**A Thesis Proposal**
**on**

# Photographic Text-to-Image Synthesis via Multi-turn Dialogue using Attentional GAN

For Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Computer Engineering Awarded by
Pokhara University

**Submitted By**

**SHIVA KUMAR SHRESTHA**
ME Computer (MECE)
15957

Department of Graduate Studies
Nepal College of Information Technology
**Balkumari, Lalitpur, GPO Box 8975 Kathmandu**
**Nepal**

Jan, 2019

# ABSTRACT

This proposal attempts to solve a novel task - interactive image generation via conversational language, where users can guide an agent to modify images features via multi-turn dialogue in natural language. In the initial step, user inputs text-description and the agent generates an initial image and then in each dialogue turn, the agent takes a source image and a natural language description from the user as the input and generates a target image following the textual description. For this task, an Attentional Generative Adversarial Network (AttnGAN) framework will be used, which applies a neural state tracker to encode both source image and textual descriptions, and generates high-quality images in each dialogue turn. To achieve better region specific text-to-image generation, a new attention mechanism model will be used. Experiments will be on the three datasets, including quantitative evaluation and user study to show that this approach performs best in both image quality and text-to-image consistency.

*Keywords: Realistic Image Synthesis, Text-to-Image Synthesis, Image Generation, Generative Adversarial Network, GAN, Deep Learning*

# Contents

# LIST OF FIGURES

# Chapter 1

# INTRODUCTION

## 1.1 BACKGROUND

The promise of deep learning is to discover rich, hierarchical models that represents probability distribution over the kinds of data encountered in artificial intelligence applications [1]. A key challenge in image understanding is to correctly relate natural language concepts to the visual content of images. In recent years there has been significant progress in learning visual-semantic embedding [11], e.g. zero-shot learning and automatically generating image captions for general web images.

The process of generating a photo-realistic image from the text, is one of the important problems [1] and is the emerging technology and has various tremendous applications including photo editing, computer-aided design, interactive computational graphic design, image fine-tuning etc. The process of generating truly plausible looking image has not been easy. The majority of various advanced methods do not produce photo-realistic details that describe given text description. The Generative Adversarial Network (GAN) has shown the promising result in synthesizing the real world images. Conditional GANs are able to generate images that are extremely related to the text meanings but it is very hard to train GAN to generate high-resolution photo-realistic images from text descriptions. Here adding more upsampling layers in the state-of-art GAN models for generating high-resolution images results in training instability and produces nonsensical outputs. The main difficulty for generating high-resolution images by GANs is that supports of natural image distribution and implied model distribution may not overlap in high dimensional pixel space. This problem is extremely severe when the image resolution increases.

For the text-to-image generation, the limited numbers of text-image pairs often result in sparsity in the text conditioning and such sparsity makes it difficult to train GAN. Thus conditioning augmentation technique is used to encourage smoothness in the latent conditioning. This method increases the diversity of the synthesized images. The basic concept of this thesis work is to convert the given text description into vectors and generating the images from the given vectors using GAN.

Typically, a GAN consists of two networks: generator and discriminator [4]. A Generator (the one who produces interesting data from noise), and a Discriminator (the one who detects fake data fabricated by the Generator) as shown in the figure below. The duo is trained iteratively:

- The Discriminator is taught to distinguish real data (Images/Text whatever) from that created by the Generator. At this step, the Generator is not being trained only the Discriminator's detective skills are improved.
- The Generator is trained to produce data that can sufficiently fool the (now-improved) Discriminator. The random input ensures that the Generator keeps coming up with novel data every-time essentially acting as inspiration.

The key insight is in the dual-objective: As (and because) the discriminator becomes a better detective, the generator becomes a better faking-artist. After a sufficient number of epochs, the Generator can create surprisingly realistic images:

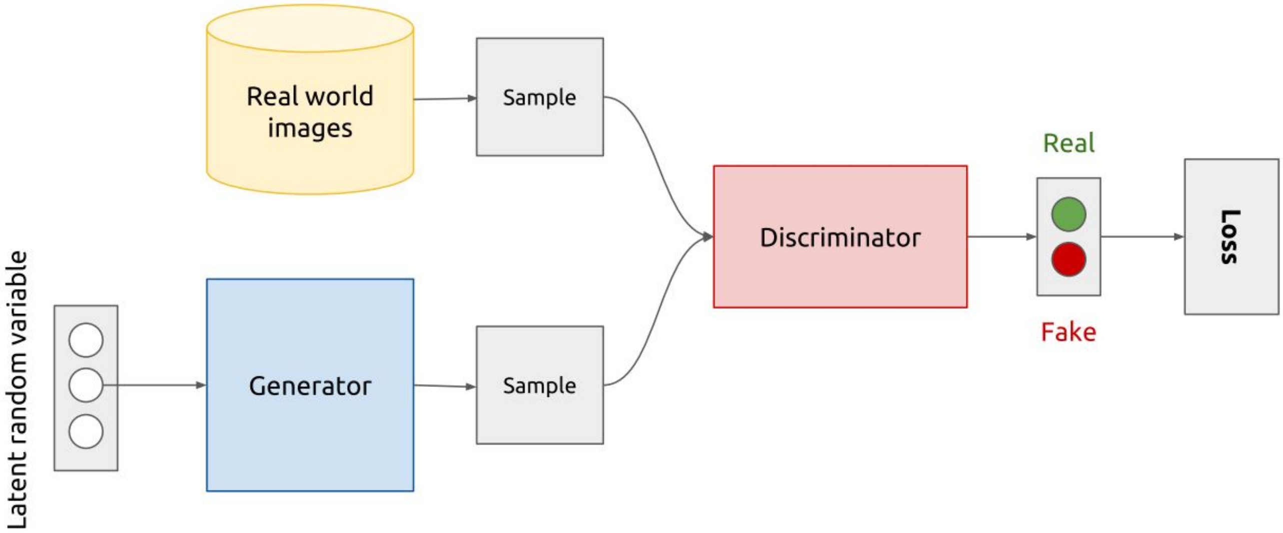

*Figure 1.1: GAN Framework*

## 1.2 PROBLEM STATEMENTS

Most recently proposed text-to-image synthesis methods are based on GANs. A commonly used approach is to encode the whole text description into a global sentence vector as the condition for GAN-based image generation [3, 6, 10]. Although impressive results have been presented, stacked GAN lacks to use conversational text which can help to generate photo-realistic image with visual details.

To address this issue, my research work attempts to solve the generation of high-resolution photo-realistic images from the given text description via multi-turn dialogue.

## 1.3 MOTIVATION

Nowadays, the term GAN is a hot topic which is introduced by Ian Goodfellow in 2014 at his Ph.D. Thesis. Before the discovery of GAN, there were various approaches such as Maximum Likelihood Estimation, Variational Autoencoders, etc. used for modeling. Many researchers are working in this field to build a better model which can generate or predict images, videos, anime character, object, etc and this could be a model to transfer our imagination to computer-generated objects. So, the concept of using GAN to generate an image as per human's conversation motivated me to do this thesis.

## 1.4 RESEARCH OBJECTIVE

A skilled painter or highly trained craftsmen can even create paintings that approach photo-realism [8]. Can we train computational models that have this ability? Given a semantic via conversational text of imagination, can an artificial system synthesize an image that depicts image and looks like a real photo?

The objective of this thesis work is as follows:

- To generate a photo-realistic image from given text description via multi-turn dialog using Attentional GAN.

## 1.5 SCOPE OF THESIS

- Text to Image Synthesis,
- High-resolution Image Synthesis,
- Intelligent Image Manipulation,
- Face Synthesis,
- Image Editing, etc.

# Chapter 2

# LITERATURE REVIEW

Researches and studies have been performed earlier on GANs. Research papers/works that is closely related to the proposed work are as follows:

**1. Generative Adversarial Nets for Text-to-Image Synthesis:** Generating high-resolution images from text descriptions, though very challenging, is important for many applications such as art generation and computer-aided design. Recently, great progress has been achieved in this direction with the emergence of deep generative models [12, 16, 19]. Goodfellow et at. [12] proposed a new framework for estimating generative models via an adversarial process, in which they simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G.

**2. Image Generation and Editing:** Language-based image editing [21, 22] is a task designed for minimizing labor work while helping users create visual data. Specifically, systems that can perform automatic image editing should be able to understand which part of the image that the user is referring to. This is a very challenging task, which requires a comprehensive understanding of both natural language and visual information. Cheng et al. [20] proposed a model for image editing via conversational language.

**3. Image-to-Image Synthesis:** Image-to-image translation [7, 9, 10] is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. Zhu et al. [7] presented an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples.

**4. Realistic Image Synthesis with Stacked Generative Adversarial Networks:** Many works have been proposed to use multiple GANs [3, 6, 13] to improve sample quality to model details of natural images. Wang et al. [23] utilized a structure GAN and a style GAN to synthesize images of indoor scenes. Yang et al. [17] factorized image generation into foreground and background generation with layered recursive GANs. Huang et al. [13] added several GANs to reconstruct the pre-trained discriminative model were unable to generate high-resolution images with photo-realistic details. Durugkar et al. [24] used multiple discriminators along with one generator to increase the chance of the generator receiving effective feedback. However, all discriminators in their framework are trained to approximate the image distribution at a single scale. Some methods [6, 13, 24] follow the same intuition as StackGAN++ [3]. The authors of LAPGAN (Denton et al., 2015) motivated to develop an alternative approach to iteratively upscale low resolution generated images which can be modeled more reliably i.e. scale up GANs using CNNs [15].

**5. Hierarchically-nested Adverserial Network for High-Quality Image Generation:** Zhang et al. [2] presented a novel method to deal with the challenging task of generating photographic images conditioned on semantic image descriptions. This paper pays attention to two major difficulties for text-to-image synthesis with GANs. The first is balancing the convergence between generators and discriminators [12, 19], which is a common problem in GANs. The second is stably modeling the huge pixel space in high-resolution images and guaranteeing semantic consistency [6]. An effective strategy to regularize generators is critical to stabilize the training and help capture the complex image statistics [13]. Dong et al. [5] proposed a way of synthesizing realistic images directly with natural language description. Their model synthesizes images to meet two requirements: 1) being realistic while matching the target text description; 2) maintaining other image features that are irrelevant to the text description.

**6. Story Visualization:** Learning to generate meaningful and coherent sequences of images [18] from a natural language story is a challenging task that requires understanding and reasoning in both natural language and images. Li et al.[14] proposed a Story Visualization task. After giving a multi-sentence paragraph, the story is visualized by generating a sequence of images, one for each sentence.

**7. Conditional GANs:** Isola et al. [9] investigated conditional adversarial networks as a general-purpose solution to image-to-image translation problems. They demonstrated that this approach was effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks.

All the above Papers have focused on the generation of artificial data/images. Works done till now on this topic can be used as a reference to design an agent to generate a photo-realistic image using multi-turn dialogue of the user.

# Chapter 3

# RESEARCH METHODOLOGY

## 3.1 BACKGROUND

Synthesizing high-quality images from GANs is challenging and has many practical applications, including interactive image generation, anime character generation, three-dimensional object generation, image editing, face aging, human pose estimation, photo inpainting, video generation/prediction, etc. Photographic text-to-image synthesis, in generative research, aims to learn a mapping from a semantic text space to a complex RGB image space.

## 3.2 PROPOSED SOLUTION

For the simplicity only two steps shown in figure 3.1. The attention module automatically retrieves the dialogue context for generating different sub-regions of the image. And the DMS function provides a fine-grained image-text matching loss. The target image will be generated through the model with a forward pass.

To full-fill objective of the proposed solution, a novel model will be introduced to generate new images by inputting dialogue description, while preserving coherency to the natural language description, visual quality, and naturalness. The context encoder encodes the user response(s) and passes it to the state tracker. The state tracker then aggregates the representation with the dialogue history from previous turns. Base on the joint representation of user response(s) and previous intermediate images, the image generator produces a new image.

### 3.2.1 ATTENTION MODULE

Attention Module is used to perform compositional mapping [1, 23], i.e., enforcing the model to produce regions and associated features that conform to the textual description. First, the user feedback is converted to the common semantic space via a transform layer. Then, a word-context vector is computed for each sub-region of the image based on its hidden features. Finally, this module will produces a word-context matrix which is passed to the neural tracker

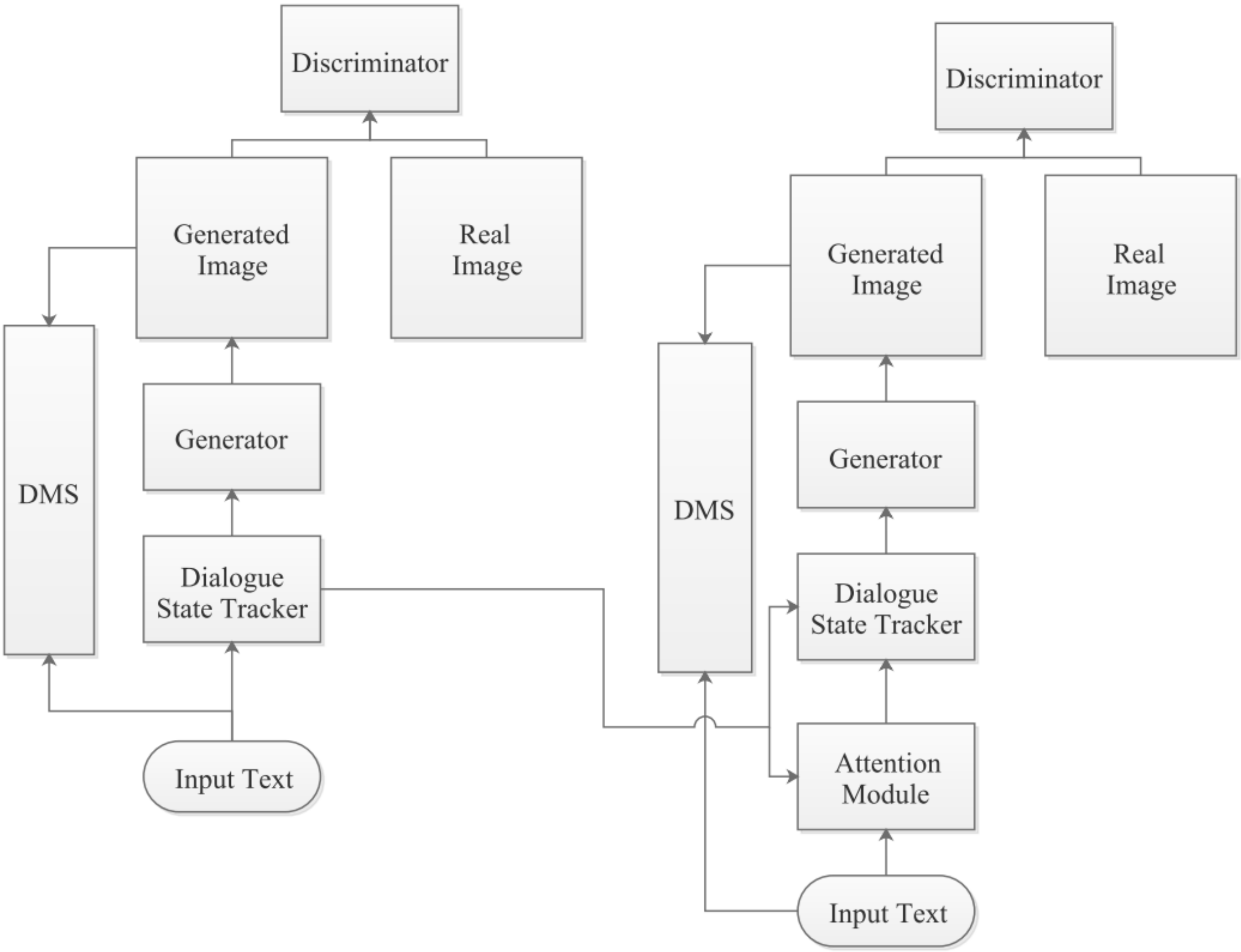

*Figure 3.1: Block Diagram of Proposed Work*

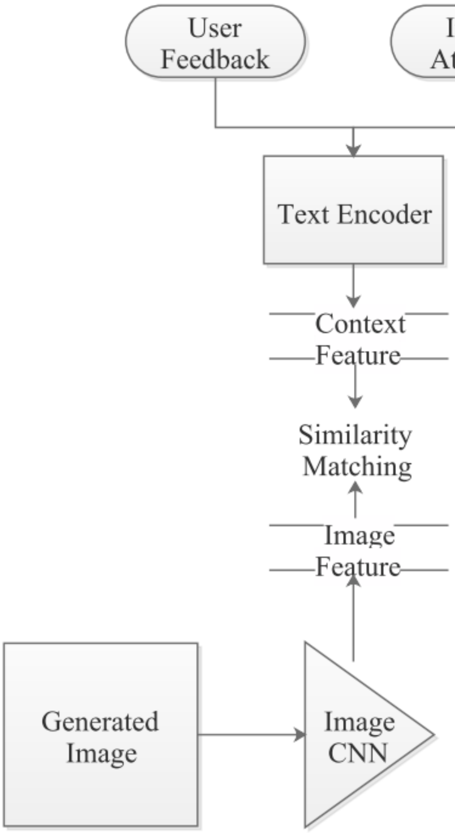

*Figure 3.2: Deep Multimodal Similarity Regularize (DMS) part of Proposed Work*

to generate an image in that step. All the dialogue turn share the same generator, while AttnGAN [1] has disjoint generators for different scales.

### 3.2.2 DEEP MULTI-MODAL SIMILARITY REGULARIZER

This work will adopt the deep attentional multimodal similarity model (DAMSM) used in [1] to: 1) maximize the utility of all the input information (such as attributes) to boost the model performance; 2) regularize the model in order to stabilize the image generator. DAMSM is to match the similarity between the synthesized images and user input sentences, acting as an effective regularizer.

The DMS function will be trained using dialogue data. Specifically, for any dialogue sample, first image retrieval will be done then concatenate the attribute value and the annotated description to have a new text. Attributes and reference feedback as the text will be combined, which is different from [1, 3, 6]. Following [1], the posterior probability of description $D_i$ matching the image $I_i$ is defined as:

$$P(D_i/I_i) = \frac{exp(\gamma R(I_i, D_i))}{\sum_{j=1}^{M} exp(\gamma R(I_i, D_i))} \tag{3.1}$$

where $\gamma$ is a smoothing factor. R() is the word-level attention-driven image-text matching score [1] (i.e., the attention weights are calculated between the sub-region of an image and each word of its corresponding text) in word level.

In summary, similar idea of DAMSM in [1] will be used to form the DMS regularizer. The training pairs will be created by concatenating attributes and user description. The image-description matching score will also be calculated in each step. By bringing in the discriminator power of DMS, the model will generate region-specific features that align well with the descriptions as well as improve the visual diversity.

## 3.3 ALGORITHM

The step-wise procedure to generate a photo-realistic image from the given user input text via multi-tun dialogue is given below:-

Step 1: Start

Step 2: Input Text/User Feedback

Step 3: Dialogue State Tracker

Step 4: Text Encoder

Step 5: Context Feature

Step 6: Generate Image

Step 7: Similarity matching with image features

Step 8: If user wants more realistic image, goto step 2.

Step 9: Stop

## 3.4 DATASET

The following datasets will be used in the proposed work:

1. CelebA, Large-scale CelebFaces Attributes Dataset, will be used in the proposed work which contains face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations.
2. The Caltech-UCSD Bird(CUB) dataset will also be used for the proposed system which contains 200 bird species with 11,788 images.

3. The COCO dataset is also selected for the proposed work because it is known to be much more challenging than the CUB dataset which contains 330K images, 80 object categories, 5 captions per image, 250,000 people with key points. Existing methods struggle in generating realistic high-resolution images on this dataset.

## 3.5 TOOLS AND PLATFORMS

The tools, programming language, and software that will be used in this thesis work is listed below:

1. Python Programming Language

2. TensorFlow Framework

3. Python Packages

   - Pillow
   - SciPy
   - python-dateutil
   - easydict

## 3.6 RESEARCH CHALLENGES

The research challenges of this thesis work are listed below:

1. Identity of Real Work: There are so many citations of my main papers. It shows that there are so many researchers working in this field. There may be parallel research works in this direction.

2. Need More Computing Power: This thesis work requires lots of computing power. May need more powerful CPU & GPUs.

# Chapter 4

# EXPECTED OUTCOME

At the end of this thesis, the following outcomes are expected:

1. A photo-realistic image will be generated during conversational phases.
2. This work will be tested using three datasets: CelebA, CUB 200 and MS COCO.

# REFERENCES

[1] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. CoRR, abs/1711.10485, 2017.

[2] Chen, Q., Koltun, V. (2017, October). Photographic image synthesis with cascaded refinement networks. In IEEE International Conference on Computer Vision (ICCV) (Vol. 1, No. 2, p. 3).

[3] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017) Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916, 2017. 1, 2, 3, 5, 7, 8

[4] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In ICLR, 2018. 2, 3, 11

[5] Dong, H., Yu, S., Wu, C., Guo, Y. (2017). Semantic image synthesis via adversarial learning. arXiv preprint arXiv:1707.06873.

[6] 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. 1, 2, 3, 5, 7

[7] Zhu, J. Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint.

[8] Chen, Q., Koltun, V. (2017, October). Photographic image synthesis with cascaded refinement networks. In IEEE International Conference on Computer Vision (ICCV) (Vol. 1, No. 2, p. 3).

[9] Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. arXiv preprint.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017. 2

[11] Reed, S., Akata, Z., Lee, H., Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 49-58).

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[13] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. Stacked generative adversarial networks. In CVPR, 2017. 2

[14] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., ... Gao, J. (2018). StoryGAN: A Sequential Conditional GAN for Story Visualization. arXiv preprint arXiv:1812.02784.

[15] Radford, A., Metz, L., Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[16] Berthelot, D., Schumm, T., Metz, L. (2017). BEGAN: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717.

[17] Yang, J., Kannan, A., Batra, D., Parikh, D. (2017). LR-GAN: Layered recursive generative adversarial networks for image generation. arXiv preprint arXiv:1703.01560.

[18] Singh, A., Agrawal, S. (2018). CanvasGAN: A simple baseline for text to image generation by incrementally patching a canvas. arXiv preprint arXiv:1810.02833.

[19] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In NIPS, 2016. 1, 2, 8

[20] Cheng, Y., Gan, Z., Li, Y., Liu, J., Gao, J. (2018). Sequential Attention GAN for Interactive Image Editing via Dialogue. arXiv preprint arXiv:1812.08352.

[21] Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X. (2017). Language-Based Image Editing with Recurrent Attentive Models. arXiv preprint arXiv:1711.06288.

[22] Manuvinakurike, R., Bui, T., Chang, W., Georgila, K. (2018). Conversational image editing: Incremental intent identification in a new dialogue task. In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (pp. 284-295).

[23] Wang, X., Gupta, A. (2016, October). Generative image modeling using style and structure adversarial networks. In European Conference on Computer Vision (pp. 318-335). Springer, Cham.

[24] Durugkar, I., Gemp, I., Mahadevan, S. (2016). Generative multi-adversarial networks. arXiv preprint arXiv:1611.01673.

[25] Denton, E. L., Chintala, S., Fergus, R. (2015). Deep generative image models using aï£ij laplacian pyramid of adversarial networks. In Advances in neural information processing systems (pp. 1486-1494).

# Appendix A

# THESIS TIME-LINE

*Table A.1: Detailed Schedule for the Thesis*

| **Tasks** | **Month/Year**<br>**2018/2019** | | | | | |
|---|---|---|---|---|---|---|
| | **Dec** | **Jan** | **Feb** | **Mar** | **Apr** | **May** |
| Preliminary Investigations | ■ | | | | | |
| Literature Review | ■ | ■ | ■ | ■ | | |
| Proposal Defense | ■ | ■ | | | | |
| System Design & Coding | | ■ | ■ | ■ | ■ | |
| Mid-Term Defense | | | ■ | ■ | | |
| Final Submission of Thesis | | | | | ■ | ■ |
| Documentation of Thesis | | ■ | ■ | ■ | ■ | |
| Research and Experiments | | | ■ | ■ | ■ | ■ |