Text to Image Synthesis

Andrew Drozdov

NYU Department of Computer Science

August 8, 2017

Overview

Paper Summary: Reed et al., 2016

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)

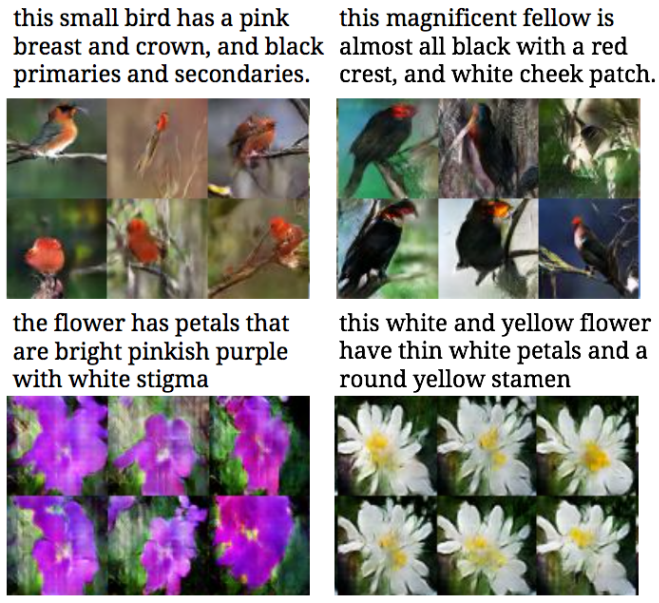


Figure 1. Examples of generated images from text descriptions. Left: captions are from zero-shot (held out) categories, unseen text. Right: captions are from the training set.

What is the problem?

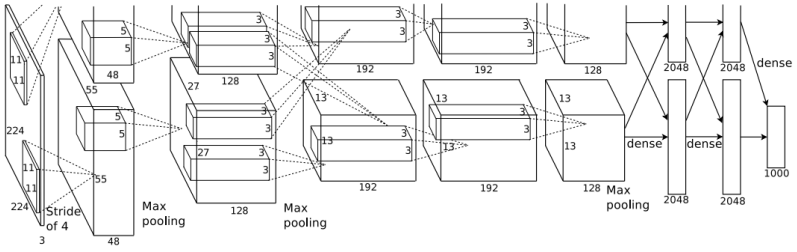
However, one difficult remaining issue not solved by deep learning alone is that the distribution of images conditioned on a text description is highly multimodal, in the sense that there are very many plausible configurations of pixels that correctly illustrate the description. The reverse direction (image to text) also suffers from this problem but learning is made practical by the fact that the word or character sequence can be decomposed sequentially according to the chain rule; i.e. one trains the model to predict the next token conditioned on the image and all previous tokens, which is a more well-defined prediction problem.

Background

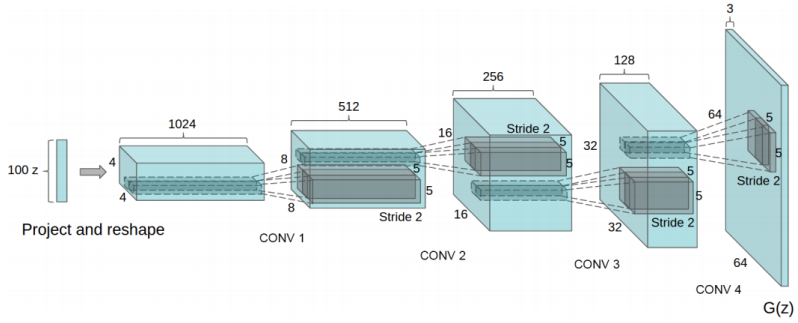
Generative Adversarial Networks (Goodfellow et al. 2014)

Play this minimax game until reaching a Nash equilibrium:

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



(Krizhevsky et al. 2015)



(Radford et al. 2015)

Joint Embedding (Reed et al. 2016)

$$\frac{1}{N}\sum_{n=1}^{N}\Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \qquad (2)$$

- v_n - image
- y_n - image class
- t_n - image description

$$f_v(v) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_{t\sim\mathcal{T}(y)}[\phi(v)^T\varphi(t))] \qquad (3)$$

$$f_t(t) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}_{v\sim\mathcal{V}(y)}[\phi(v)^T\varphi(t))] \qquad (4)$$

- ϕ - image encoder
- ψ - text encoder
- f_v - image classifier
- f_t - text classifier

Method

Model

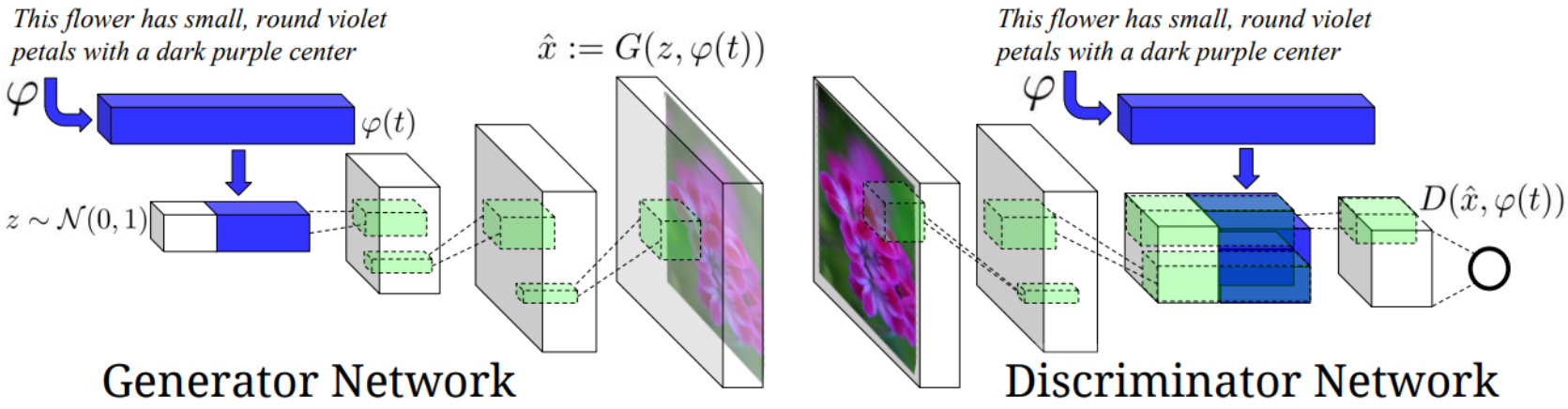


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

Joint Conditioning of Image and Description (GAN-CLS)

Easily discriminate early in training but later...

What we have:

- Real Image + Matching Description
- Fake Image + Arbitrary Description (Matching or Not Matching)

What we may need:

- Real Image + Not Matching Description (Learn to match)

Data Augmentation by Interpolation (GAN-INT)

Create additional *text embeddings* (using interpolation between existing embeddings).

Assumption: "interpolations between embedding pairs tend to be near the data manifold" (Bengio et al., 2013; Reed et al., 2014)

$$\mathbb{E}_{t_1,t_2\sim p_{data}}[\log(1 - D(G(z, \beta t_1 + (1-\beta)t_2)))] \quad (5)$$

Style Transfer

The text encodes the content; the image encodes the style.
The text is easy to manipulate; the image not so.

Convolutional Style Network:

$$S : R^D \rightarrow R^Z$$

Style Loss:

$$\mathcal{L}_{style} = \mathbb{E}_{t,z\sim\mathcal{N}(0,1)}||z - S(G(z, \varphi(t)))||_2^2 \qquad (6)$$

Style Transfer:

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

Results

Datasets

- Birds - CUB dataset
  - 11,788 images
  - 200 categories
  - 150 train+val classes
  - 50 test classes
- Flowers - Oxford-102 dataset
  - 8,189 images
  - 102 categories
  - 82 train+val classes
  - 20 test classes
- 5 captions per image
- Augment images with random transformation: scale, crop, etc.

Model Summary

- GAN
- GAN-CLS (real image, fake label)
- GAN-INT (text interpolation)
- GAN-INT-CLS
- Style Transfer

Zero Shot: Birds

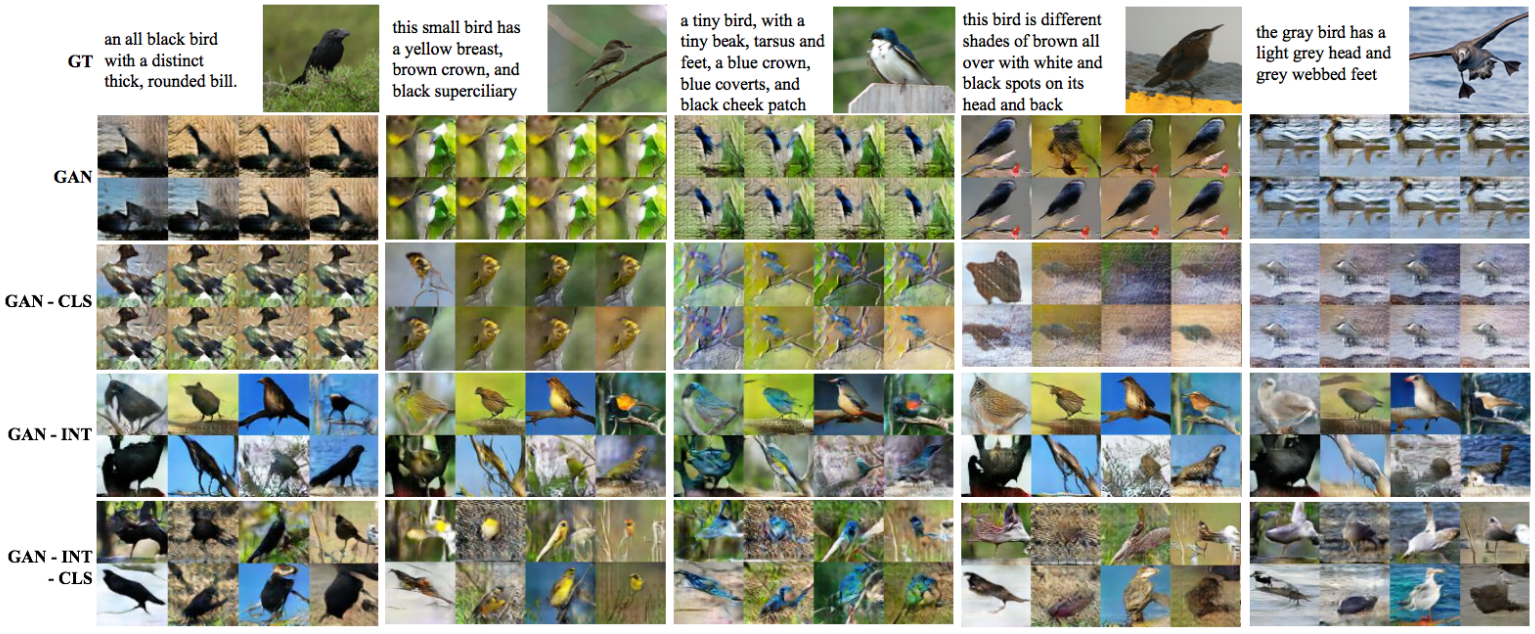


Figure 3. Zero-shot (i.e. conditioned on text from unseen test set categories) generated bird images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. We found that interpolation regularizer was needed to reliably achieve visually-plausible results.

Zero Shot: Flowers

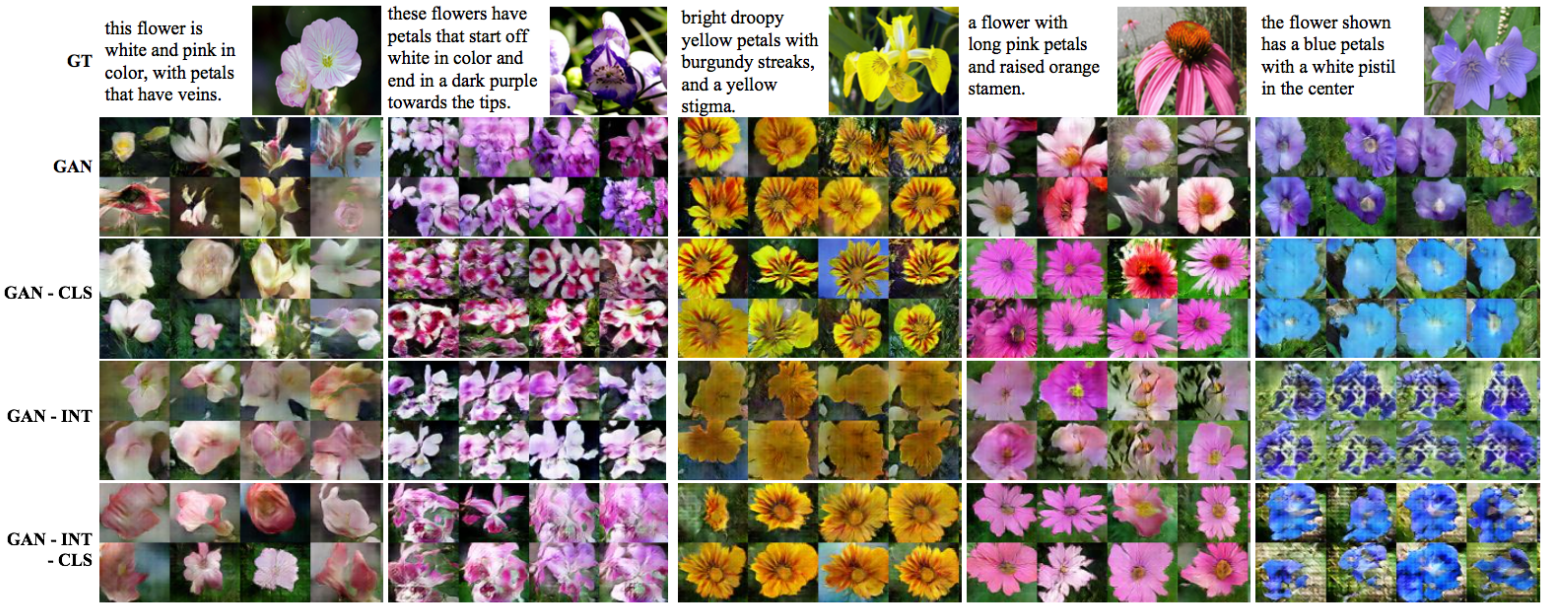
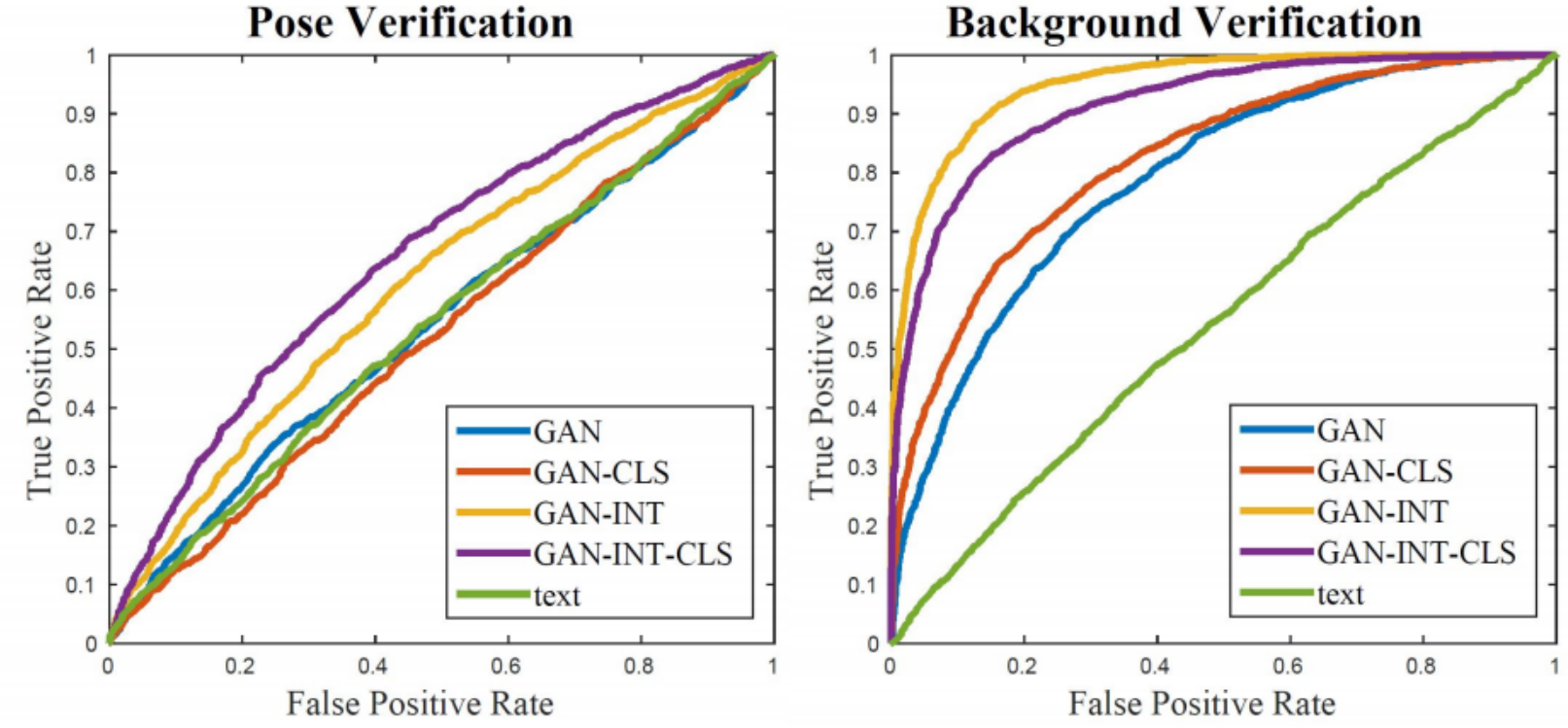


Figure 4. Zero-shot generated flower images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. All variants generated plausible images. Although some shapes of test categories were not seen during training (e.g. columns 3 and 4), the color information is preserved.

Style Transfer: Birds

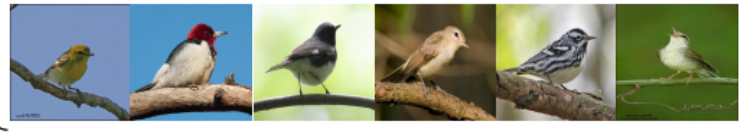
- 100 Clusters (K-means)
- Background: RGB
- Pose: 6 keypoint coordinates (beak, belly, breast, crown, forehead, and tail)



Style Transfer: Birds

Text descriptions (content)

Images (style)



The bird has a **yellow breast** with **grey** features and a small beak.

This is a large **white** bird with **black wings** and a **red head**.

A small bird with a **black head and wings** and features grey wings.

This bird has a **white breast**, brown and white coloring on its head and wings, and a thin pointy beak.

A small bird with **white base** and **black stripes** throughout its belly, head, and feathers.

A small sized bird that has a cream belly and a short pointed bill.

This bird is **completely red**.

This bird is **completely white**.

This is a **yellow** bird. The **wings are bright blue**.

Text 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.

Beyond birds and flowers

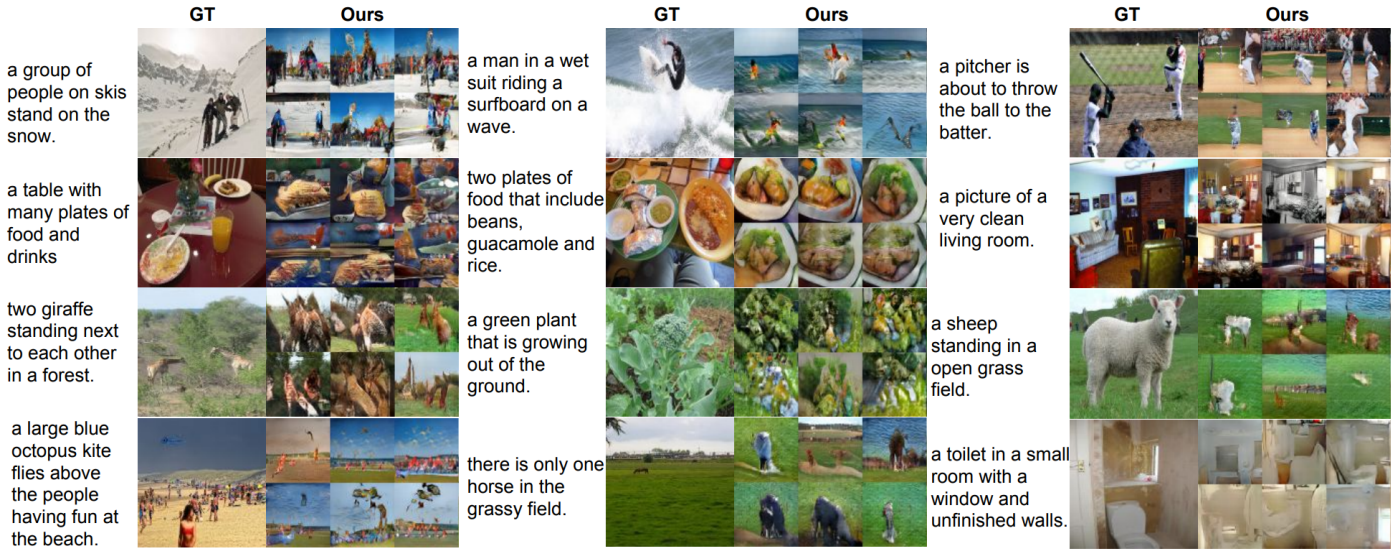


Figure 7. Generating images of general concepts using our GAN-CLS on the MS-COCO validation set. Unlike the case of CUB and Oxford-102, the network must (try to) handle multiple objects and diverse backgrounds.

Discussion

Question for the Audience

What is the closest representation to an image, in pure 100% text?

Followup

- Can we do text to image synthesis without z?
- Can we do text to image synthesis with longer descriptions?
- Can we do text to text synthesis (inverse summarization)?

References

1. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems. 2014.
2. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
3. Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015). (ICLR 2016)

References

4 Reed, S., Akata, Z., Lee, H., and Schiele, B. Learning deep representations for fine-grained visual descriptions. In CVPR, 2016.

5 Gauthier, J. Conditional generative adversarial nets for convolutional face generation. Technical report, 2015.

6 Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better mixing via deep representations. In ICML, 2013.

References

7 Reed, S., Akata, Z., Lee, H., and Schiele, B. Learning deep representations for fine-grained visual descriptions. In CVPR, 2016.

8 Reed, S., Sohn, K., Zhang, Y., and Lee, H. Learning to disentangle factors of variation with manifold interaction. In ICML, 2014.