TEXT-TO-IMAGE EDITING USING GENERATIVE AI WITH CROSS ATTENTION CONTROL

WONG KAI JUN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: **TEXT-TO-IMAGE EDITING USING GENERATIVE AI WITH CROSS ATTENTION CONTROL**

SESI PENGAJIAN: **2022/2023**

Saya: **WONG KAI JUN**

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan unituk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. * Sila tandakan (✓)

______	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
______	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan)
✓	TIDAK TERHAD	

Wong Kai Jun

(TANDATANGAN PELAJAR)

Alamat tetap: 76, Laluan Pakatan Jaya 3
Taman Pakatan Jaya, 31150, Ulu Kinta, Perak

Tarikh: 29/09/2023

(TANDATANGAN PENYELIA)

Assoc. Prof. Dr. Choo Yun Huoy
Nama Penyelia

Tarikh: 29/9/2023

CATATAN: * Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

TEXT-TO-IMAGE EDITING USING GENERATIVE AI WITH CROSS ATTENTION CONTROL

WONG KAI JUN

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Artificial Intelligence) with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I hereby declare that this project report entitled

TEXT-TO-IMAGE EDITING USING GENERATIVE AI WITH CROSS ATTENTION CONTROL

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : Wong Kai Jun Date : 29/09/2023
(WONG KAI JUN)

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of Computer Science (Artificial Intelligence) with Honours.

SUPERVISOR : Date : 29/9/2023
(ASSOC. PROF. DR. CHOO YUN HUOY)

DEDICATION

This final year project is especially dedicated to my supervisor, Assoc. Prof. Dr. Choo Yun Huoy, who helped and guided me to successfully complete this project. Special dedication also to my beloved parents who never failed to give me financial and moral support. Lastly, I also dedicate this project to my friends who have supported me throughout the project development. I appreciate all they have done for helping me develop the project.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my special thanks of graduate to my supervisor, Assoc. Prof. Dr. Choo Yun Huoy for giving me guidance and motivation to complete this final year project. Her guidance and advice carried me through all the stages of developing the project.

Bearing in mind previous, I am using this opportunity to express my deepest gratitude and special thanks to my beloved parents for their continuous support and understanding when undertaking my final year project, despite their busy schedule. At the same time, to all my colleagues and my fellow friends, it was a pleasure to know them all as I was assisted by them in terms of solving problems and collecting ideas.

ABSTRACT

Recently, text-to-image models, quickly garnered attention for their incredible generating potential in both semantics and composition. It can save a significant amount of time and resources in generating realistic and detailed images using the text-to-image model instead of manual artwork creation. However, editing is challenging for these generative models, in the text-based models, even a small modification of the text prompt often leads to a completely different outcome. Hence, this project proposes a prompt editing insights solution for text-to-image editing using cross attention control. The cross-attention maps associated in this solution empower users to grasp the connection between the text prompt and the generated image. This helps users pick an accurate word in image generation and editing. At last, the developed tool is able to provide meaningful editing insight and edit the image accordingly for word within noun word class.

ABSTRAK

Baru-baru ini, model teks-ke-gambar telah mendapat perhatian daripada komuniti awam kerana potensi penjanaan imej yang luar biasa dalam kedua-dua semantik dan komposisi. Ia boleh menjimatkan banyak masa dan sumber dalam menghasilkan imej yang realistik dan terperinci dengan menggunakan model tersebut tanpa menghasilkan imej secara manual. Walau bagaimanapun, penyuntingan adalah mencabar bagi model ini kerana pengubahsuaian kecil dalam teks sering membawa kepada hasil yang berbeza sama sekali. Oleh itu, projek ini mencadangkan penyelesaian pandangan penyutingan untuk penyutingan model teks-ke-gambar dengan menggunakan *cross-attention control*. *Cross-attention maps* yang dikaitkan dalam penyelesaian ini memperkasakan permahaman pengguna dalam hubungan antara teks dan imej yang dijana. Ini membantu pengguna memilih perkataan yang tepat dalam penjanaan dan penyuntingan imej. Akhirnya, alat yang dibangunkan dapat memberikan pandangan penyuntingan yang bermakna dan menghasilkan imej dengan sewajarnya untuk perkataan dalam kelas kata nama.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

CHAPTER 1: INTRODUCTION

1.1 Introduction

Text-to-image model is type of generative AI model which takes a natural language description input and produces an image matching that description. Recently, Text-to-image models, such as Imagen, DALL·E 2 and Stable Diffusion, quickly garnered attention for their incredible generating potential in both semantics and composition. It has numerous potential use cases in various industries that are needed for visual content creation, including E-commerce, Gaming, Advertising, etc. It can save a significant amount of time and resources in generating realistic and detailed images using the text-to-image model instead of manual artwork creation.

1.2 Problem Statement

Editing poses a considerable challenge for generative models, particularly in text-based ones, where even minor alterations to the input text prompt can yield drastically different results. Current state-of-the-art techniques address this issue by requiring users to explicitly specify a region of the image that needs inpainting. Then guide the edited image to change exclusively within the marked area while preserving the background from the original image. However, this approach is somewhat cumbersome and overlooks the original structure and content within the masked region. Consequently, it limits the practical utility of text-to-image models across various industries by constraining their customization and editing capabilities in generating the desired images.

1.3 Objective

- To propose a prompt editing insights solution for text-to-image editing using cross attention control

- To develop a text-to-image generator based on prompt editing with insights.

- To evaluate prompt editing with insights solution through the demonstration tool.

1.4 Scope

The project scope is focused on image generation using text-to-image generative model, providing image editing insights to the users, and image editing based on the concept of prompt editing.

1.4.1 Module to be developed.

1. **Web application:** Platform for targeted users to access the developed system.

2. **Image generation module:** Text-to-image generative AI module.

3. **Image editing module:** Image Editing module based on the concept of prompt editing.

1.4.2 Targeted User

Individuals that used text-to-image model to generate images and edit those generated images.

1.4.3 Limitation

- **Tool Capability:** Graphics editor functionality is not available.

- **Image Generation:** The model may not be able to generate images that have never been trained before.

- **Prompt Language**: The model was trained mainly with English captions and will not work as well in other languages.

- **Image Editing:** The prompt editing is only able to edit images generated by the system, and a single word can be replaced from the original input prompt only for each editing.

- **Performance:** System may have performance issues due to limited resources (based on stated hardware requirement).

1.5 Project Significance

- **Provide Insight:** Prompt editing insights provide user guidance in editing images and guide user to input correct prompt to generate desired images.

- **Image Editing:** No graphics editing skill is required to edit the generated images.

- **Resource and Time**: Stable Diffusion saves a significant amount of time and resources in generating graphical content compared to manual artwork creation.

1.6 Expected Output

The end-product of the project should be able to:

- Generates images based on user’s text input.

- Provide prompt editing insights by visualizing the cross-attention maps.

- Generate the edited images accurately based on user text input.

1.7 Conclusion

This chapter includes the project background in brief, problem statement, objectives, scopes, significances and expected output of the purposed project. Next chapter will explain literature review and project methodology.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter will preview the literature review and project methodology by identifying facts and findings, including relevant domain, comparing existing systems to system to be developed as well as discussing the most appropriate techniques to use.

2.2 Generative Artificial Intelligence (GAI)

Machine learning models can be categorized into two main types: discriminative and generative. Discriminative models make predictions on unseen data by considering conditional probabilities and are suitable for tasks such as classification or regression. Examples of discriminative models include logistic regression, decision trees, and certain deep neural networks like image classifiers. In contrast, generative models focus on understanding the underlying data distribution to return a probability for a given example. This is a more complex task, but it enables the generation of new samples, such as images. A proficient generative model aims to maximize the likelihood of the data distribution or approximate it when exact computation is challenging.

Generative Artificial Intelligence (GAI) harnesses the power of generative modeling and the advancements in deep learning (DL) to create a wide range of content at scale, including text, graphics, audio, and video, by building upon existing media sources (Gui, et al., 2020). It excels at producing highly realistic and intricate content that closely emulates human creativity, thus finding utility in numerous applications.

The evolution of GAI spans several stages in its history. It had its origins in early AI research during the 1950s and 1960s when rule-based and expert systems were developed to generate content (Flemming, et al., 1986). However, these early systems had limitations in terms of creativity. Subsequently, in the 1990s and 2000s, probabilistic models like Hidden Markov Models and Markov Random Fields were explored for generating sequences and images. While these models captured data dependencies, they struggled with generating truly novel content. The transformative breakthrough for GAI occurred in the 2010s with the ascent of deep learning.

In 2014, Generative Adversarial Networks (GANs) (Goodfellow, et al., 2014) were introduced, revolutionizing generative modeling by pitting a generator against a discriminator to produce realistic content. Variational Autoencoders (VAEs) (Kingma & Welling, 2013) gained popularity around 2014-2017, using latent variable spaces to generate meaningful samples. Language models like GPT-2 and GPT-3 (Radford, et al., 2018) showcased impressive text generation capabilities from 2017-2019. Recent focus has been on multimodal generative models that generate diverse outputs across domains (Srivastava, 2023). Ongoing research aims to tackle challenges in control, interpretability, and ethics, propelling Generative AI toward even more sophisticated applications in the future.

Currently, there are several notable approaches within the realm of Generative Artificial Intelligence which include GANs, VAEs, flow-based models, and finally diffusion models. It should be noted that these approaches mentioned are not mutually exclusive, and recent models have started to combine various concepts. For instance, autoregressive diffusion models (Hoogeboom, et al., 2021), or denoising diffusion GANs (Xiao, et al., 2021). A brief description for each of the generative models, as well as their upsides and downsides are provided here.

2.2.1 Generative Adversarial Networks (GANs)

GANs, which stands for Generative Adversarial Networks, are comprised of two key components: a generator and a discriminator. The generator's task is to create synthetic data that closely resembles real data, whereas the discriminator's role is to distinguish between authentic and counterfeit data. These two components engage in adversarial training, essentially entering into a competitive relationship. The generator's objective is to continually improve its ability to generate highly realistic data, while the discriminator endeavors to precisely determine whether the data it encounters is real or generated.

GAN finds extensive use in deep learning for tasks like data augmentation and data pre-processing in various domains such as image processing and biomedicine. Nevertheless, a notable challenge is the occurrence of mode collapse, which results in limited diversity among generated samples (Zhao, et al., 2018). Additionally, the inherent adversarial process of training makes the latter inherently unstable and frequently demanding (Roth, et al., 2017).

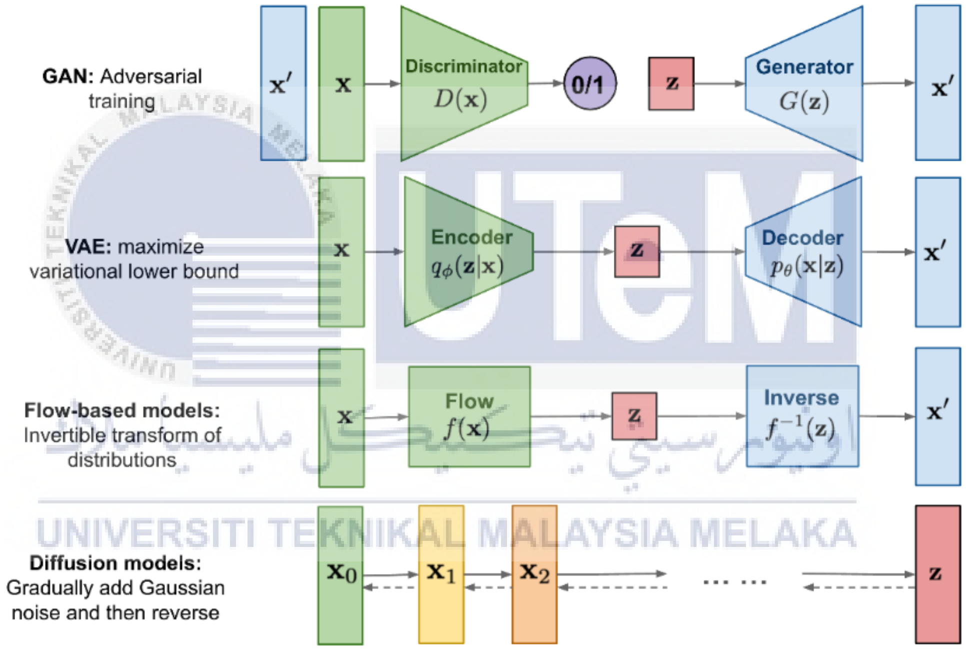
GAN: Adversarial training
x'
x
Discriminator $D(x)$
0/1
z
Generator $G(z)$
x'
VAE: maximize variational lower bound
x
Encoder $q_\phi(z|x)$
z
Decoder $p_\theta(x|z)$
x'
Flow-based models: Invertible transform of distributions
x
Flow $f(x)$
z
Inverse $f^{-1}(z)$
x'
Diffusion models: Gradually add Gaussian noise and then reverse
x_0
x_1
x_2
… …
z

Figure 2.1 Overview of different types of generative mode (Weng, 2021)

2.2.2 Variational Auto-Encoders (VAEs)

VAE, or Variational Autoencoder (Kingma & Welling, 2013), provides a streamlined method for capturing essential low-dimensional insights from data, which can then be harnessed to produce fresh samples by manipulating these acquired low-dimensional representations using a decoder. The VAE framework comprises two primary elements: an encoder and a decoder. The encoder employs a neural network to extract both the mean and variance of the latent variables that dictate the characteristics of the data.

In contrast to traditional autoencoders (Michelucci, 2022), which compress the image into a single point in the latent space only. Instead, VAEs enforce the latent variables to follow a standard normal distribution. By shifting from a single point to a distribution, the encoder in VAE is allowed to learn a smooth latent state representation of the input data, allowing for coverage of unseen samples in the input data. While VAEs often achieve high log-likelihood values, they encounter challenges in generating non-blurry high-quality samples.

2.2.3 Flow-based generative models

Flow models are designed to establish a mapping between the distribution of data representations and the distribution of latent variables. This mapping is achieved by employing a series of reversible transformations (Dinh, et al., 2014; Rezende & Mohamed, 2015). In contrast to GANs and VAEs, which aim to approximate the true data distribution, flow models focus on directly solving the mapping transformation between these two distributions through the manipulation of the Jacobian determinant.

By sampling from the distribution of latent variables, flow models enable the generation of new content. Two noteworthy characteristics define flow models: Firstly, they have the capability to compute the distribution of latent variables that govern data representation, ensuring that the generated content aligns with the distribution of the training data. Secondly, the entire model is constructed using a series of reversible operations, transforming the data into a prior distribution through a bijective function. This unique approach enhances the interpretability of the model.

Notably, unlike other generative models, flow models rely solely on a reversible encoder for model construction, leading to intricate design considerations. However, this design choice introduces challenges such as increased parameterization and higher computational costs, which remain critical areas requiring further attention and resolution.

2.2.4 Diffusion models

Denoising Diffusion Probabilistic Model (DDPM) (Ho, et al., 2020), upon its release, outperformed Generative Adversarial Networks (GANs) (Dhariwal & Nichol, 2021) in the realm of image synthesis. DDPM accomplishes this by learning the distribution of crucial parameters with the help of a neural network.

During the training process, the model introduces incremental noise to the original data in a forward pass. At each step, the model updates its parameters through a Markov process until the data undergoes a transformation into a state represented by pure Gaussian noise. Following the forward pass, the model initiates a backward pass, during which the acquired representation is gradually purified through decoding, ultimately resulting in the generation of new data.

The primary objective of the diffusion model is to enhance the similarity between the newly generated data and the original data, resulting in continuous optimization of the model's parameters. This ongoing improvement leads to higher quality of the generated content. In comparison to GANs, diffusion models offer greater stability during training and surpass Variational Autoencoders (VAEs) in generating a diverse set of high-quality samples. However, the original diffusion model faced three main limitations: slow sampling speed, subpar maximum likelihood, and weak data generalization ability.

2.3 Text-to-image Generative Models

Text-to-image models fall under the category of Generative AI (GAI) models that have the remarkable capability to take textual descriptions or inputs and produce corresponding images illustrating the described content. In recent times, Text-to-image models like Imagen (Saharia, et al., 2021), DALL·E 2 (Ramesh, et al., 2021), Parti (Yu, et al., 2022), and Stable Diffusion (Rombach, et al., 2021) have garnered substantial public attention due to their extraordinary ability to generate semantically rich and compositionally intricate images. These models undergo training using immensely large language-image datasets and employ cutting-edge image generative techniques, including autoregressive models, diffusion models, and GANs.

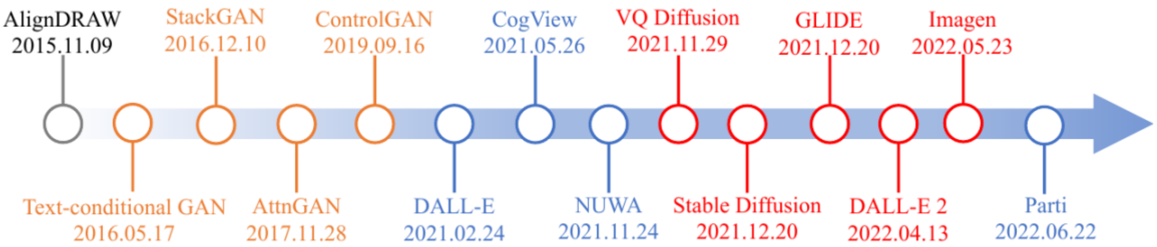

Figure 2.2 Representative works on text-to-image task over time (Zhang, et al., 2023).

The concept of generating images based on textual captions originated from the initial research on AlignDRAW (Mansimov, et al., 2016). However, the early attempts yielded poor image quality, with the generated scenes and objects being hardly recognizable. Subsequently, Text-conditional GAN (Reed, et al., 2016) is introduced as the first end-to-end differential architecture, operating at both the character and pixel levels. Subsequent to that, there were five years of gradual advancements driven by the progress of GANs, which led to the development of models like AttnGAN (Xu, et al., 2017) and ControlGAN (Li, et al., 2019). These models showed some progress in representing certain aspects of the captions, but the generated images still lacked realism, except for specific and simple datasets like the CUB dataset, which focused solely on bird images. By incorporating contrastive learning into the pipeline and particularly by expanding the size of the dataset, XMC-GAN (Zhang, et al., 2021) was able to generate improved images that depicted clearer scenes.

Likewise, in 2021, Ramesh et al. showcased the potential of scaling up their dataset to a staggering 250 million image-text pairs in their creation, DALL-E. This expansion enabled the remarkable feat of zero-shot learning. What sets DALL-E apart is its capacity to blend various objects, concepts, and locations to generate unconventional and non-iconic images, such as an "avocado chair." Notably, DALL-E takes a departure from conventional GAN-based approaches and instead employs a combination of a VQ-VAE, as introduced by Oord et al. in 2017, along with two Transformers, capitalizing on the success of the Transformer model.

In the same year, (Dhariwal & Nichol, 2021) demonstrated that diffusion models could surpass GANs in the generation of class-conditional images, building upon this achievement. Subsequently, GLIDE (Nichol, et al., 2022), extended