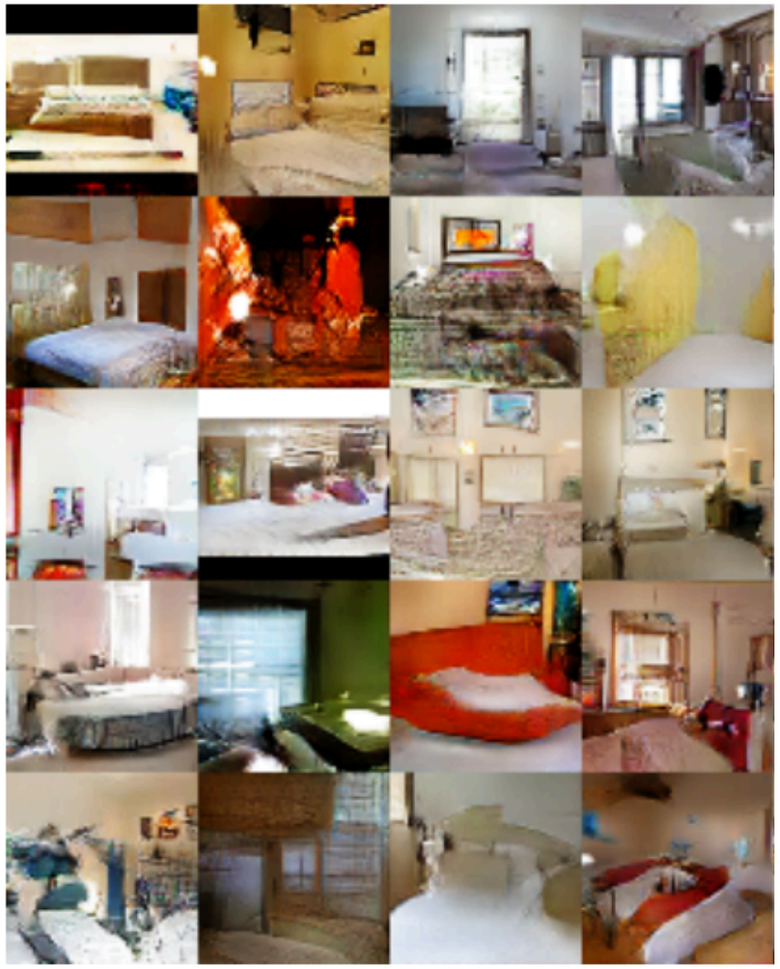
- The goal is to synthesize images that are mistakable for real from textual description. The method is built upon:
  - Text encoding that captures important visual details
  - Generative Adversarial Networks (GAN) and GAN-CLS
  - Manifold interpolation
  - Style transfer

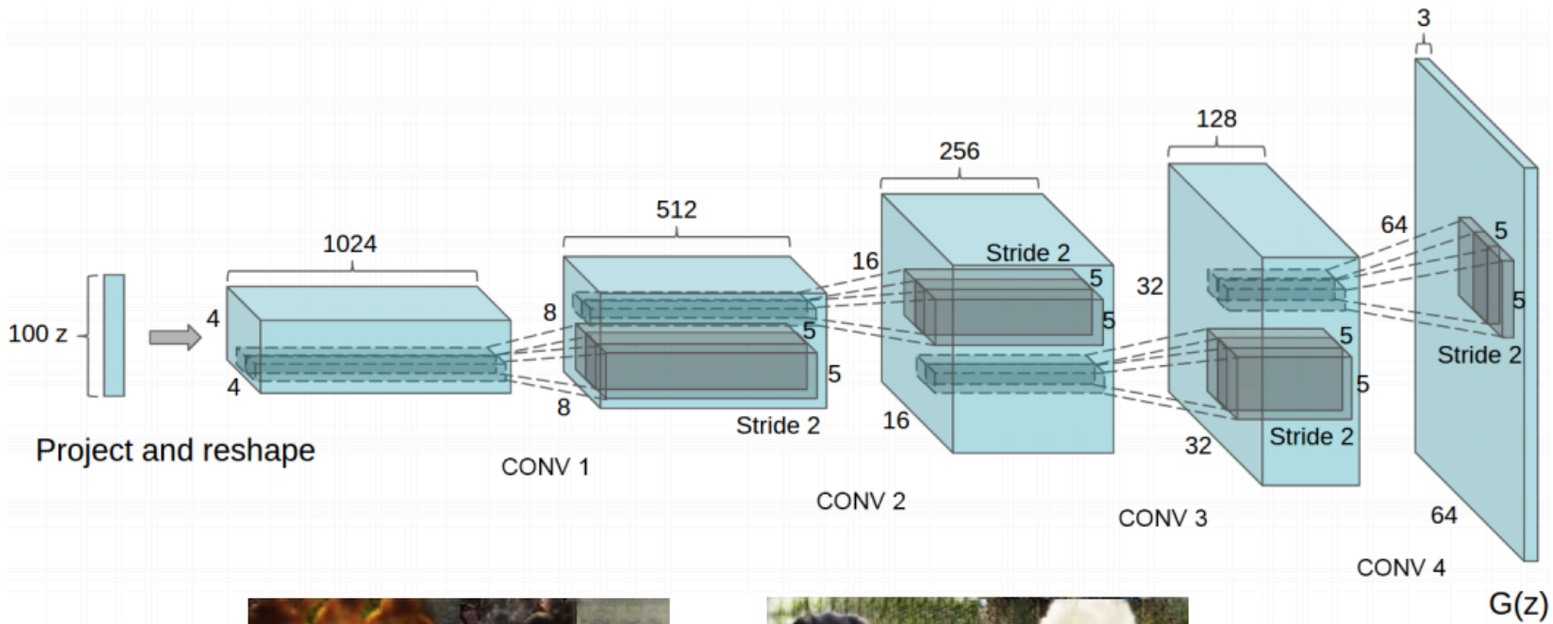
Similar Work

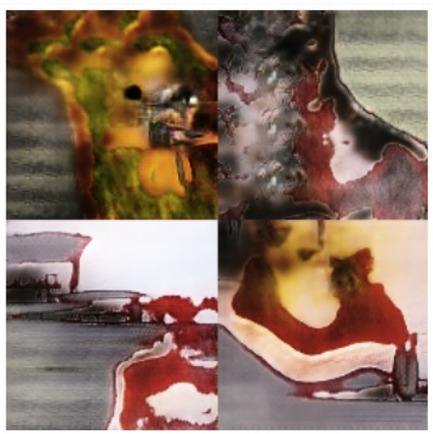


Dosovitskiy, Alexey, Jost Tobias Springenberg, and Thomas Brox. "Learning to generate chairs with convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

Similar Work





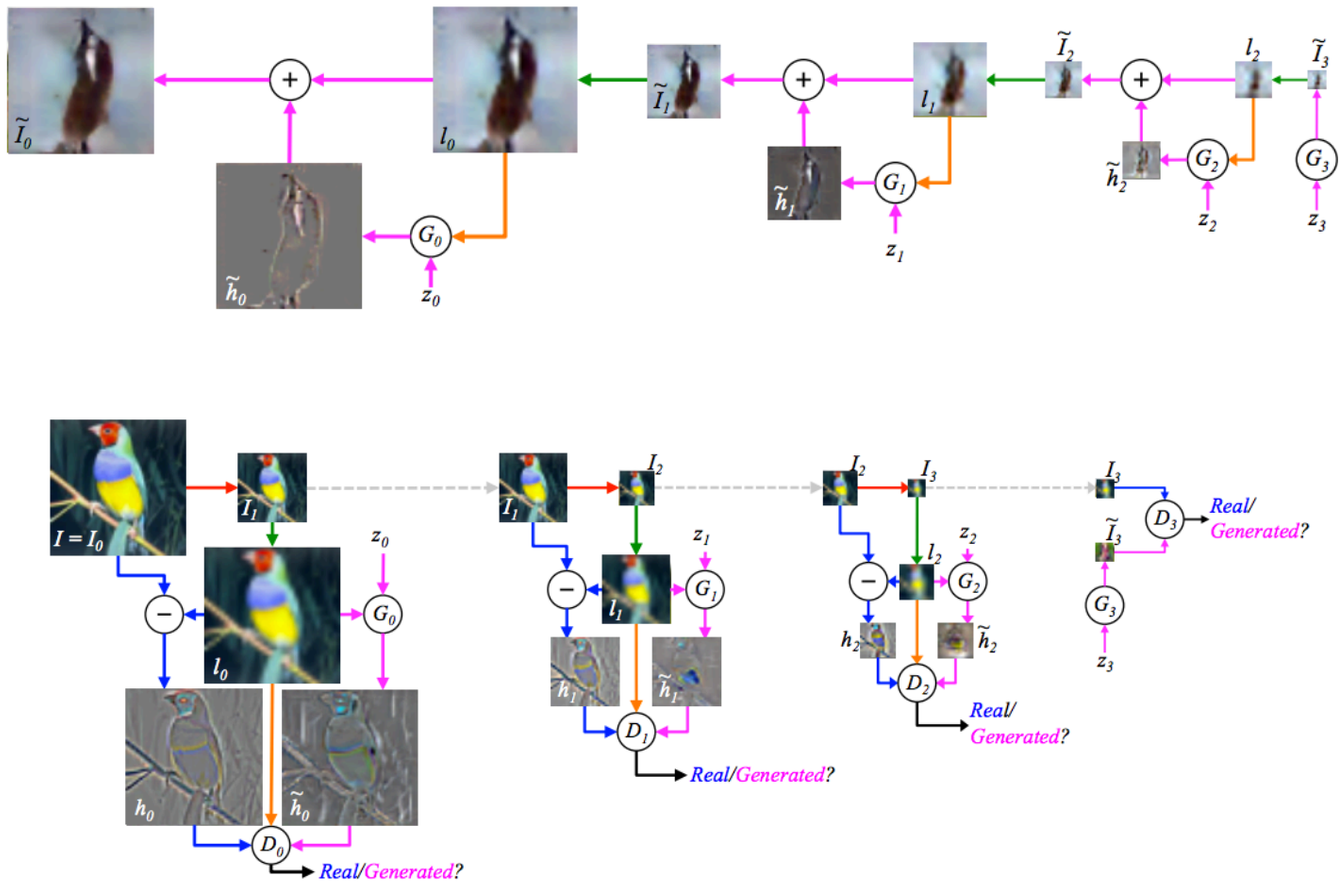


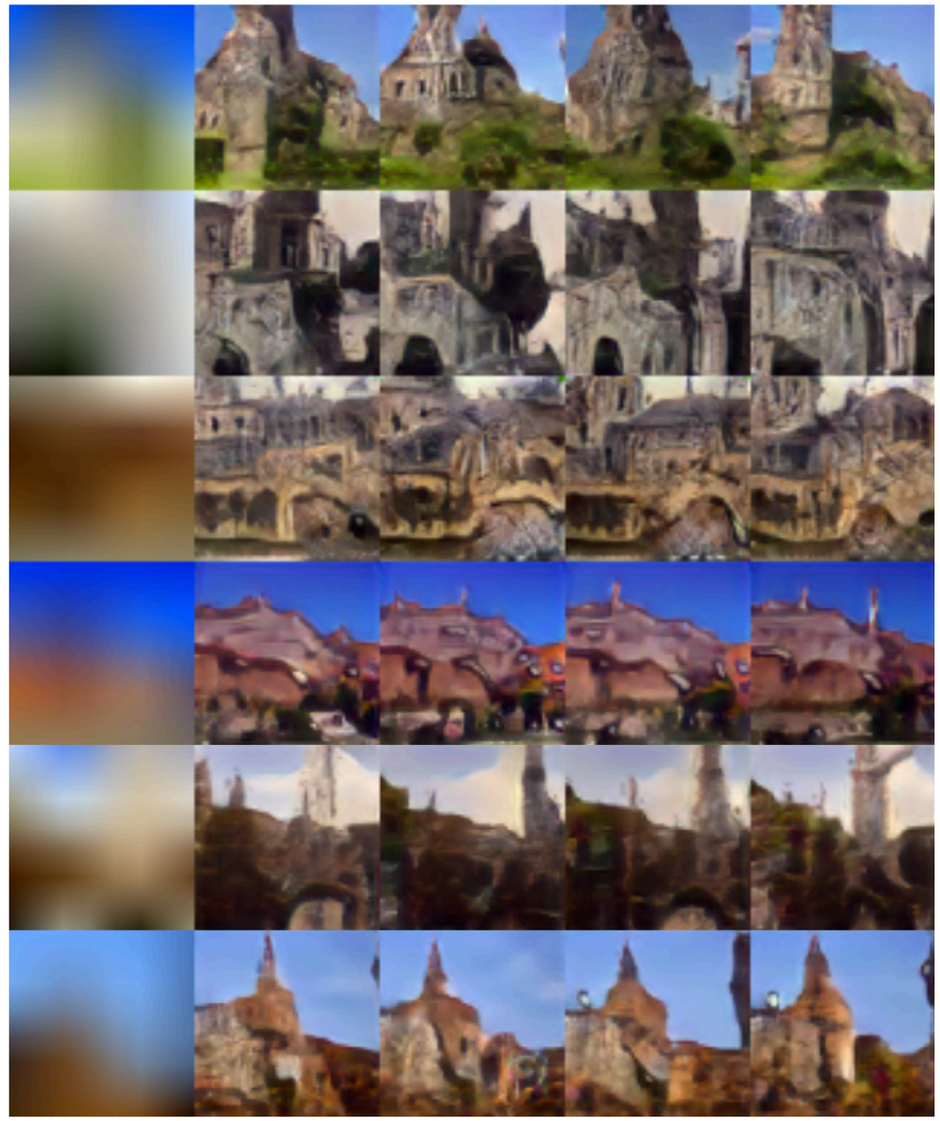
DCGAN

Improved DCGAN

- Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).
- Salimans, Tim, et al. "Improved techniques for training gans." Advances in Neural Information Processing Systems. 2016.

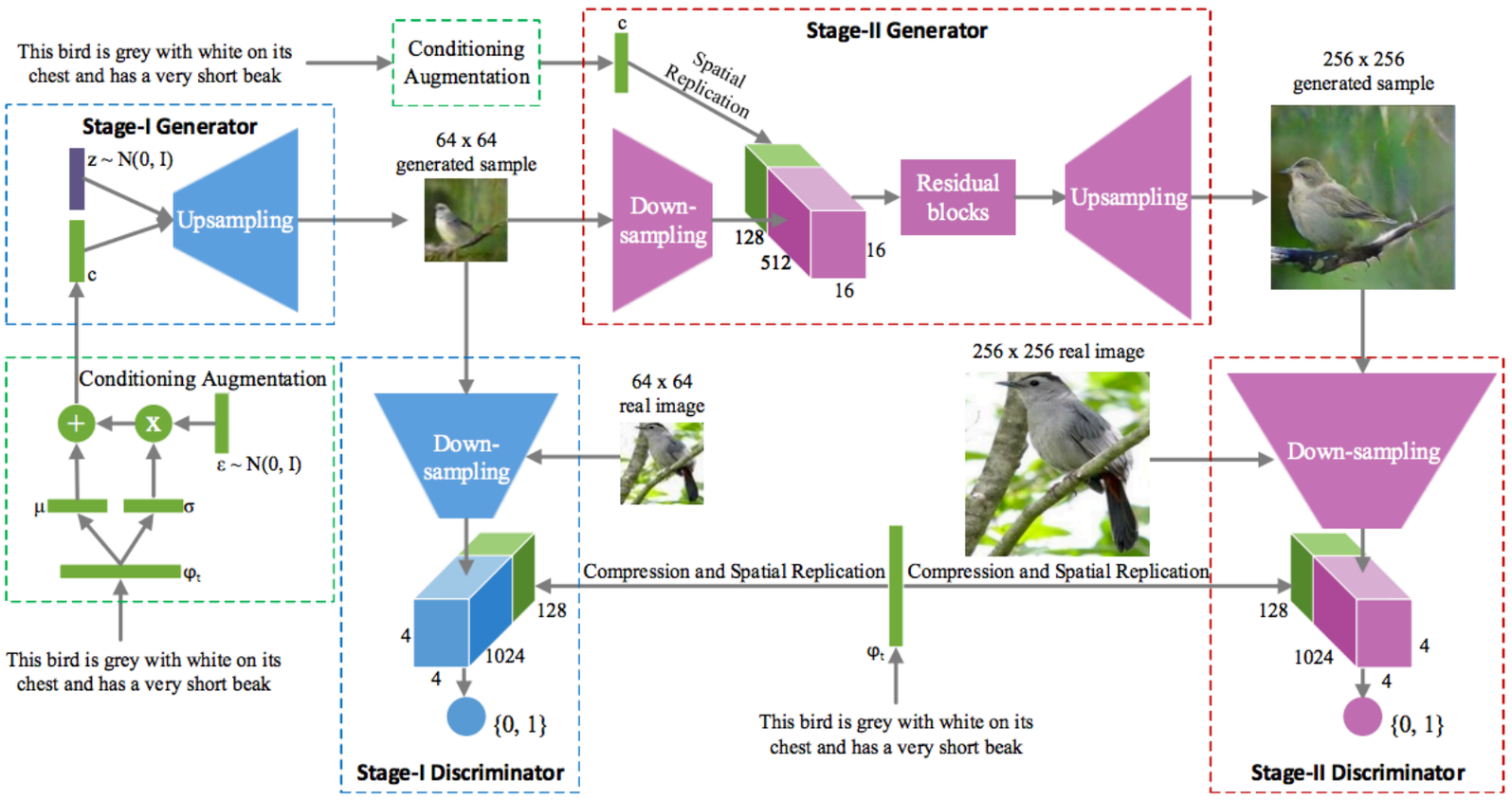
Similar Work

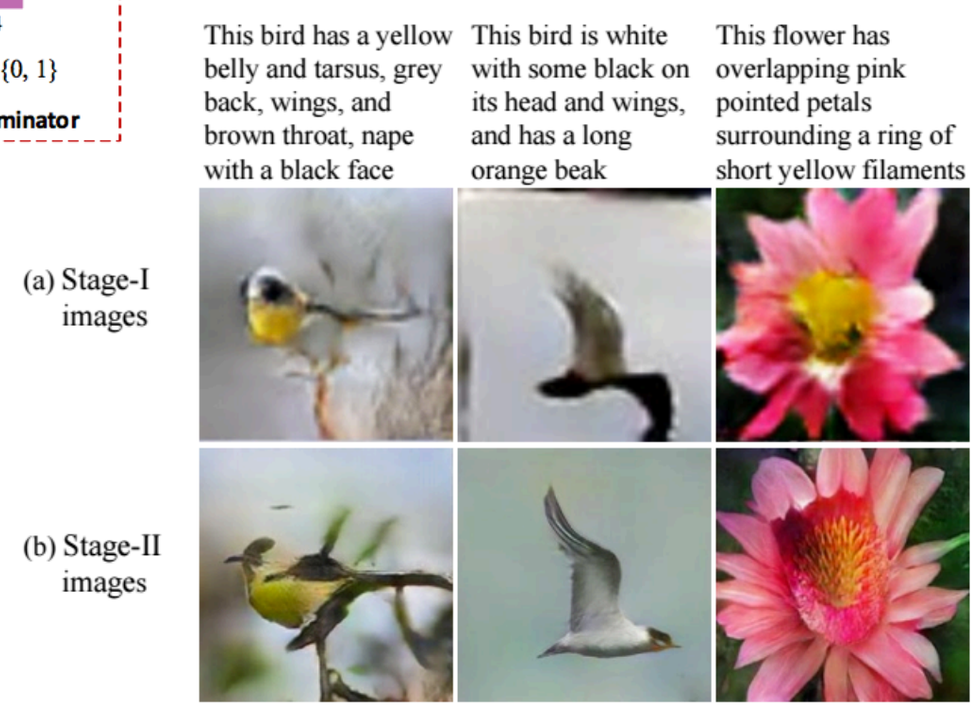




Denton, Emily L., Soumith Chintala, and Rob Fergus. "Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks." Advances in neural information processing systems. 2015.

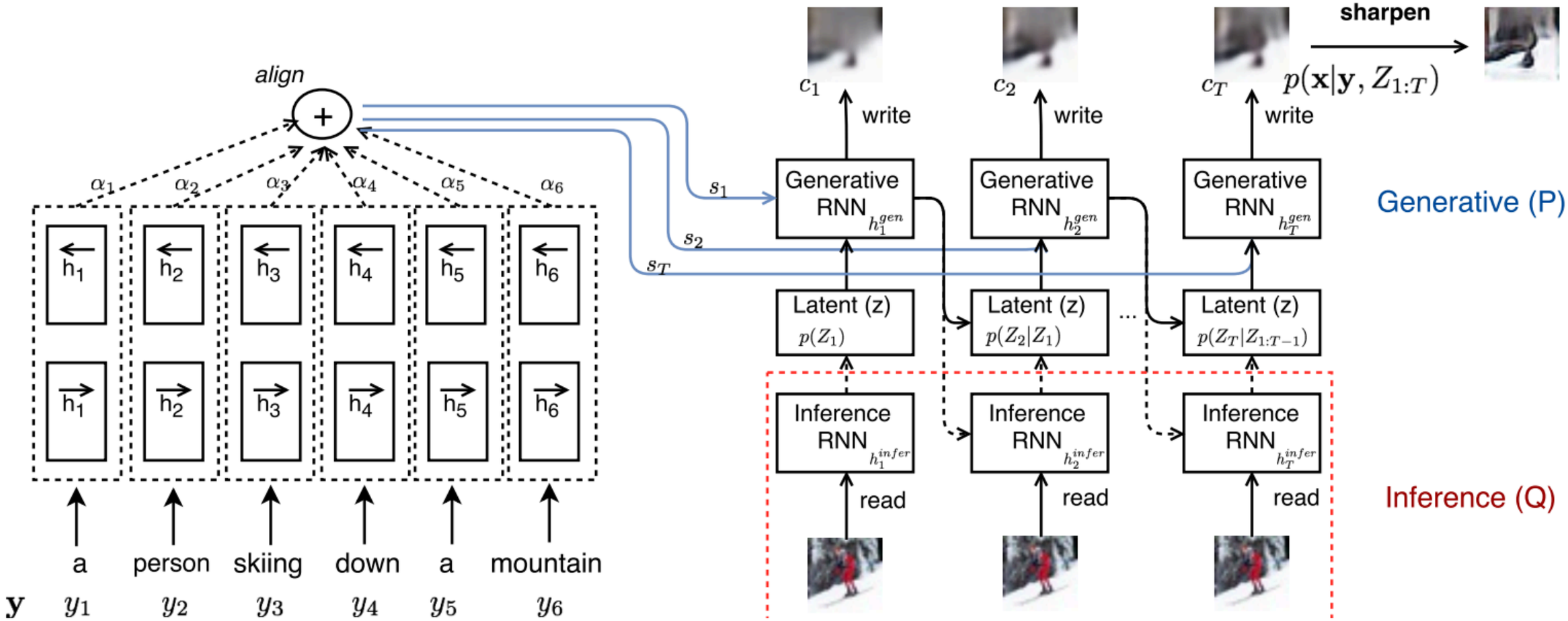
Similar Work

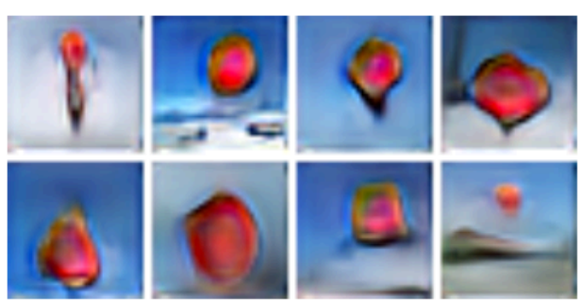




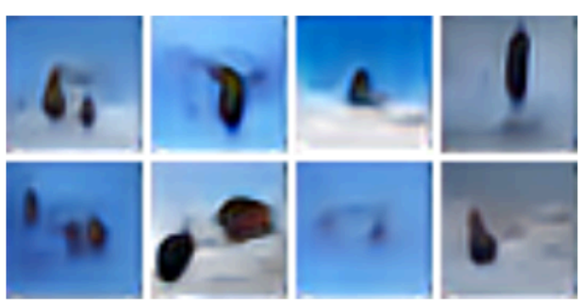
Zhang, Han, et al. "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks." arXiv preprint arXiv:1612.03242 (2016).

Similar Work



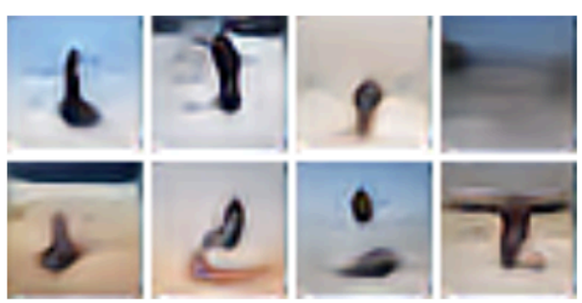


A stop sign is flying in blue skies.



A herd of elephants flying in the blue skies.

A toilet seat sits open in the grass field.



A person skiing on sand clad vast desert.

Mansimov, Elman, et al. "Generating images from captions with attention." arXiv preprint arXiv:1511.02793 (2015).

Text Feature Representation

- The representation should capture important visual details
- Word/character based convolutional recurrent network is used

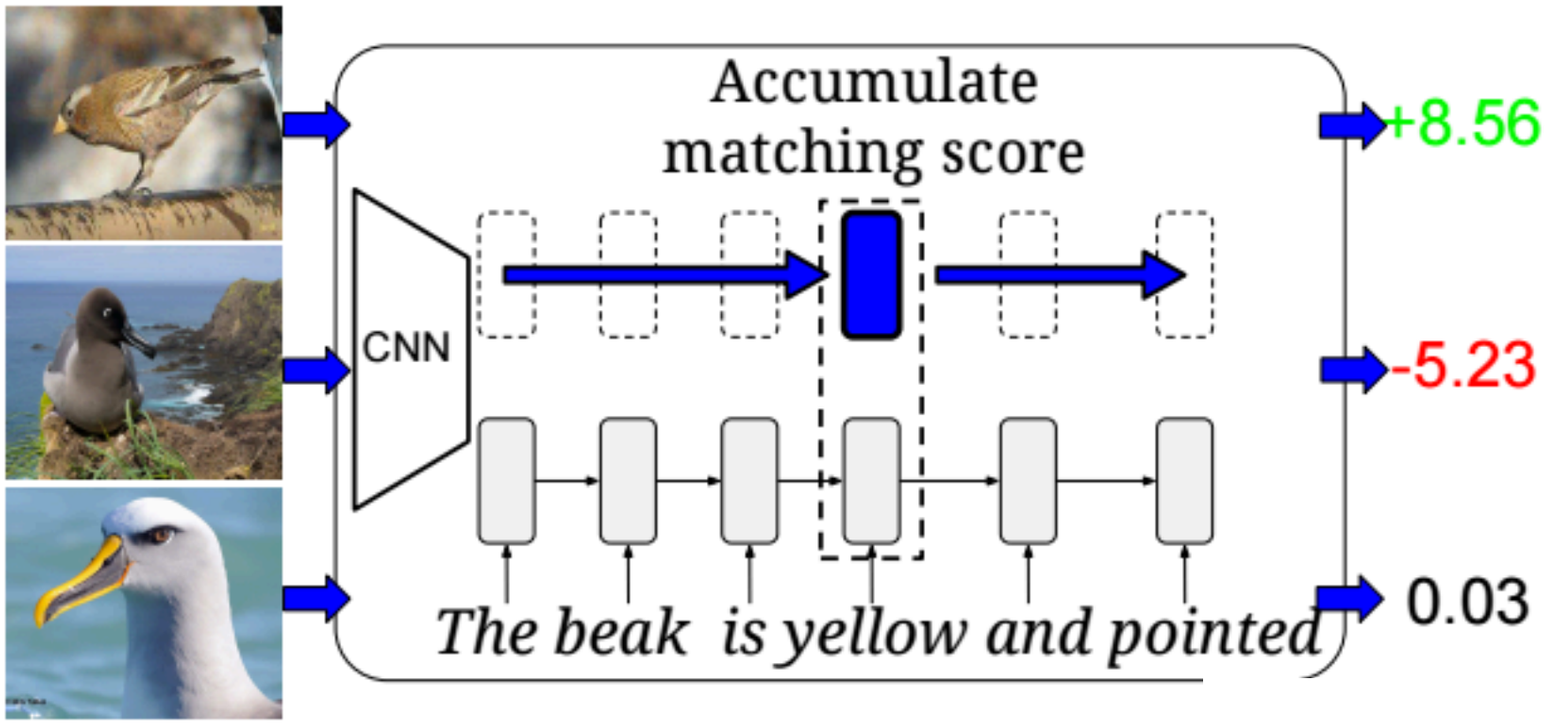
$$\frac{1}{N}\sum_{n=1}^{N}\Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n))$$

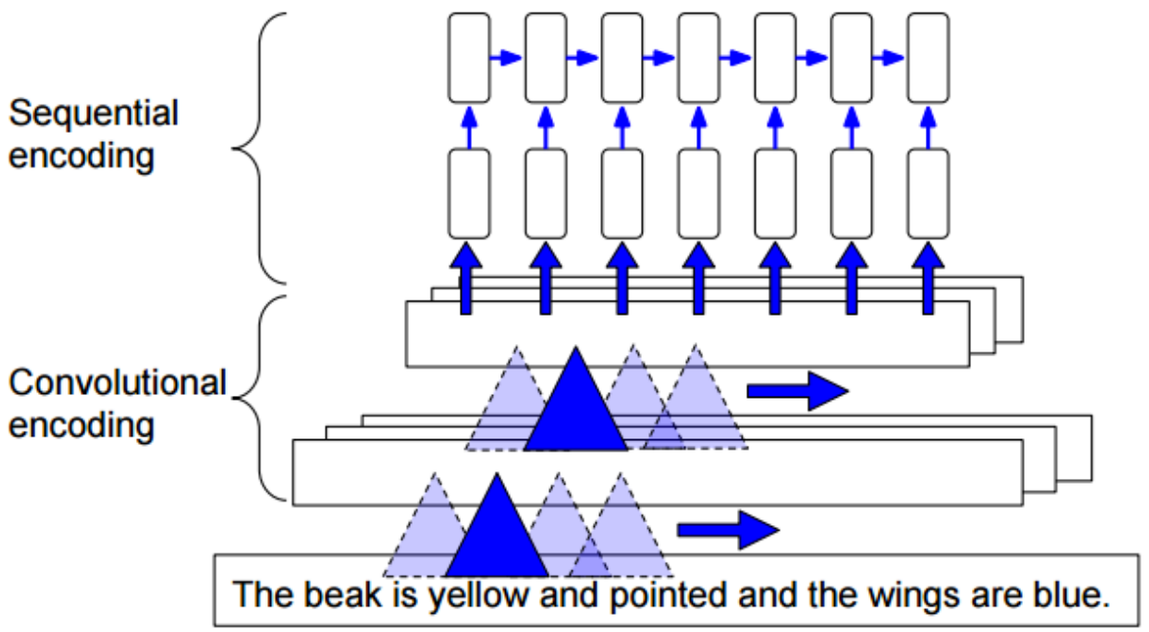
$$f_v(v) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_{t\sim\mathcal{T}(y)}[\phi(v)^T\varphi(t))]$$

$$f_t(t) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_{v\sim\mathcal{V}(y)}[\phi(v)^T\varphi(t))]$$

Reed, Scott, et al. "Learning deep representations of fine-grained visual descriptions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Text Feature Representation



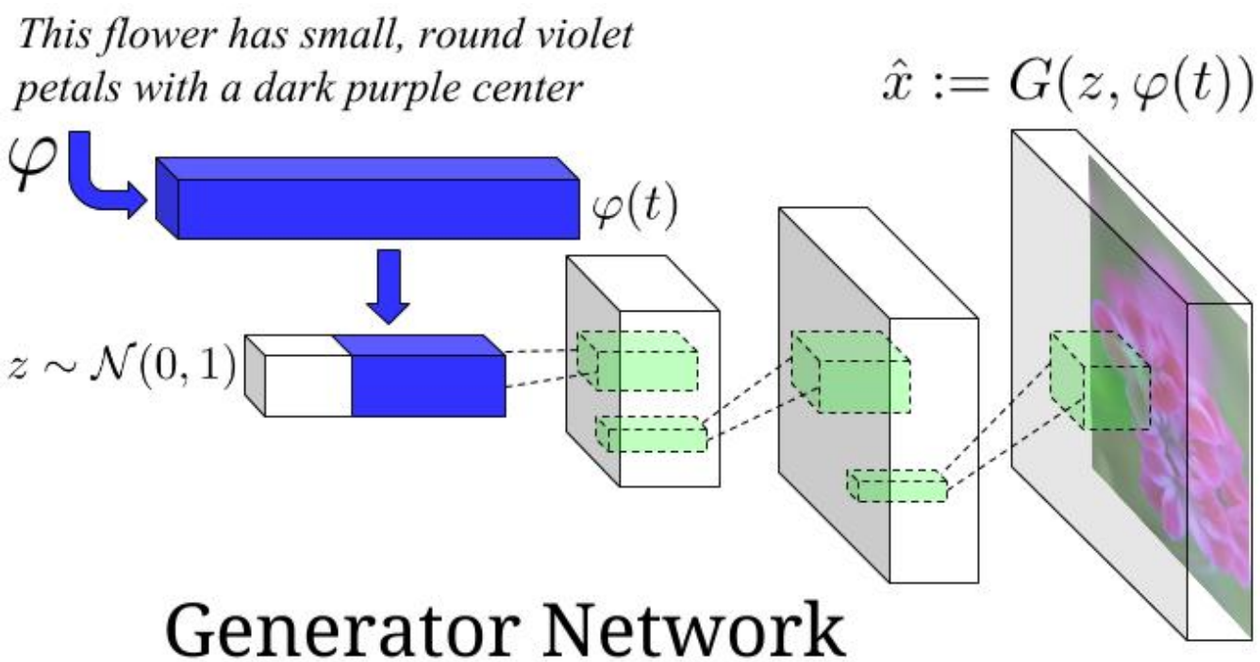


Multimodality

- A mapping between text and pixels should be learned: GAN is used
- In GAN, the generator network tries to fool adversarially trained discriminator network
  - both are conditioned on text
  - Discriminator acts as a smart adaptive loss function

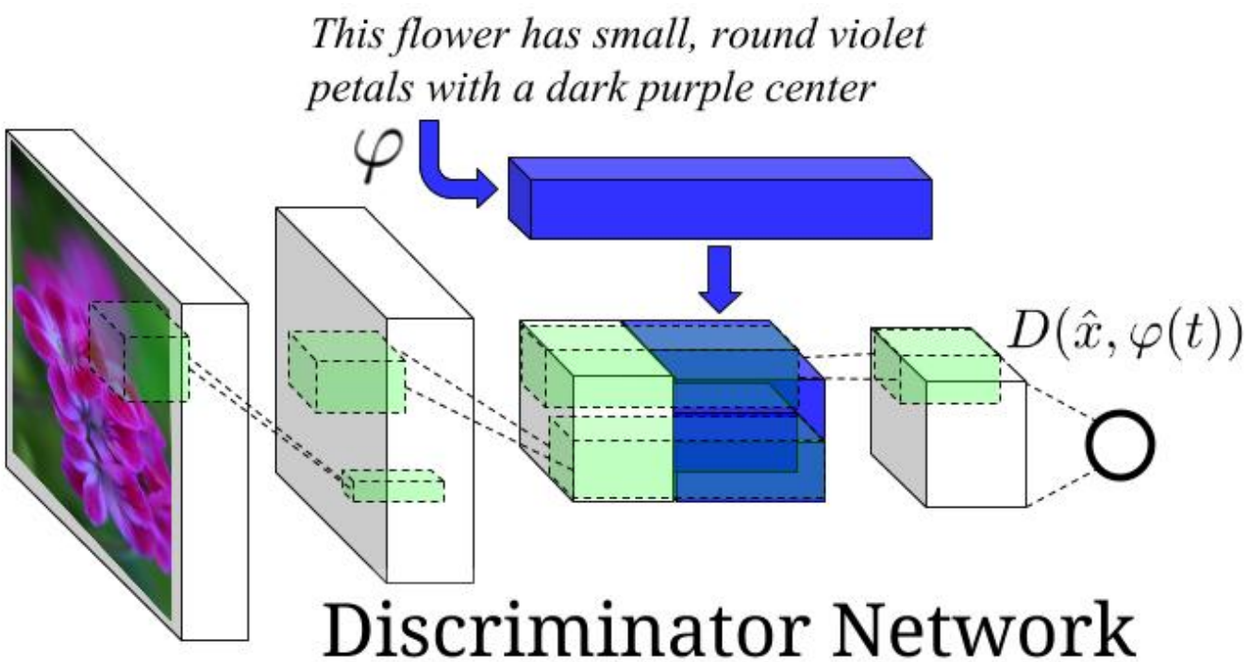
GAN

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



Generator Network

- Fully connected layer (dim-reduc.)
- Leaky ReLU
- Concatenation
- Deconvolution



Discriminator Network

- Several layers of stride-2 conv. (with spatial batch normalization)
- Leaky ReLU
- Fully connected layer (dim-reduc.) + rectification (text)
- Depth Concatenation
- conv, rectification, conv.

[**]Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

GAN - CLS

- Naive GAN : **<real img,matching text> :** unrealistic images contribute learning, **<synthetic img, arbitrary text>:** wrong class contributes learning

- GAN CLS: **GAN + <real image, mismatched text> :** should be scored as fake, an additional signal provided by discriminator

GAN - CLS

Algorithm 1 GAN-CLS training algorithm with step size α, using minibatch SGD for simplicity.

Input: minibatch images x, matching text t, mismatching $\hat{t}$, number of training batch steps S
for $n = 1$ **to** S **do**
  $h \leftarrow \varphi(t)$ {Encode matching text description}
  $\hat{h} \leftarrow \varphi(\hat{t})$ {Encode mis-matching text description}
  $z \sim \mathcal{N}(0, 1)^Z$ {Draw sample of random noise}
  $\hat{x} \leftarrow G(z, h)$ {Forward through generator}
  $s_r \leftarrow D(x, h)$ {real image, right text}
  $s_w \leftarrow D(x, \hat{h})$ {real image, wrong text}
  $s_f \leftarrow D(\hat{x}, h)$ {fake image, right text}
  $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$
  $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ {Update discriminator}
  $\mathcal{L}_G \leftarrow \log(s_f)$
  $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ {Update generator}
end for

GAN - INT

- Based on the observation that interpolations between embeddings tend to be near the data manifold, extra amount of text embeddings can be generated (although they don't have a matching text/images, they are useful for D)

- A term added to generator objective:

$$\mathbb{E}_{t_1,t_2\sim p_{data}}[\log(1-D(G(z,\beta t_1+(1-\beta)t_2)))]$$

Style Transfer

$$s \leftarrow S(x), \hat{x} \leftarrow G(s, \varphi(t))$$

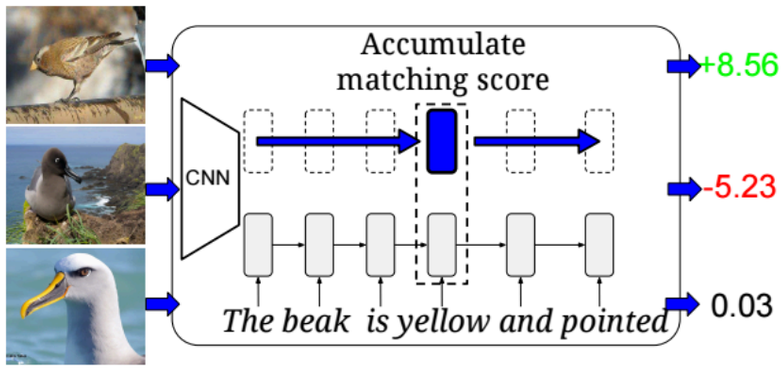
$$\mathcal{L}_{style} = \mathbb{E}_{t,z\sim\mathcal{N}(0,1)}||z - S(G(z, \varphi(t)))||_2^2$$

Experiments

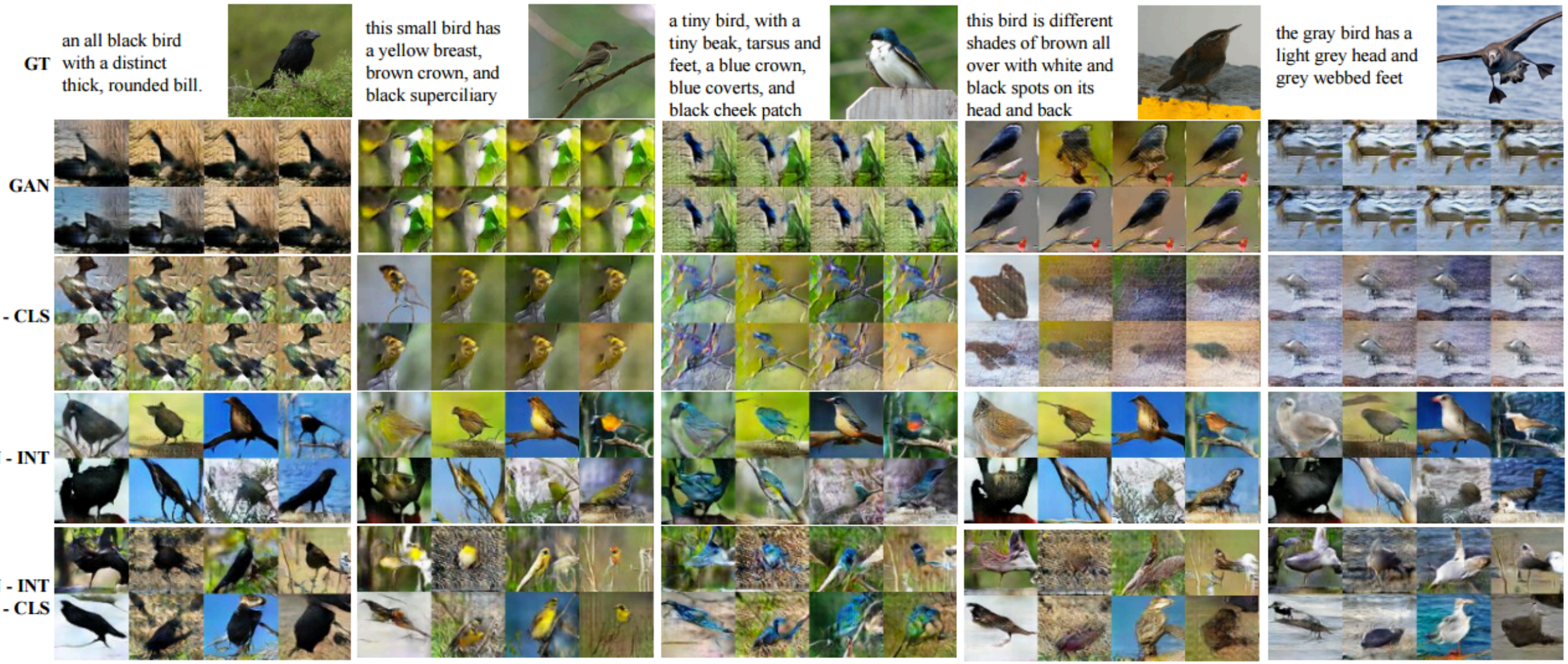
- Datasets:
    - CUB birds (11788 images, 200 classes, 5 captions per image)
        - Split to disjoint classes:150 train+val, 50 test
    - Oxford-102 flowers (8189 images, 102 categories, 5 captions per image)
        - 82 train+val, 20 test

Experiments

- Text features
    - Pre-training on deep deep convolutional-recurrent text encoder (char level) with Google LeNet image embeddings



Qualitative Results



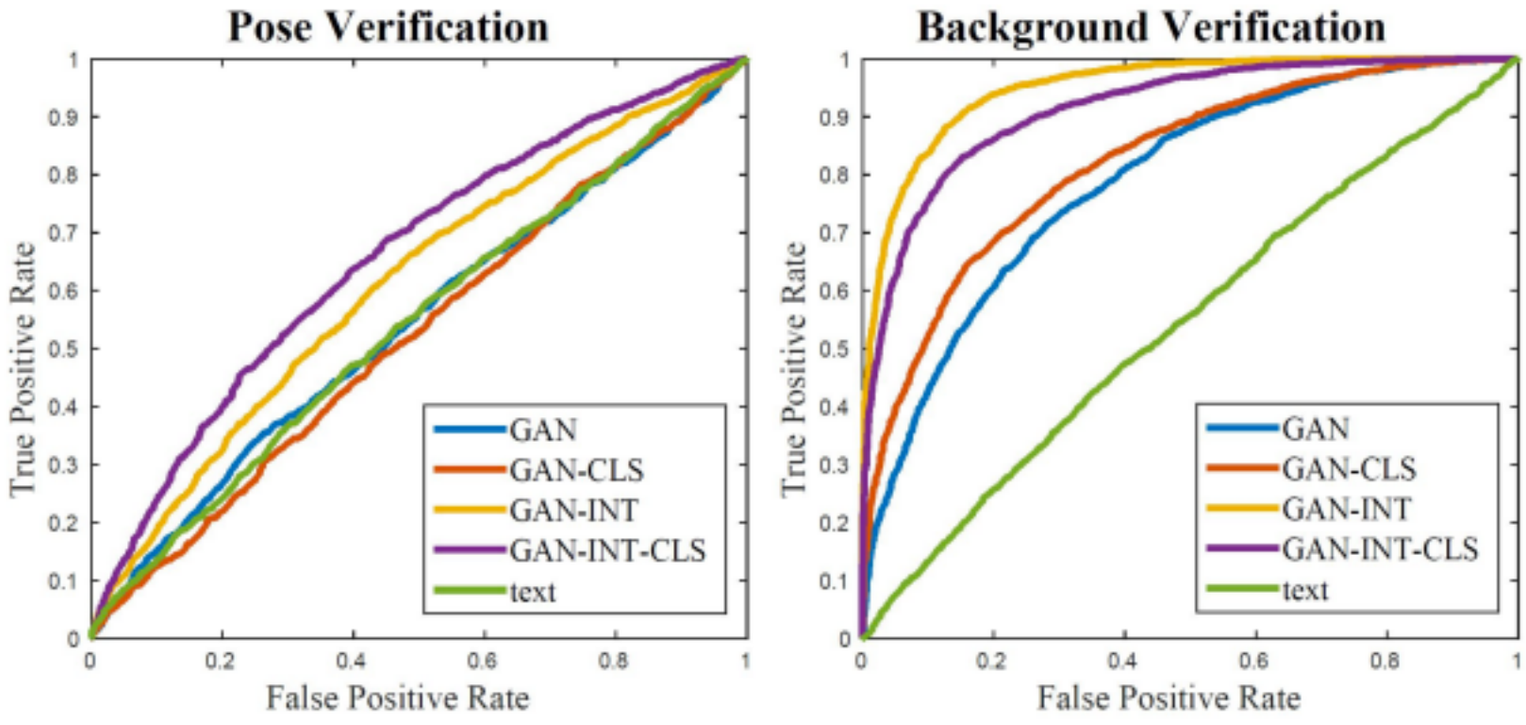
Qualitative Results

Disentangling Style and Content

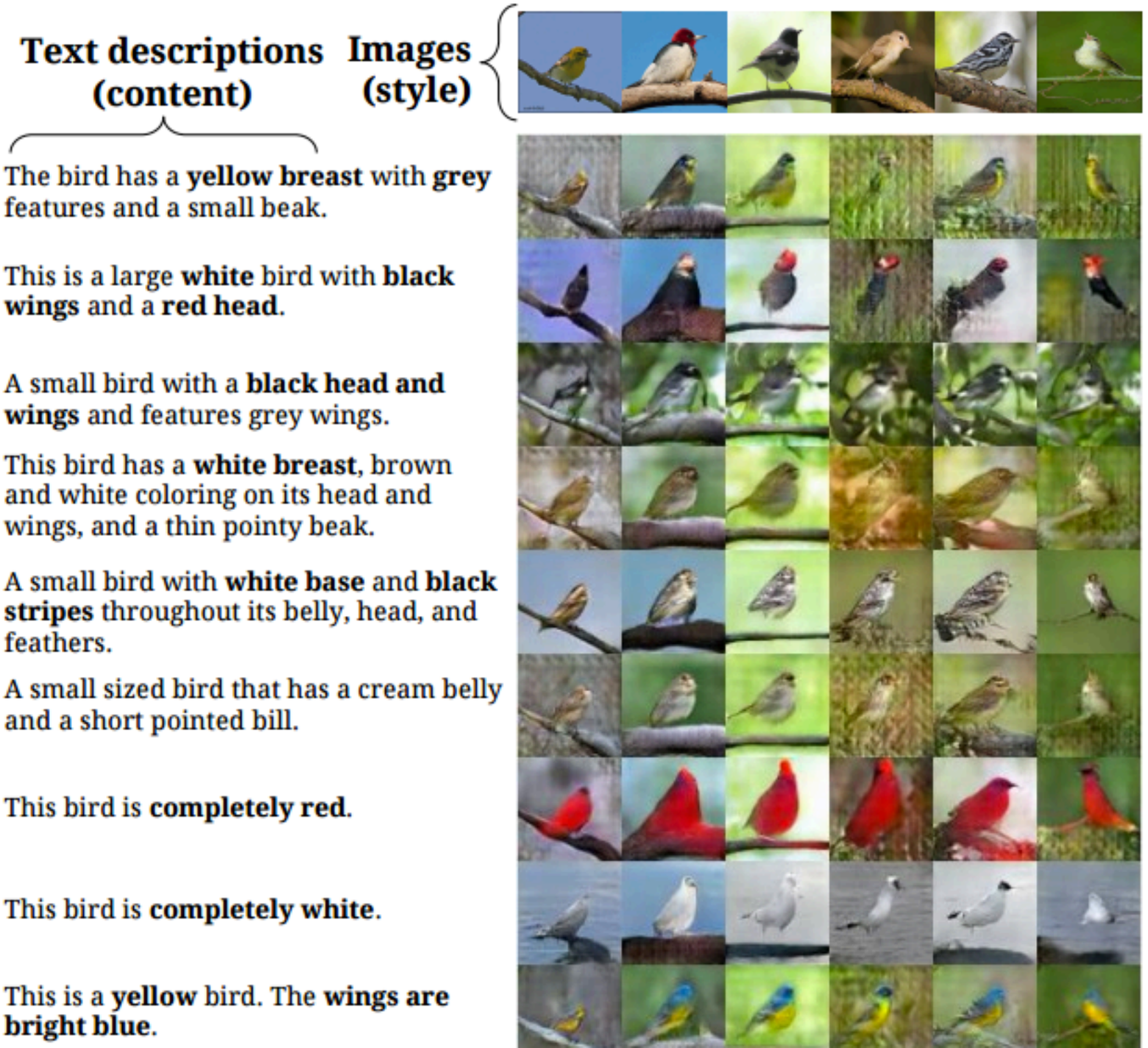
- Quantification of success is based on pose verification and background verification
- Similar pairs of images constructed for each task via K-means:
    - Avg. RGB for background color
    - Keypoint coordinates for pose

Disentangling Style and Content

- Similar and different images are fed into Style network, then cosine similarity is calculated based on the resulting encodings:



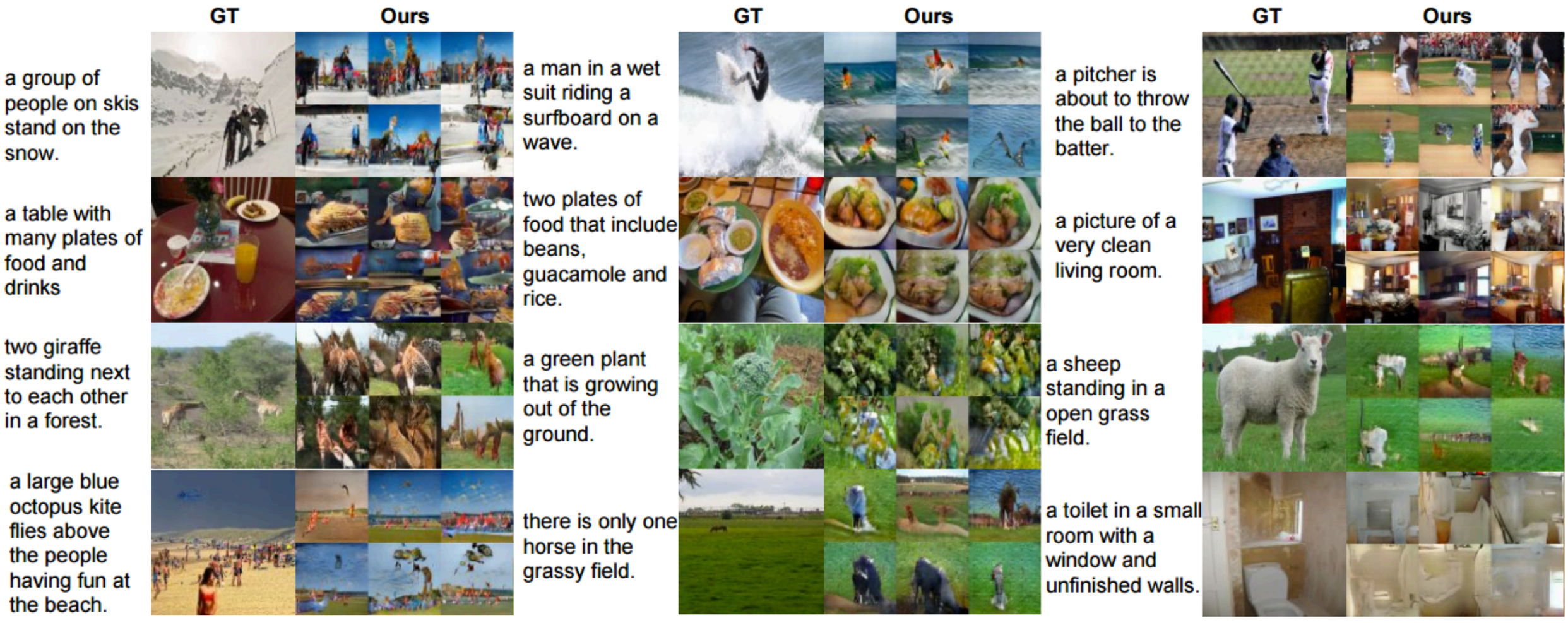
Pose and Background Style Transfer



Sentence Interpolation

Figure 8. Left: Generated bird images by interpolating between two sentences (within a row the noise is fixed). Right: Interpolating between two randomly-sampled noise vectors.

GAN CLS on MS COCO



THANKS FOR YOUR ATTENTION