# CMU RI 16-995: Independent Study: Listening Generative Models

Advisor: Jean Oh
The Robotics Institute, Carnegie Mellon University

Fall 2018

## 1 Course Description

**Objective:** The goal of this independent study is to develop an algorithm for generating and interactively modifying an image based on an input sequence of language descriptions. The idea is related to Interactive Generative Adversarial Networks (i-GANs) [20] where a user can modify an image interactively using a graphical user interface (GUI). Instead of using a GUI to interact with the system, we propose a system that listens to natural language descriptions to interactively (re-)generate images.

**Background research:** This study requires solid understanding of basic deep learning approaches including feed-forward and recurrent neural networks that can be reviewed in textbooks such as [4] (Part I and II). As background research for this study, a literature survey will be conducted on recent progress on image synthesis, in particular Generative Adversarial Networks (GANs) models [5, 10, 18, 20, 7, 21, 13, 14, 22, 16, 6, 15, 17, 11, 8, 1, 19], Variational Autoencoders (VAEs) [9], and Flow-based models [2].

**Datasets:** For this study, we plan to use sketches rather than photo-realistic imagery. This study will utilize publicly available datasets including the human sketch dataset [3] and Google Quick! Draw dataset [1]. The student is

[1] https://github.com/googlecreativelab/quickdraw-dataset

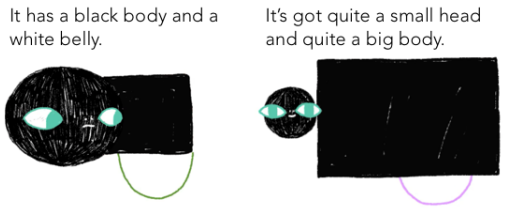


Figure 1: An motivational example of incremental hand drawing [12]

expected to do further research on additional datasets as needed. Additionally, the student will be responsible for collecting language description data over the chosen sketch dataset using Amazon Mechanical Turk. The cost of this new data collection will be paid by the course advisor.

**Experiments:** On one set of experiments, we will evaluate the image synthesis on the final image only as follows. The algorithm will be evaluated on the duel-learning manner. We first train a multi-class classifier that maps a sketch to a label using the sketch dataset. Next, we use this classifier to classify the generated sketch from our system. Finally, we compare the accuracy of the classifier on the generated images against that of human drawings, e.g., if the accuracy is close to that on the test dataset that includes human drawings then it indicates comparable performance.

On the second set of experiments, we will create an online/offline game version of our system and collect user statistics, e.g., how long each participants engage in the game and their reactions, for future study.

**Evaluation:** The student and the advisor will co-author a technical paper that includes a formal problem definition, related work, detailed technical approach, experiments and results, and conclusion and future directions. The report will be written incrementally we we keep track of the progress.

# Reading list


[1] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *ArXiv e-prints*, November 2017.

[2] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. *CoRR*, abs/1410.8516, 2014.

[3] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10, 2012.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2017. http://www.deeplearningbook.org/.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] David Ha and Douglas Eck. A neural representation of sketch drawings. *CoRR*, abs/1704.03477, 2017.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[9] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

[10] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5969–5978, 2017.

[11] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *CoRR*, abs/1706.00409, 2017.

[12] Songeun Lee. My favorite animals. https://vimeo.com/17456055, 2011.

[13] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[14] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 217–225. Curran Associates, Inc., 2016.

[15] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. Chatpainter: Improving text to image generation using dialogue. *CoRR*, abs/1802.08216, 2018.

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. arXiv preprint arXiv:1711.11585.

[17] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017.

[18] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915, 2017.

[19] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017.

[20] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[21] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.

[22] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems 30*. 2017.