INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Transformation of Text to Image for In-Vogue Fashion Designs & Stylings

A Project Proposal by

Mr. Aarthif Nawaz

Supervised by

Mr. Guhanathan Porovi

Submitted in partial fulfillment of the requirements for the BEng/BSc in Software Engineering degree at the University of Westminster

Date: November 2021

Key Words: text to image synthesis, generative models, image generation

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	**Acronym**
AI	Artificial Intelligence
CNN	Convolutional Neural Network
GAN/s	Generative Adversarial Network/s
GPU	Graphics Processing Unit
GUI	Graphical User Interface
LR	Literature Review
LSTM	Long Short-Term Memory
OS	Operating System
SSADM	Structured Systems Analysis and Design Method
DAMSM	Deep Attentional Multimodal Similarity Model
SOTA	State of The Art

1. PROJECT PROPOSAL

1.1 Introduction

The present era has become a trending fashion design warehouse. People around the world spend thousands of dollars on products like clothing, makeup, and other fashion accessories. But there is even more special attention given to in-vogue style designs where the upper-class individuals step foot. Fashion products that have unique designs focused more on in-vogue styling like Haute couture, Label, Ensemble, Silhouette, Off-the-rack, Hemline, In-vogue, Fashion-forward, Monochrome, Peplum, etc. have been paid a high price to get delivered to the market. Nowadays, AI has played a great role in generating valuable outputs in various industries which in turn become quality units sold in the market. Generating high-resolution images from natural language has become one of the main applications as of today. Nevertheless, these problems are entirely different because text-image conversions are highly multimodal problems that have a good impact on multiple domains (Bodnar, 2018). The generation of such images in the fashion in-vogue industry makes the job of fashion designers extremely easier and less time-consuming where a simple phrase or a large storyline can draw out a two-dimensional image that is useful as an end product (Zaidi, 2017). Therefore, the motivation for addressing this domain for this project is the direct impact it would have on the growing fashion design industry.

1.2 Problem Domain

1.2.1 Generative adversarial networks

According to (Goodfellow *et al.*, 2014) Generative adversarial networks are types of generative models that generate images, audios, videos, text-based on different training characteristics of the dataset. In other words, the ultimate goal is to generate similar and new output based on the training data. Different types of GANs have been trained, tested, and evaluated to outperform the SOTA models that existed in various domains. The GAN network features a generator and discriminator model where the generator simultaneously trains itself until it can outperform the discriminator. Thus, able to generate high-quality synthesized outputs based on the training dataset. There are many use-cases of GANs and generating realistic-looking photographs, generating visual emotions of human beings is a widely known scenario. Generating fake data for data augmentation, generating audio, generating videos, text generation, paraphrasing, semantic image-to-photo translation, and text-to-image generation are some of the examples.

1.2.2 Text-to-Image Generation

Text to image generation aims to generate high-quality photosynthetic images that are semantically consistent with the input text description (Hu *et al.*, 2021). The existing systems which are built upon using the unsupervised learning approach such as GANs have demonstrated amazing results, which in turn has been given a lot of importance as it has been usefully applicable in many industries and various other domains. Previous studies and SOTA have shown commendable results on the visual quality of the generated images. Text to image generation has been experimented with, tested, and evaluated by several datasets. The inception score of the GANS has been a remarkable result achieved so far in this domain. But the limitations of mode collapse and the ability of GANs to generate images on text descriptions on a fine-grained level taking certain phrases/keywords from the text makes the image visually incomplete and not useful for the domains which prefer to input text descriptions of substantial size.

The existing systems for text-to-image synthesis, first generate an initial image with rough shape and color and then refine the initial image to a high-resolution one (Zhu *et al.*, 2019). The input text is converted to an embedding vector combined with random noise and passed to a generator, whereby the generator produces better images and tries to outperform the discriminator. The limitations not only subside in the visual semantic consistency of the image but also in the length of the text. The ability to guarantee semantic consistency between the textual descriptions and visual image remains very challenging (Qiao *et al.*, 2019). The ideology is to make sure that the text descriptions can generate the image in all aspects of the way that’s been described such that the semantic consistency is uniform and preserved throughout the training process.

1.2.3 In-Vogue Fashion Design Industry

The fashion design industry is a large-scale manufacturing business that has created plenty of demands and supply as of the 21st century. Statistics have proven that every human being uses fashion in some way or the other despite what class they belong to. It has been paid much attention to evolve better livelihood. It has also proven the fact that it has been a consumer need that people spend hundreds of dollars on it.

Understanding the significance this domain produces. There is also a lot of problems that the fashion design industry encounters. Though, multi-million-dollar companies hire fashion experts to draw out creative ideas and implicate them into fashion designs that are to be used in the market and to be accepted in the industry. It takes a lot of hard work, creativity, and time to outcast production-level fashion designs.

The fashion design industry, where generation of high-quality outputs with a lot of information and features to the design is needed, and for which multiple reasonings and descriptions are deeply necessary to produce such high-level quality. There is no such solution that will help us achieve the above at present.

Based on the above viewpoints and aspects. It is very clear that manually creating fashion designs and in-vogue stylings can be time-consuming with a lot of human effort that will involve a lot of raw materials and substantial in-house energy. With all of that being said, the problem definition for this problem was bought out.

1.3 Problem Definition

Generating fashion designs and in-vogue styles by the experts in the industry is a lot time-consuming and it needs a lot of human effort like creativity, ability to reason out, ability to comprehend various phenomenal factors (nature, background, luminosity). Though these can be carried out manually. It’s still a tedious task that needs technical intervention.

Transformation of text to image has been applied to modern multimodal applications to facilitate the easiness of generating images for a wide variety of textual inputs such as keywords and phrases. It has been used as a generic approach in CAD designing, scientific engineering, graphic designing, image fine-tuning, and perhaps even animation (Zaidi, 2017). But these transformations only have produced outputs for simple phrases, clauses, and pre-defined trained text towards the image. As for the basics, generative adversarial networks (GAN) (Goodfellow et al. 2014) have successfully demonstrated the learning of a probability distribution to synthesize realistic examples of images from textual descriptions which has been vividly used in the text to image domain. Recent progress in generative models, especially Generative Adversarial Nets (GANs) in many forms and aspects has made significant improvement in synthesizing images and generating plausible and approvable samples. But it has only proven successful for text description having a span of less than max 15-20 words which also uses only the terms specific to the domain. But if we are to input a larger span of text like storylines, paragraphs, detailed descriptions. The GANs haven’t been able to pull out versatile quality images. In simple terms, only fine-grained texts are of considerable inputs to produce images specific to the domain. (Nasr, Mutasim and Imam, 2021).The current model boosts the original GAN Inception score by +4.13% and the Frechet Inception Distance score by +13.93%. Which is not quite good compared to generating images for a larger span of texts.

1.3.1 Problem Statement

Designing in-vogue stylings and fashion designs at a mass scale through manual crafting, usage of in-house energy and raw materials, which is a lot time-consuming, facilitating more manpower, and usage of additional unwanted raw materials. To avoid all of these the author proposes a generic text-to-image model that takes in fashion terms along with a substantial description in three natural languages which will produce fashion designs of various in-vogue styles of different designs.

1.4 Research Motivation

Ian Goodfellow's revolutionary introduction of the initial GAN (Goodfellow *et al.*, 2014) and its variants built by other researchers are being applied for image generation in many domains. GANS has been considered as the primary approach for unsupervised learning. It also has great potential in the image generation domain which has been paid a lot of attention to in the recent past (Salimans *et al.*, 2016). Especially the fashion design industry where they could simply visualize the in-vogue designs of unique standards in a variety of ways in just a few pieces of texts. Thereby the author believes in developing an AI-based solution that will reduce manpower, time consumption, cost, and other raw materials needed to design an in-vogue fashion design. This is indeed a major landmark for the fashion designers and experts to change the entire fashion design industry.

1.5 Existing Work

The text to image synthesis has been used in the past three years using several GANs as an experimental prototype and has been tested among a few datasets (COCO, CUB, Flowers, Birds) to determine how effective it is in terms of resolution, the length of text it could process, performance time, non-blurriness of the image and the number of different varieties of image it could generate per train. Different papers have based their training for the above features using different types of GANs namely (Stack GANs – 2 stage, DF GANS, Mirror GANs, DM-GANs), variational text encoders, image encoders, and contrastive learning approaches. The below papers have explained in detail how they managed to achieve the text to image transformation experimented for different datasets.

Citation	**Technique**	**Improvement**	**Limitations**
(Bodnar, 2018)	Wasserstein's GAN-CLS approach uses a deep convolutional architecture for both the generator and the discriminator model training. Input Text embedded as a vector and passed as random noise to the generator with a series of CNN layers for the image to get generated. Whilst, real image and caption are trained by the discriminator separately so that generator can outperform the discriminator to achieve good results.	The model boosts by 7.07%, the best Inception Score (on the Caltech birds' dataset) of the models which use only the phrase-level keywords to generate visual semantics (Image).	• Resolution of images limited to 64*64 pixels. • Only English natural language. • Phrase level keywords considered for training. • Blurry images with a lot of background noise.
(Zaidi, 2017)	The two-stage procedure, each of which is its own manageable GAN implementation. Stage 1 GAN - Using GAN to generate a rough blurry sketch of the image based on the input text Stage 2 GAN -Super resolution of the sketched image based on the combination of stage one embedded text and the rough image generated from stage one.	The evaluation of a Stack-GAN productively generates highly realistic synthetic images from text phrase. Model Fretchet Inception score boosts by 14.4%.	• Resolution limited to 256*256 pix • Only English natural language. • Phrase level keywords considered for training.
(Zhu *et al.*, 2019)	Employs a dynamic memory component to refine the initial generated image based on the input text, a memory writing	• Propose the dynamic memory	• Refines initial images with wrong color

	gate to highlight important text information, and finally, a response gate to fuse the image and memory representation of the highlighted text, So the image gets generated based from the fusion.	module to select the most relevant words from a sentenced-based text to visually represent the global image feature in lesser time. • Produces high-level synthesized quality images.	and rough shapes. • Sentenced level text. But only a few relevant words are taken into consideration, thereby generating some images outside the scope. • Generates images from varied scopes irregular of the input image scope. • Only English natural language.
(Zhang, Li, and Zhou, 2021)	Textual description of real images is embedded in the noise vector as a constraint to the discriminator. Based on Deep Convolutional Generative Adversarial Networks (DCGAN), multiple generators are incorporated to capture high	• Focused on generating images more diverse, thus avoiding mode collapse to some extent.	• Phrase-level textual descriptions. • Only English natural language.

	probability among the target distribution exposed from the discriminator loss and the image gets generated.	• The images synthesized have relatively high resolution and high quality. • Multiple generators are used to avoid mode collapse. • faster training time and good inference speed.	• Uses local word embedding, thereby subsides to a limitation of words for the training set.
(Mishra *et al.*, 2020)	Follows a RGAN architecture that generates images of reasonable visual details from the given text query. The RGAN mainly focuses on the residual loss generated by the generator and discriminator as a combination to improve the image generation.	Stabilizing the learning process and new ways for the model to reach convergence faster, when generating images from textual descriptions.	- Texts are more query-related and fine-grained relevant words to be taken as input. - The quality of images is low 64*64 pixels. - Usage of only English natural language.

Table 1 - Comparison Of Text To Image Synthesis Models

1.6 Research Gap

The research gap that the author offers is to achieve the ability for text to image transformation models to take in larger descriptions of multiple sentences (storylines, paragraphs) as inputs to produce images that are taken as a future work (Mishra *et al.*, 2020). Moreover, the author is applying this research on a domain towards fashion in-vogue design industry which has not been done so far. Also, another theoretical gap that the author will address in the project is the ability to take in three natural languages (English, Sinhala, and Tamil) as training inputs using machine translation and language modeling. That way the model will be able to generate images from three natural languages.

1.7 Contribution to The Body of Knowledge

Based on the gap that the author wishes to achieve, which is to extend the models architecture to take in descriptions with multiple sentences and pertain it to the trending in-vogue style and design industry. Thereby contributing immensely to the technical as well as the domain proportions in multiple ways.

1.7.1 Technical Contribution

The text to image transformation models built so far has been proved successful and unsuccessful in different aspects of their model origins (Dhivya and Navas, 2020). A Basic two-staged GAN or a dynamic memory GAN only takes texts in a fine-grained level. Even if larger sentences are provided only the most relevant words are taken into considerations to produce images, thus leading the image to be less detailed with the least features that are not expected. To avoid all of these encounters the most recommended Stack GAN which has a high inception score in terms of textual semantics (Xu *et al.*, 2018) with a set of pre-trained texts will be ensembled with a contrastive learning approach to produce a novel framework that will learn and train images with larger textual descriptions of three natural languages.

1.7.2 Domain Contribution

This research will be contributing to the in-vogue fashion design domain where users will have the ability to input a description and the system will generate varieties of diverse outputs of in-vogue stylings based on the given description. This way it will help the fashion designers to easily generate images and provide immense benefits like gaining extensive ideas, additional pattern creativity that they could use in their designs. This will also help in saving time, usage of less manpower, usage of less mental effort, and usage of less raw material. These productive outcomes will result in a huge favorite amongst the fashion designers in the market. And with just mere pieces of textual descriptions in your desired natural language (English, Sinhala, Tamil), the in-vogues are ready to be used and it will provide numerous benefits.

1.8 Research Challenges

The text to image synthesis is a research area that has been actively producing a lot of valuable outcomes in different aspects (Cha, Gwon, and Kung, 2019). Most of the current existing text-to-image synthesis models are developed using GANs, GAN-CLS, and other GAN networks specific to the text-to-image domain. But various domains using text to image synthesis have had limitations in different ways particularly from the GAN networks. Some of them include blurry images, phrase-level texts, training time, capacity, outputs out of scope, etc. (Mishra *et al.*, 2020). According to my investigation on the research, there are plenty of challenges to be faced during the research process and they are listed below.

- Analyzing and choosing the appropriate technology to solve the process of taking input of larger descriptions of substantial size and synthesize them to high quality images.
- Designing and developing a framework that will overcome the limitations like mode collapse, blurry image, inconsistency of visual and textual semantics incurred from the existing SOTA text-to-image synthesis models.
- Analyzing and choosing the best text-to-image synthesis unsupervised approach with contrastive learning to synthesize textual description to high quality two-dimensional images.
- Using techniques and identifying the appropriate tools, libraries and technologies to develop the text-to-image synthesis model which will accept larger text descriptions.
- Analyzing and identifying tools and technologies to train multi language-based text-to-image synthesis models.

1.9 Research Questions

RQ1: What are the current technology frameworks/algorithms/network architectures used in the text-to-image synthesis models?

RQ2: What are the current problems being faced and areas that should be improved when developing text-to-image synthesis models to generate better-detailed images for the descriptions given?

RQ3: Why is GANs the most preferred architecture to implement a text to image synthesis model and how far has it progressed in terms of textual description and image quality/resolutions?

RQ4: What are the challenging problems that need to be solved when transforming text to images?

RQ5: What existing text to image synthesis models can be used as a partial resolution to solve the challenging problems faced when transforming text to images?

1.10 Research Aim

This research aims to design, implement and evaluate a text to image synthesis model that will absorb larger spans of in-vogue fashion-related text (storylines, paragraphs) of substantial size in any three natural languages (English, Sinhala, Tamil) and generate highly synthesized in vogue fashion design images with better details, resolution, and features.

To further elaborate on the aim. This research project will produce a text-to-image synthesis model for fashion designers. This system will allow my stakeholders to input descriptions explaining the desired in-vogue fashion style they wish to visualize and it will produce high-quality two-dimensional images with a lot of details present in the image. Outputs of state-of-the-art methods in the domain will be compared using proper evaluation metrics.

The research domain and the technological body will be properly researched to obtain knowledge before proceeding towards the development of this project. The knowledge gained will be used in developing components needed for the model and several other areas which will achieve my outcome. The system will be open source, deployed onto a dedicated server and it will have the ability to run across any device including mobile and desktop.

1.11. Research Objectives

After investigating research aims and questions, subsequent objectives were allotted in the perspective of the research.

Research Objectives	**Description**	**Learning Outcome**
Problem Identification	Carry out in-depth research to identify a potential problem that needs to be solved. • **RO1:** Research on the fashion design domain and analyze the problems encountered. • **RO2:** Research on how to overcome the ability to visualize in-vogue images simply from fashion textual descriptions. • **RO3:** Research on why the current algorithms that transform text to images take only phrase-level textual inputs. • **RO4:** Research on why the current algorithms only pursue the ability to transform text to images using the English natural language.	**LO1, LO4**
Literature Review	Conduct extensive research to evaluate how to achieve the target outcome. • **RO1:** Research and analyze existing text to image synthesis models. • **RO2:** Analyze techniques and technologies used to transform text to images. • **RO3:** Research on existing Generative Adversarial Networks in the text to image domain. • **RO4:** Research on how the current text to image models train phrase-level texts and produce images.	**LO1, LO4, LO8**

	- **RO5:** Research on how the current text to image synthesis models has been applied to different industries. - **RO6:** Elaborating the research gap applied to the fashion design industry and training larger spans of texts with three natural languages. - **RO7:** Provide an analysis document on the critically evaluated system.	
Data Gathering and Analysis	Carry out a requirement gathering analysis. - **RO1:** Gathering feedback on building a text to image transformation model that will take larger spans of texts in three natural languages applied for the first time to a fashion design domain. - **RO2:** Evaluate the requirements gathered to develop a GAN-based network that will allow the transformation of text to images. - **RO3:** Determine the frontend, backend, and deployment perspectives of the system. - **RO4:** Gather, evaluate and define the end-user requirements through questionnaires.	**LO3, LO4, LO6, LO8**
Research Design	Plan the timeline, design, a text-to-image synthesis model that takes larger spans of fashion-based textual descriptions. - **RO1:** Design a text-encoder network that will convert texts to noise vectors and create text embeddings. - **RO2:** Design an image-encoder network that will convert images to vectors. - **RO3:** Design a generator to perform upsampling for the text got from the text-encoder passed as a noise vector. And a discriminator network to perform downsampling	**LO2, LO5, LO7, LO8**

	of the images got from the image-encoder passed for model training. • **RO4:** Design a DL model architecture using CNN to combine the images and text vectors into batch layers. • **RO5:** Design the frontend for my system using the React JS Framework. • **RO6:** Design the backend of my system using Python Flask and expose API calls to the model execution.	
Implementation	Developing the prototype with the appropriate software and hardware resources to fulfill the requirements and design aspects identified above. • **RO1:** Develop the text-encoder network that will take in larger spans of texts and converts them to vectors. • **RO2:** Develop the image-encoder network that will take an image and convert that to a vector. • **RO3:** Develop the generator and discriminator network. • **RO4:** Develop the deep learning model using CNN with the relevant tensor layers. • **RO5:** Develop the frontend using React JS. • **RO6:** Develop the backend using Python Flask. • **RO7:** Deployment of the system to the cloud.	**LO2, LO5, LO7, LO8**
Testing and Evaluation	Evaluation of generative models includes inception score and visual quality comparisons. Test and evaluate the prototype. • **RO1:** Evaluate and test the created text-to-image model and compare against current existing models in terms of inception score and Frechet Inception Distance score. • **RO2:** Create a test plan to perform unit, integration, and functional requirement testing of the prototype.	**LO8, LO9**

	• **RO3:** Produce a detailed report for the academic and research community.	

Table 2 - Research Objectives

1.12 Project Scope

Based on the initial literature review and research objectives, focused and non-focused points of research are defined.

1.12.1 In-Scope

Parts that will be focused on during the research process are as follows:

- Reviewing and analyzing the SOTA generative adversarial network models applied to the text to image domain and other text to image synthesis models.
- Deciding on using a proper approach to transform larger spans of text to a detailed image with all the features mentioned in the text.
- Deciding on using a proper technique to train text descriptions for Tamil and Sinhala language.
- Effectively reading in-vogue fashion-based text descriptions and mapping each feature word from the text to an image and finally generating an in-vogue fashion design image with all the relevant details needed from the text.
- Evaluating results of the system based on common evaluation metrics.
- Developing a full-stack application that will take in-vogue fashion text descriptions and generate a two-dimensional high-quality image.

1.12.2 Out-Scope

Parts that will not be focused on during the research process are as follows:

- The project will be limited to only taking in-vogue fashion-based textual descriptions and the model will only be effective based on fashion-related texts and other supporting words to generate in-vogue fashion design images.
- Projects focus is only to extend the ability to train larger spans of textual description and the ability to work on multiple natural languages. But not to generate videos from texts or generate images outside of the fashion design scope.

- The author will focus only on credible and novel related works having promising results in respective domains and technology.

1.12.3 Diagram Explaining Prototype Features

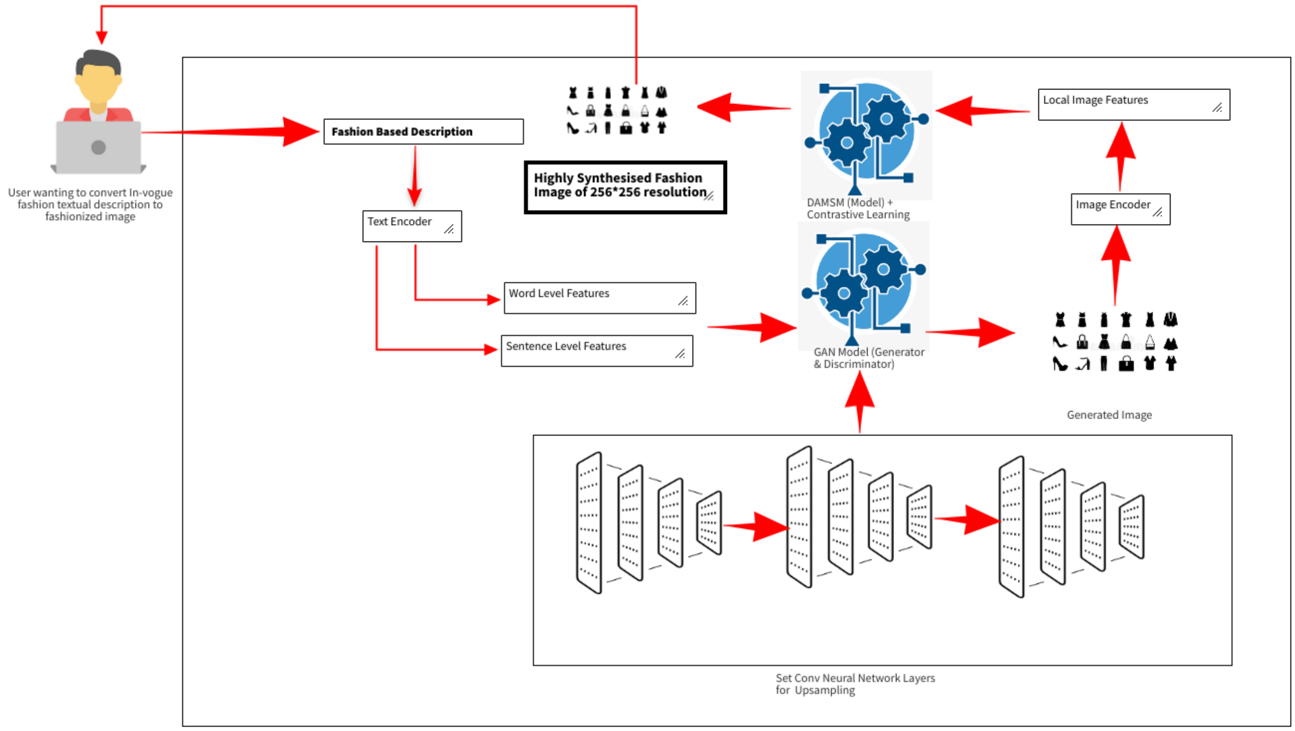

User wanting to convert In-vogue fashion textual description to fashionized image
Fashion Based Description
Text Encoder
Word Level Features
Sentence Level Features
Highly Synthesised Fashion Image of 256*256 resolution
DAMSM (Model) + Contrastive Learning
Local Image Features
Image Encoder
GAN Model (Generator & Discriminator)
Generated Image
Set Conv Neural Network Layers for Upsampling

Figure 1 : - Diagram explaining the features of the prototype

2. METHODOLOGY

2.1 Research Methodology

Research Methodology	
Philosophy	**Pragmatism** was chosen as the philosophy because research is based on data to develop a hypothesis and this research is comparing both **qualitative** and **quantitative** results produced by different text-to-image synthesis models using GANs. This is also an **applied research** on the domain of In-vogue fashion designs and stylings.
Approach	This research aims to test and prove a **hypothesis** that needs to be solved. This is to input larger spans of In-vogue fashion-based textual descriptions in three natural languages (English, Sinhala, Tamil) and achieve high-quality detailed In-vogue fashion design images as output. A **deductive** approach was selected to follow, as the research chooses to apply an existing theory to the domain of interest.
Strategy	The strategy of research is how the answers to the research questions are proposed**. Interviews, documents, and research analysis, experiments**, and **surveys** were chosen as strategies to fit the research strategy.
Choice	Choice of research will depend on the research paradigm that is chosen. Among the mono, multi and mixed methods for choice, **mixed-method** was chosen, as text to image synthesis model had both quantitative and qualitative results which were gained through interviews, survey papers, and other documents such as journal articles and conference papers that could be used as a comparison for the model prototype that is going to be developed in this research.
Time Horizon	Data needs to be gathered at a single point in time to do evaluations. Hence, out of longitudinal and cross-sectional, **cross-sectional** time horizon looks the most convenient and was chosen for the research.
Techniques and Procedures	For the collection and analysis of data, techniques such as observations, documents, conversations, evaluation reports, interviews, and questionnaires will be used.

Table 3 - Research methodology selection and reasoning

Based on the above research methodologies, the below mentioned aspects of the research were determined.

Research Hypothesis: With the ability of text to image transformation models to take in larger spans of texts with multiple natural languages by training and validating to provide high-quality fashion design images. It is possible for developers and other stakeholders to now train larger text descriptions with multiple natural languages and apply them to any other domain for productive outcomes.

Research Process: Finding out the best possible way to train larger corpora of texts in multiple languages whilst simultaneously applying those trained texts description on the images to extend the process for the unsupervised GAN network to produce images for larger textual descriptions.

Prototype Input: Text descriptions of in-vogue fashion-based designs that you desire to visualize.

Prototype Output: High-quality two-dimensional synthesized In-vogue fashion design images having all the detailed features given as input.

Prototype Features:

1. A novel framework to generate high-quality images from larger spans of texts in multiple languages.
2. It will be an open-source application to be used by end-users to input in-vogue fashion descriptions and achieve high-quality in-vogue fashion design images as output.
3. A GUI for better user experience.

2.2 Development Methodology

Among the various software lifecycle methodologies available. The author will be using the prototype methodology where my system is built and tested repeatedly as there will be different evaluations at different stages. The author decided to choose the prototype model as he has to repeatedly make changes and test until an achievable outcome is accumulated and until a positive outcome is attained.

2.3 Design Methodology

The author will be using SSADM as he will be following a methodological approach in designing the system's architecture. The author decided to choose this methodology so that he can improve project management control, make more effective use of experienced developmental staff, and develop a better-quality system that is versatile, accurate, and efficient in producing outputs.

2.4 Evaluation Methodology

The system will be tested and evaluated using the evaluation metrics for GAN inception score, R precision score, and Frechet Inception Distance. Also, the developed system will be benchmarked against all other SOTA text to image synthesis models with the Birds dataset to understand the comparison to where our system stands. The author will benchmark the current SOTA models in terms of data capacity, training speed, model precision, and inference speed.

2.3 Project Management Methodology

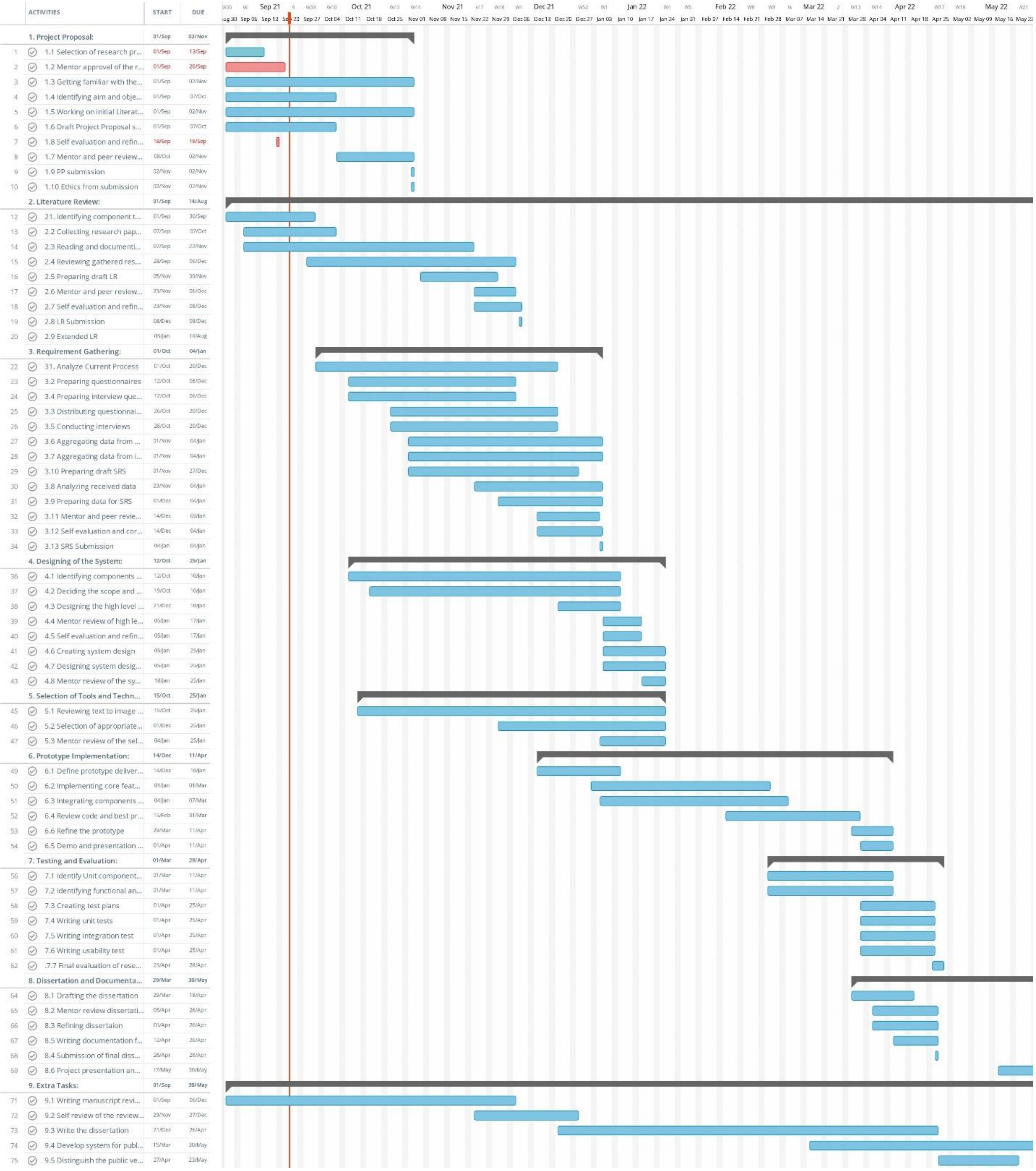
Among many candidates, Agile Prince2 which allows focusing on management, recursive planning, and flexible delivery while adapting to risks was chosen for project management. The author decided to choose agile prince2 methodology because it allows you to focus on both management and delivery at the same time, helps to be on time and hit deadlines consistently, cooperate friendly, increased stakeholder confidence, and the ability to change tools and technology if the requirements constantly keep changing.

2.3.1 Project Deliverables

Deliverable Component	**Tentative Delivery Date**
Project Proposal	1st Nov 21
Review Paper	15th Oct 21
Literature Review Document	18th Oct 21
Software Requirement Specification	22nd Nov 21
System Design Document	6th Dec 21
Prototype	25th Apr 22
Thesis	25th Apr 22
Project Research Paper	1st May 22

Table 4 - Project Deliverables

2.3.2 Gantt Chart



		ACTIVITIES	START	DUE
		1. Project Proposal:	01/Sep	02/Nov
1	✓	1.1 Selection of research pr...	01/Sep	13/Sep
2	✓	1.2 Mentor approval of the r...	01/Sep	20/Sep
3	✓	1.3 Getting familiar with the...	01/Sep	02/Nov
4	✓	1.4 Identifying aim and obje...	01/Sep	07/Oct
5	✓	1.5 Working on initial Literat...	01/Sep	02/Nov
6	✓	1.6 Draft Project Proposal s...	01/Sep	07/Oct
7	✓	1.8 Self evaluation and refin...	18/Sep	18/Sep
8	✓	1.7 Mentor and peer review...	08/Oct	02/Nov
9	✓	1.9 PP submission	02/Nov	02/Nov
10	✓	1.10 Ethics from submission	02/Nov	02/Nov
		2. Literature Review:	01/Sep	14/Aug
12	✓	21. Identifying component t...	01/Sep	30/Sep
13	✓	2.2 Collecting research pap...	07/Sep	07/Oct
14	✓	2.3 Reading and documenti...	07/Sep	22/Nov
15	✓	2.4 Reviewing gathered res...	28/Sep	06/Dec
16	✓	2.5 Preparing draft LR	05/Nov	30/Nov
17	✓	2.6 Mentor and peer review...	23/Nov	06/Dec
18	✓	2.7 Self evaluation and refin...	23/Nov	08/Dec
19	✓	2.8 LR Submission	08/Dec	08/Dec
20	✓	2.9 Extended LR	05/Jan	14/Aug
		3. Requirement Gathering:	01/Oct	04/Jan
22	✓	31. Analyze Current Process	01/Oct	20/Dec
23	✓	3.2 Preparing questionnaires	12/Oct	06/Dec
24	✓	3.4 Preparing interview que...	12/Oct	06/Dec
25	✓	3.3 Distributing questionnai...	26/Oct	20/Dec
26	✓	3.5 Conducting interviews	26/Oct	20/Dec
27	✓	3.6 Aggregating data from ...	01/Nov	04/Jan
28	✓	3.7 Aggregating data from i...	01/Nov	04/Jan
29	✓	3.10 Preparing draft SRS	01/Nov	27/Dec
30	✓	3.8 Analyzing received data	23/Nov	04/Jan
31	✓	3.9 Preparing data for SRS	01/Dec	04/Jan
32	✓	3.11 Mentor and peer revie...	14/Dec	03/Jan
33	✓	3.12 Self evaluation and cor...	14/Dec	04/Jan
34	✓	3.13 SRS Submission	04/Jan	04/Jan
		4. Designing of the System:	12/Oct	25/Jan
36	✓	4.1 Identifying components ...	12/Oct	10/Jan
37	✓	4.2 Deciding the scope and ...	19/Oct	10/Jan
38	✓	4.3 Designing the high level ...	21/Dec	10/Jan
39	✓	4.4 Mentor review of high le...	05/Jan	17/Jan
40	✓	4.5 Self evaluation and refin...	05/Jan	17/Jan
41	✓	4.6 Creating system design	05/Jan	25/Jan
42	✓	4.7 Designing system desig...	05/Jan	25/Jan
43	✓	4.8 Mentor review of the sy...	18/Jan	25/Jan
		5. Selection of Tools and Techn...	15/Oct	25/Jan
45	✓	5.1 Reviewing text to image ...	15/Oct	25/Jan
46	✓	5.2 Selection of appropriate...	01/Dec	25/Jan
47	✓	5.3 Mentor review of the sel...	04/Jan	25/Jan
		6. Prototype Implementation:	14/Dec	11/Apr
49	✓	6.1 Define prototype deliver...	14/Dec	10/Jan
50	✓	6.2 Implementing core feat...	01/Jan	01/Mar
51	✓	6.3 Integrating components ...	04/Jan	07/Mar
52	✓	6.4 Review code and best pr...	15/Feb	31/Mar
53	✓	6.6 Refine the prototype	29/Mar	11/Apr
54	✓	6.5 Demo and presentation ...	01/Apr	11/Apr
		7. Testing and Evaluation:	01/Mar	28/Apr
56	✓	7.1 Identify Unit component...	01/Mar	11/Apr
57	✓	7.2 Identifying functional an...	01/Mar	11/Apr
58	✓	7.3 Creating test plans	01/Apr	25/Apr
59	✓	7.4 Writing unit tests	01/Apr	25/Apr
60	✓	7.5 Writing integration test	01/Apr	25/Apr
61	✓	7.6 Writing usability test	01/Apr	25/Apr
62	✓	.7.7 Final evaluation of rese...	25/Apr	28/Apr
		8. Dissertation and Documenta...	29/Mar	30/May
64	✓	8.1 Drafting the dissertation	29/Mar	18/Apr
65	✓	8.2 Mentor review dissertati...	05/Apr	26/Apr
66	✓	8.3 Refining dissertaion	05/Apr	26/Apr
67	✓	8.5 Writing documentation f...	12/Apr	26/Apr
68	✓	8.4 Submission of final diss...	26/Apr	26/Apr
69	✓	8.6 Project presentation an...	17/May	30/May
		9. Extra Tasks:	01/Sep	30/May
71	✓	9.1 Writing manuscript revi...	01/Sep	06/Dec
72	✓	9.2 Self review of the review...	23/Nov	27/Dec
73	✓	9.3 Write the dissertation	21/Dec	26/Apr
74	✓	9.4 Develop system for publ...	15/Mar	30/May
75	✓	9.5 Distinguish the public ve...	27/Apr	23/May

Figure 2 - Gantt Chart

2.3.3 Resource Requirements

Required hardware, software, dataset, and skills requirements are decided based on research objectives and the plan to develop and test.

2.3.3.1 Hardware Requirements

Hardware Requirement	Justification
Core i7 9th generation, 6 core processor	To provide required intensive processing power.
16GB RAM	To load heavy image datasets and save them in memory.
Graphics Processing Unit	A powerful enough GPU to train models will be necessary. Can use Google Collab for this.
Storage space of 40 GB or more	Disk space is required to save application code, data sets, and testing files in non-volatile memory.

Table 5 - Hardware Requirements

2.3.3.2 Software Requirements

Software Requirement	Justification
Operating System (Windows 10, Linux)	The OS is required to handle all heavy processes. Linux has been the favored OS for the deep learning community. However, a 64-bit version of either Windows 10 or Linux can be used.
PyTorch	A Pythonic framework by Facebook’s AI research lab for deep learning.
TensorFlow GPU	A deep learning framework built by Google to handle deep learning development.
Google Collab	To research and experiment with different available models, with GPU power in the cloud.
MS Office	To create all documentation and reports.

Google Drive	To keep backups of application code and datasets.
Zotero	To manage all citations and references.

Table 6 - Software Requirements

2.3.3.3 Dataset Requirements

- Available datasets required for the project (DeepFashion and FashionGen)

2.3.3.4 Skill Requirements

- Skills and knowledge in GANs.
- Knowledge in using neural networks and evaluating them
- Knowledge in evaluating GANs.
- Knowledge about techniques and loss functions while training with GANs.

2.3.4 Risk Management

Risk	Severity	Frequency	Mitigation Plan
Constant changes for the requirement of the project. Like any other research, this research will also be incumbent to the evolving requirements with the recurrence of the prototype. Such changes need to be addressed.	4	5	Following the prototype methodology will help overcome frequent requirement changes.
In-depth knowledge of the domain knowledge needs to be acquired as the in-vogue fashion domain is subjected to many changes.	4	5	Thorough research about the domain and existing technologies will mitigate the risk of having a lack of domain and technical knowledge.
Limited Availability of the hardware resources as the prototype system may require advanced	5	5	To mitigate this issue, Google Collab and cloud services from

computational power and processing speed			Amazon, Google or Microsoft can be used.
Losing of previous and existing related work can be a big risk when compared to the project.	4	4	The author needs to stay updated about the domain and technologies. If such new contributions are made, it is important for those to be analyzed and to adjust the project accordingly.

Table 7 - Risk Management

References

Bodnar, C. (2018) ‘Text to Image Synthesis Using Generative Adversarial Networks’, *ArXiv*, abs/1805.00676.

Cha, M., Gwon, Y. and Kung, H.T. (2019) ‘Adversarial Learning of Semantic Relevance in Text to Image Synthesis’, *AAAI* [Preprint].

Dhivya, K. and Navas, N.S. (2020) ‘Text to Realistic Image Generation Using Stackgan’, *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, pp. 1–7.

Goodfellow, I. *et al.* (2014) ‘Generative Adversarial Nets’, in *NIPS*.

Hu, K. *et al.* (2021) ‘Text to Image Generation with Semantic-Spatial Aware GAN’, *ArXiv*, abs/2104.00567.

Mishra, P. *et al.* (2020) ‘Text to Image Synthesis using Residual GAN’, *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pp. 139–144.

Nasr, A., Mutasim, R. and Imam, H. (2021) ‘SemGAN: Text to Image Synthesis from Text Semantics using Attentional Generative Adversarial Networks’, *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pp. 1–6.

Qiao, T. *et al.* (2019) ‘MirrorGAN: Learning Text-To-Image Generation by Redescription’, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1505–1514.

Salimans, T. *et al.* (2016) ‘Improved Techniques for Training GANs’, in *NIPS*.

Xu, T. *et al.* (2018) ‘AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks’, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324.

Zaidi, A. (2017) ‘Text to Image Synthesis Using Stacked Generative Adversarial Networks’, in.

Zhang, M., Li, C. and Zhou, Z.-P. (2021) ‘Text to image synthesis using multi-generator text conditioned generative adversarial networks’, *Multimedia Tools and Applications*, 80, pp. 7789–7803.

Zhu, M. *et al.* (2019) ‘DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis’, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5795–5803.