cFineGAN: Unsupervised multi-conditional fine-grained image generation

Gunjan Aggarwal[*]
Adobe Inc, Noida, India
guaggarw@adobe.com

Abhishek Sinha[*†]
Stanford University
a7b23@stanford.edu

Abstract

We propose an unsupervised multi-conditional image generation pipeline: cFineGAN, that can generate an image conditioned on two input images such that the generated image preserves the texture of one and the shape of the other input. To achieve this goal, we extend upon the recently proposed work of FineGAN [8] and make use of standard as well as shape-biased pre-trained ImageNet models. We demonstrate both qualitatively as well as quantitatively the benefit of using the shape-biased network. We present our image generation result across three benchmark datasets- CUB-200-2011[9], Stanford Dogs[5] and UT Zappos50k[10].

1 Introduction

Recent developments in deep learning and generative adversarial networks(GAN) have made it possible to generate realistic looking images of high resolution. The image generation techniques generally come in two forms : i) Unconditional image generation – starting from a noise vector, the generator generates an image [2]. ii) Conditional image generation – given a condition, the aim is to generate an image adhering to some condition [6, 4, 11].

While a lot of work has been done in the domain of single-conditional image synthesis, the domain of unsupervised multi-conditional image synthesis is relatively new. We aim to generate an image conditioned on two inputs such that the generated image contains texture of the first and shape of the second conditioned image.

2 Approach

Our work is based upon the recently released work FineGAN. The authors propose a GAN based framework that learns to disentangle the background, shape and texture of an image in an unsupervised manner. The network generates an image conditioned on input background, shape and texture codes.

Our pipeline takes in two images I_1 and I_2 as input and generates an output image(O). The pipeline consists of three steps - i) Compute the texture code(T) that describes the first input image(I_1), ii) Compute the shape code(S) that describes the second input image(I_2), and iii) Feed the computed codes(T and S) as input to the pre-trained FineGAN network to get the desired output O.

To compute the codes(T and S), we take a trained FineGAN and iterate over all the possible combinations of shape and texture codes for 10 different noise vectors(different noise vector lead to different orientations of the generated image). We denote as G the set of all such generated images. To compute the texture code(T), we compute the nearest neighbour of I_1 amongst G in the embedding

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[*]Authors contributed equally
[†]work done while author was working at Adobe, India

space of ImageNet [7] pre-trained ResNet50 model [3]. The embedding space is defined by the Global Average Pooling layer output of the ResNet50 model.

We repeat the same process for image I_2 to compute the shape code(S) with the exception of using a shape biased pre-trained ResNet50 network. The motivation for using a shape biased network stems from [1], where the authors show that the ImageNet trained models are biased towards texture details of image. The authors use stylized variants of the ImageNet dataset to train the network resulting in a shape-biased network. We hypothesize that the shape biasness of the network would allow it to better capture the shape details of the image, leading to correct identification of the shape code of an input image. We verify both quantitatively and qualitatively this design choice in the following section.

3 Results and Discussions

Figure 1: cFineGAN results - columns 1-3 show results for CUB-200-2011, 4-6 for UT Zappos50k and 7-9 for Stanford Dogs datasets respectively.

Our cFineGAN results over the three datasets - CUB-200-2011 [9], UT Zappos50k [3] [10] and Stanford Dogs [5] are shown in figure 1. Additional results can be found in the Appendix section.

ResNet Model	Accuracy (%)
Standard ResNet50	70.75
Shape-biased ResNet50	86.90

Table 1: Quantitative analysis of models for shape code prediction of generated images.

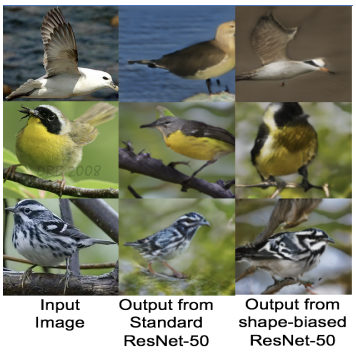

Figure 2: Nearest neighbour image for standard and shape-biased networks.

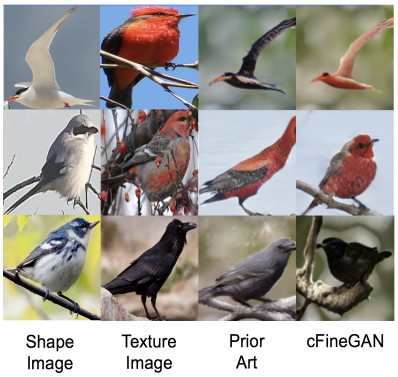

Figure 3: Qualitative comparison of cFineGAN against the prior art.

To quantitatively evaluate the benefit of using a shape-biased pre-trained model for extracting the shape code, we compute the nearest neighbour in the embedding space for each generated image in G. We define accuracy as the fraction of times the query image and its nearest neighbour have the same shape code. As the shape code of all the generated images is known, we can compute this metric. Table 1 shows that the accuracy achieved by the shape-biased model is much better than that of a standard model. Some qualitative results have been shown in figure 2.

We baseline our method against the approach mentioned in [8] where the authors train classifiers over the domain of generated images to predict the shape and texture codes given image as an input. Since the classifier is trained over the domain of generated images but is expected to predict the codes of natural images during evaluation time, the huge domain shift encountered between train and test settings lead to incorrect outputs. We show some qualitative comparisons against this baseline in figure 3. As can be seen cFineGAN better captures the shape and texture details of input images.

[3] We trained our own FineGAN model over this dataset

References

[1] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. 2015.

[6] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[8] K. K. Singh, U. Ojha, and Y. J. Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. *arXiv preprint arXiv:1811.11155*, 2018.

[9] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010.

[10] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

4 Appendix

4.1 Additional Results

We show additional results over the three datasets in Figures 4, 5 and 6.

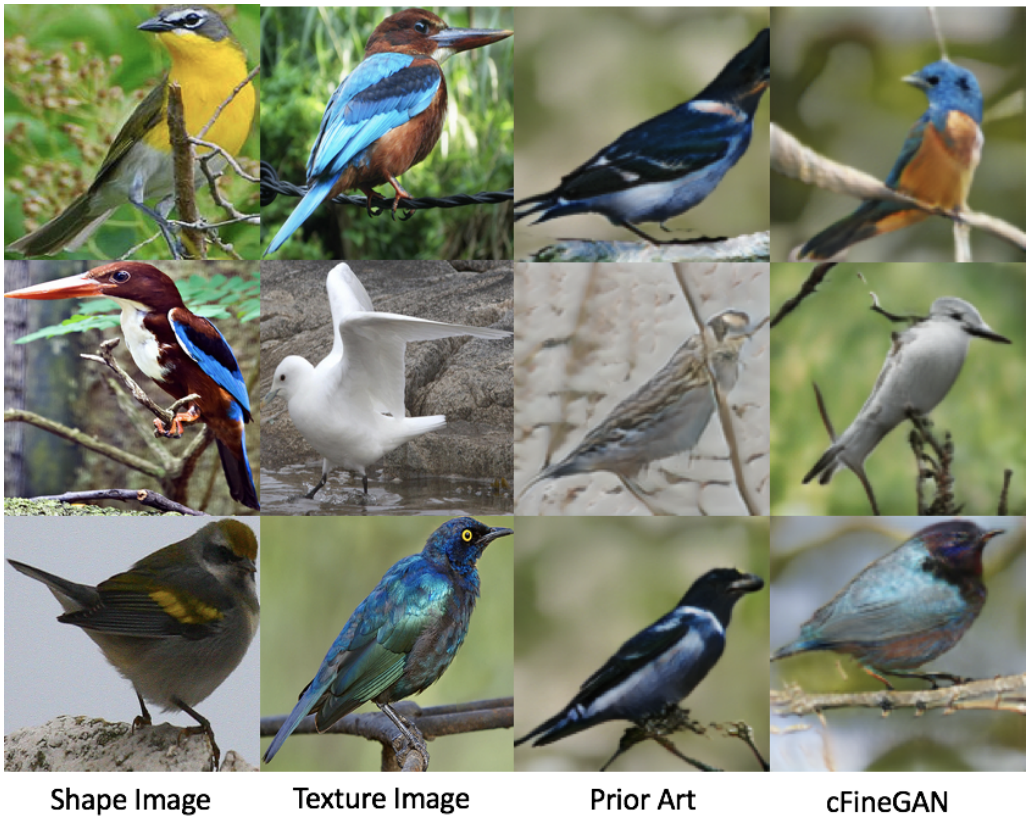

Figure 4: Additional results over the CUB-200-2011 dataset.

Figure 5: Additional results over the Stanford dogs dataset.

Figure 6: Additional results over the UT Zappos50k dataset.