NOVEL TEXT-TO-IMAGE SYNTHESIS MODELS: OBJ-SA-GAN AND SWINV2-IMAGEN

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF COMPUTER AND INFORMATION SCIENCES

Supervisor

Dr. Weihua Li

Prof. Quan Bai

Dr. Yi Yang

January 2023

By

Ruijun Li

School of Engineering, Computer and Mathematical Sciences

Abstract

The rapid development of deep learning techniques has considerably boosted the development of text-to-image synthesis. Nowadays, a large number of generation models based on deep learning algorithms have emerged in the field of image generation. In this thesis, I recognise the importance of text semantic layout and propose two novel generation models, Obj-SA-GAN and Swinv2-Imagen. These two models use different algorithms to mine the text semantics, and both improve the quality of the generated images compared to baselines.

In recent years, text-to-image synthesis techniques have made considerable breakthroughs, but the progress is restricted to simple scenes. Such techniques turn out to be ineffective if the text appears complex and contains multiple objects. To address this challenging issue, I propose a novel text-to-image synthesis model called Object-driven Self-Attention Generative Adversarial Network (Obj-SA-GAN), where self-attention mechanisms are utilised to analyse the information with different granularities at different stages, achieving full exploitation of text semantic information from coarse to fine. Complex datasets are used to evaluate the performance of the proposed model. The experimental results explicitly show that our model outperforms the state-of-the-art methods. This is because the proposed Obj-SA-GAN model improves text utilisation and provides a better understanding of complex scenarios.

In addition, diffusion models have been proven to perform remarkably well in text-to-image synthesis tasks in a number of studies, immediately presenting new study

opportunities for image generation. Google's Imagen follows this research trend and outperforms DALLE2 as the best model for text-to-image generation. However, Imagen merely uses a T5 language model for text processing, which cannot ensure learning the semantic information of the text. Furthermore, the Efficient-Unet used by Imagen is not the best choice for image generation. To address these issues, I propose the Swinv2-Imagen, a novel text-to-image diffusion model based on a Hierarchical Visual Transformer. In the proposed model, the feature vectors of entities and relationships are extracted and involved in the diffusion model, effectively improving the quality of generated images. On top of that, I also introduce a Swin-Transformer-based Unet architecture, called Swinv2-Unet, which can address the problems stemming from the CNN convolution operations. Extensive experiments are conducted to evaluate the performance of the proposed model by using three real-world datasets, i.e., MSCOCO, CUB and MM-CelebA-HQ. The experimental results show that the proposed Swinv2-Imagen model outperforms several popular state-of-the-art methods.

Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of student

Publications

Li, R., Li, W., Yang, Y., Bai, Q. (2022). Obj-SA-GAN: Object-Driven Text-to-Image Synthesis with Self-Attention Based Full Semantic Information Mining. In: Khanna, S., Cao, J., Bai, Q., Xu, G. (eds) PRICAI 2022: Trends in Artificial Intelligence. PRICAI 2022 [1]. Lecture Notes in Computer Science, vol 13629. Springer, Cham. https://doi.org/10.1007/978-3-031-20862-1_25

Li, R., Li, W., Yang, Y., Wei, H., Jiang, J., and Bai, Q. (2022). Swinv2-Imagen: Hierarchical Vision Transformer Diffusion Models for Text-to-Image Generation. ArXiv, abs/2210.09549. https://doi.org/10.48550/arXiv.2210.09549

[1] https://pricai.org/2022/ The 19th Pacific Rim International Conference on Artificial Intelligence

Acknowledgements

This research work was completed as the part of the Master of Computer and Information Sciences (MCIS) course at the School of Computer and Mathematical Sciences (SCMS) in the Faculty of Design and Creative Technologies (DCT) at the Auckland University of Technology (AUT) in New Zealand. I would like to deeply thank my parents for the financial support they provided during my entire time of academic study in Auckland. The thesis is completed with the elaborate supervision of my supervisor, Dr. Weihua LI. My supervisor's substantial experienced knowledge, meticulous academic attitude, perfect working style and approachable personality have a deep impact on me. This thesis is completed under the guidance of my supervisor from the topic selection to the thesis completion, and he has devoted a lot of effort to it. I would like to express my deepest respect and gratitude to my supervisor. I believe that without his invaluable help and supervision, I would not be able to successfully complete my academic thesis. Furthermore, I also would like to express my sincere gratitude to A/Prof. Quan Bai, who is from the University of Tasmania, Australia (UTA), and Dr. Yi Yang, who is from Hefei University of Technology, China, for their valuable comments, which have greatly improved this paper.

Intellectual Property Rights

Chapter 1

Introduction

The widespread use of products and applications derived from the Internet has led to a dramatic expansion of unstructured data such as images and videos. Compared to text, images are more intuitive, impactful and compelling in expressing emotions, attitudes and describing matters, and can be used to generate better interaction in social media or presentations (Agnese et al., 2020a). Based on this background, researchers have started to design a model that automatically converts text into images. It extracts the main ideas of the text and represents them using images so that the readers can understand and remember this information more effectively. The text-to-image synthesis task is used to address this demand.

Text-to-image synthesis refers to the generation of a conformed image based on a textual description. Text-to-image synthesis requires a model that can first parse the semantic information in the text, and the resulting image must conform to the semantic representation of the text, which is a very challenging task. However, the emergence of Encoder-Decoder architectures, e.g., Variational Auto-Encoder (VAE) (Pavan Kumar & Jayagopal, 2021), and other deep learning architectures, e.g., Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have facilitated the development of deep learning in areas such as image processing and natural language

processing. Although it is possible to generate images using these techniques, the resulting images do not achieve the desired results in various aspects, such as image quality and diversity.

1.1 Background

1.1.1 GANs-based Text-to-Image generation

Generative Adversarial Networks (GAN)-based generative models have quickly attracted the attention of many researchers since they were first proposed by Goodfellow in 2014 (Goodfellow et al., 2014).The key theory of GAN is to build a competitive pair of generators and discriminators using a multilayer perceptron (MLP). The former generates images using noisy data with the aim of fooling the discriminator as much as possible, while the latter uses iterative learning to improve the ability to identify whether an image is generated or real (Pavan Kumar & Jayagopal, 2021). The GAN-based approach produces clearer and more realistic images than other generated models. These features make GAN models the preferred image synthesis models. However, they still have many open problems, such as model collapse (Esfahani & Latifi, 2019), diversity (Agnese et al., 2020a) and gradient disappearance (Arjovsky & Bottou, 2017).

To address these issues, many variants of GAN have been generated, such as CGAN (Mirza & Osindero, 2014), SAGAN (H. Zhang, Goodfellow, Metaxas & Odena, 2019), Stack-GAN (H. Zhang et al., 2017), AttnGAN (Xu et al., 2018) and AC-GAN (Odena, Olah & Shlens, 2017). After several years of development, GAN has now become extremely powerful and has been widely used in image synthesis, image restoration, style migration and many other research tasks. In text-to-image synthesis tasks, numerous experiments have shown that GAN-based generative models can produce excellent results on simple datasets, such as CUB (M. Zhu, Pan, Chen & Yang,

2019). However, on complex datasets with multiple objects and relationships, they still fail to achieve satisfying results. It is suspected that this is mainly due to the fact that GAN may miss some significant textual information and does not accurately extract the semantic information from the text. Based on this conjecture, an Object-driven Self-Attention GAN model (Obj-SA-GAN) is presented in Chapter 4. The self-attention is leveraged to analyse the text with different granularity at the different stages of the text processing stages, achieving full exploitation of the text. Then a reliable and fine-grained semantic layout is generated and used to guide the GAN module to generate images. Finally, the proposed model is evaluated on the MSCOCO dataset and proves our conjecture.

1.1.2 Diffusion-based Text-to-Image generation

The study organised by Prafulla et al. (Dhariwal & Nichol, 2021) in 2021 demonstrated that the diffusion model outperforms the current state-of-the-art GAN-based generative model for image synthesis. It immediately attracts the attention of many researchers and presents a new research opportunity for text-to-image synthesis. Both OpenAI's DALLE2 (Ramesh, Dhariwal, Nichol, Chu & Chen, 2022) and Google's Imagen (Saharia et al., 2022) are built on diffusion models, and both have achieved excellent results on complex datasets. DALLE2 first maps the text description to a representation space through a pre-trained CLIP (Radford et al., 2021) model to obtain a text embedding. Then, the text embedding is mapped to an image embedding through a priori model constructed with a diffusion model. In this stage, the model analyses the semantic information contained in the text embedding. Finally, the image embedding is parsed into an image using an image decoder constructed with a diffusion model.

Imagen discards the priori model and is a simpler generative model. Imagen simply uses a T5 text encoder (Raffel et al., 2020) pre-trained on a large plain text corpus. The

resulted text embedding is then used to generate an image using a base diffusion model and two super-resolution diffusion models. Although Imagen outperforms DALLE2 in terms of text-image alignment, image quality, etc., the insight gained is that Imagen cannot ensure that the model fully understands the semantic information of the text with only a text encoder. In addition, Imagen's Efficient-Unet module for building two super-resolution diffusion models contains multiple CNN blocks. Although the Unet architecture accelerates model convergence, it also makes the model restricted by the convolution operations, which may lose global and layout information.

To improve Imagen, a novel image generation model, i.e., Swinv2-Imagen, is proposed in Chapter 5. It uses not only the T5 encoder in the text processing stage, but also introduces a scene graph to analyse the object-relationship in the text. Afterwards, two additional embeddings, object and relationship embeddings, are obtained. These additional embeddings assist the T5 encoder in ensuring that the model understands the semantic information accurately. In addition, a new Unet architecture, i.e., Swinv2-Unet, is presented based on Swin-Transformer-v2 (Liu et al., 2022). It replaces the original CNN blocks with Transformer blocks. The new Unet model leverages attention to explore the relationship between features, allowing the diffusion model to focus on different granularities of features at different moments, from local to global;

1.2 Research Motivations

Text-to-image synthesis is a fascinating research area with a very promising perspective. In this section, the research motivations for this thesis will be presented.

Text-to-image synthesis requires only the input of natural language texts, then the model automatically converts it into a realistic image that meets the conditions. Text-to-image synthesis is currently used in a wide range of real-world applications, such as image edition (Tan, Liu, Li, Zhang & Yin, 2019), generating images for novels, lyrics or

newspapers (Efimova, Jarsky, Bizyaev & Filchenkov, 2022), generating recipe images for cooks (B. Zhu & Ngo, 2020), and assisting designers with fashion design (Jain, Modi, Jikadra & Chachra, 2019).

In the contemporary research field, there are numerous studies describing text-to-image synthesis and its applications. The main research gaps have been identified as follows:

- While there are a large number of research groups working on improving GANs-based generation models, they mainly focus on improving the GAN architecture and solving the problem of GAN models, such as model collapse and result diversity. Semantic layout, an important factor for the image synthesis task, has been neglected in many studies.
- GANs-based text-to-image generation models have achieved excellent results on simple datasets [1], but when the text contains multiple objects or includes complex relationships, the images generated by these models are frequently terrible.
- Diffusion models are still in a rapid development stage in the field of image synthesis. Although it outperforms GANs-based generation models in almost all aspects, studies on diffusion models are still limited to improving sampling efficiency or optimising generated results by combining with other image synthesis models. Very few studies have used semantic information as an explicit input to aid image synthesis.

Based on the aforementioned research gaps, the research motivations of this thesis are summarised as follows:

- Model a GANs-based generation model that can perform well in both simple

[1] Simple dataset means there is just one simple object in each image, such as a bird in CUB dataset or a face in CelebFaces Attributes Dataset

and complex datasets. The model increases text utilisation and generates a fine-grained semantic layout to guide image generation.

- Model a diffusion model, using semantic information as a piece of additional information to facilitate text generation.

1.3 Research Questions

According to the literature review and research motivations, the research questions are summarised as follows:

Research Question 1: How to model a GANs-based synthesis model that generates realistic images in scenes with both simple and complex object relationships?

In the last decade, many variants of GAN have been derived to solve the problems of GAN models, such as model collapse and diversity. These models have facilitated the utilise of the GAN models in industrial applications. However, these models are limited to producing satisfactory results in simple datasets. When multiple objects are included in a text description, these models do not accurately identify each object and the relationships between them, so the results generated are bad. Based on this research gap, one of the main research objectives of this paper is to explore how GAN models can efficiently recognise text objects and their relationships, especially when the text scenes are complex, namely, how to improve text utilisation.

Research Question 2: How to develop a diffusion-based generation model without CNNs but ensure that the model understands the text semantics?

Deep neural networks are black boxes, and there is no way for the user to specify which features the model should learn. In other words, using a large corpus alone to obtain text embeddings cannot guarantee that the model learns the semantic features. Based on this situation, another major research objective of this thesis is to take semantic

information as auxiliary vectors to guide the diffusion model to synthesise semantically consistent images.

In addition, currently, most of the diffusion models are based on Unet architectures with CNNs blocks as the core. These architectures suffer from the limitations of CNNs, and they are not able to extract features efficiently in terms of global and layout.

1.4 Design of study

In this subsection, the research methodology and evaluation methods used in this thesis are explained.

1.4.1 Research Methodology

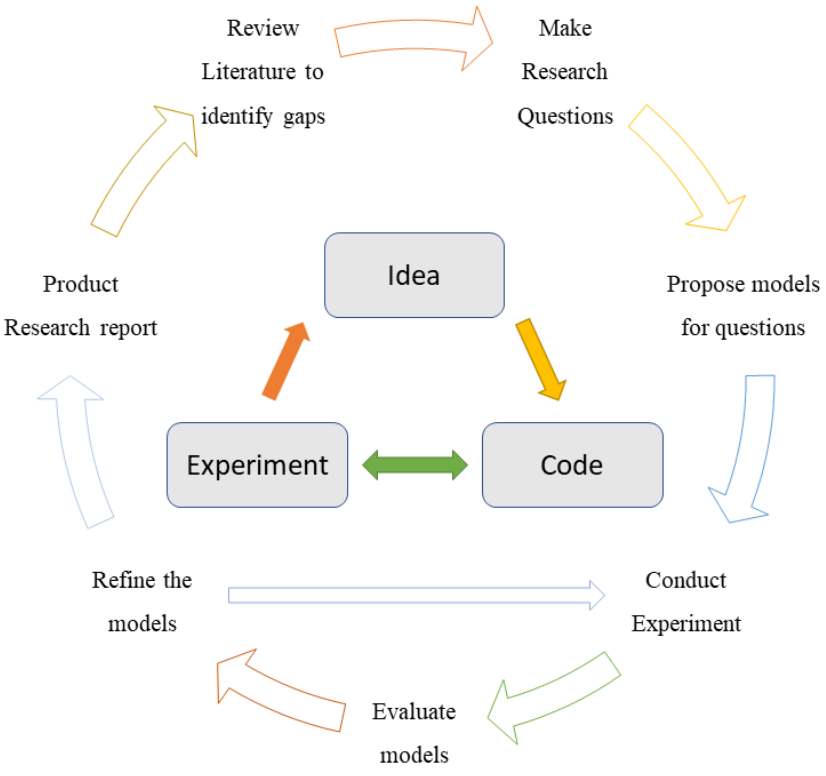

Figure 1.1: Research Methodology

The research methodology utilised in this thesis is modified from (Peffers, Tuunanen, Rothenberger & Chatterjee, 2007), which is an interactive and iterative process. The whole process includes three stages as shown in Figure 1.1: proposing an idea, implementing a model by coding and evaluating the model.

Firstly, I collected a large amount of literature on the field of text-to-image generation. It is the most significant to gain insight into the problems faced by generation models and the research gaps that remain to be filled by reviewing these studies. At the end of this phase, a research proposal is generated, which lists the proposed research questions and the corresponding solutions. Afterward, I implement these models to address these problems by using deep learning frameworks, e.g., PyTorch. Next, I evaluate the model quantitatively and qualitatively on some public datasets, such as MSCOCO. The last two phases are iterative processes, where the details of the model implementation are continuously adjusted based on the evaluation results until the generated results meet the desired outcome. The final step is to consider the thesis structure and write the research thesis.

1.4.2 Evaluation Methods

The evaluation metrics utilised in this thesis are Inception Score (IS) (Barratt & Sharma, 2018) and Fréchet Inception Distance (FID) (Borji, 2022). They are the two most commonly used evaluation measures for the tasks of text-to-image generation.

- **Inception Score (IS):** IS tests two aspects of performance: (1) the clarity of the generated images and (2) the diversity of the generated images. If the generated images are not clear enough, the model is obviously underperforming; if the generated images are clear enough, also need to see if it can generate a sufficient variety of images (Barratt & Sharma, 2018).

- **Fréchet Inception Distance (FID):** The FID score is based on the IS. The

difference between FID and IS is that IS evaluates the generated image directly, the larger the metric the better, whereas FID scores are generated by comparing the generated image with the real image, calculating a "distance value", the smaller the metric the better (Borji, 2022).

1.5 Contributions of the thesis

In this thesis, I recognise the importance of semantic layout for the text-to-image generation tasks and propose two different text-to-image synthesis models, focusing on improving the quality of the semantic layout of the models using different approaches. Based on these models, the key contributions of this thesis are summarised below.

- I propose a GANs-based generation model, i.e., Obj-SA-GAN, which uses self-attention to improve text utilisation. The Obj-SA-GAN is a typical multi-stage synthesis model, it extracts text features with different granularity at different stages, from coarse to fine. The model also addresses the performance issues of GAN models when being utilised in complex scenes. The models and relevant results are published in The 19th Pacific Rim International Conference on Artificial Intelligence (R. Li, Li, Yang & Bai, 2022).

- I propose a diffusion-based generation model, i.e., Swinv2-Imagen, which uses scene graphs to extract semantic information from text. The scene graphs construct complex text as an oriented graph, visualising the semantic relationships of the text. The model leverages attention to explore the relationship between features, allowing the diffusion model to focus on different granularities of features at different moments, from local to global (R. Li, Li, Yang, Wei et al., 2022).

1.6 Thesis Structures

The rest of the thesis is organised as follows:

- **Chapter 2** reviews related studies, such as GANs, diffusion models, scene graphs and Unet.
- **Chapter 3** introduces the related techniques, such as CNN, GANs, and diffusion models. These relevant techniques provide the theoretical foundation for subsequent research.
- **Chapter 4** presents the Object-Driven Text-to-Image Synthesis with Self-Attention Based Full Semantic Information Mining (Obj-SA-GAN) model. The self-attention mechanisms are utilised to analyse the information with different granularities at different stages, achieving full exploitation of text semantic information from coarse to fine. Extensive experiments and ablation experiments are also conducted on MSCOCO to evaluate the proposed model.
- **Chapter 5** presents the Hierarchical Vision Transformer Diffusion Models for Text-to-Image Generation (Swinv2-Imagen) model. Scene graphs are introduced in the text processing stage to parse the text semantic information. The Unet architecture of the diffusion model is also improved.
- **Chapter 6** concludes the thesis by summarising the contributions and limitations of the proposed generation models. Based on these limitations, future works are outlined.

Chapter 2

Literature Review

With the rapid development of computer vision, deep learning has become the most popular research area in recent years. Among various deep learning techniques, GANs have attracted a lot of attention due to their ability to generate realistic images. However, the training of GANs is challenging and requires a large amount of data. To address this problem, researchers have proposed various techniques such as diffusion models. Diffusion models have been used to generate high-quality images by iteratively updating the pixels in the image. In addition, scene graphs have been used to capture the relationships between objects in an image and have been shown to improve the performance of GANs and diffusion models.

In this literature review, we will review recent research papers that have explored these techniques and discuss their contributions and limitations.

2.1 GANs

The generation of realistic images from textual descriptions brings great contributions to many real-world applications, such as healthcare, education, and computer-aided systems. Nowadays, a growing number of generative methods have been proposed

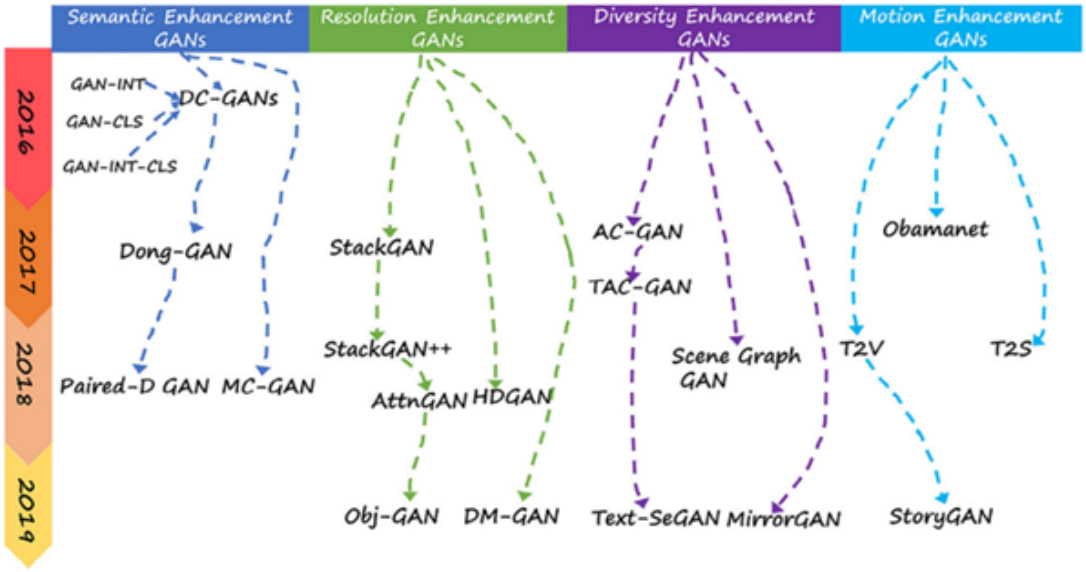

Figure 2.1: Different types of GAN models (Agnese et al., 2020a)

for text-to-image synthesis. There are many existing generative models (Bai & An, 2018; Esfahani & Latifi, 2019; Ghosh, Dutta, Totaro & Bayoumi, 2020; Karpathy & Fei-Fei, 2015; Lee, Ullah, Lee, Jeong & Choi, 2021; Ning, Nan, Xu, Yu & Zhang, 2020; Singh & Raza, 2021; Wu, Xu & Hall, 2017), where GAN-based models outperform the others in terms of the quality of the generated images and the semantic matching of the text (B. Zhu & Ngo, 2020; M. Zhu et al., 2019). However, the standard GAN model does not use mathematical models when building generators but rather a data-driven approach. The data is random, thus, the output of the GAN heavily relies on random vectors, leading to an uncontrolled process of image generation (R. Zhou, Jiang & Xu, 2021a). With the advent of the variants of GAN models, such as Conditional GAN (CGAN), the problem is being alleviated (Frolov, Hinz, Raue, Hees & Dengel, 2021; R. Zhou et al., 2021a).

CGAN and standard GAN are fundamentally the same in terms of architecture. The difference is that both the pair of generator and discriminator in CGAN receive an additionally conditional variable to obtain extra information. Due to the change, the

loss function of CGAN becomes based on conditional probabilities, thus improving the stability of the model (R. Zhou et al., 2021a; Frolov et al., 2021). However, CGAN cannot identify and extract valuable visual elements from complex text descriptions. This is mainly because it uses labels as input, and multiple labels or keywords cannot simultaneously constrain the input, turning out to be a major weakness of CGAN (Agnese, Herrera, Tao & Zhu, 2020b).

Based on CGAN, the researchers have modified and optimised a set of GAN variants, such as Stack-GAN (X. Huang, Li, Poursaeed, Hopcroft & Belongie, 2017), StackGAN++ (H. Zhang et al., 2018) and AttnGAN (Agnese et al., 2020a), which are based on the stacking or attentional architecture. Existing studies reveal that these models can produce high-quality images on simple datasets, such as CUB (Reed et al., 2016) and CelebA (H. Zhang, Goodfellow et al., 2019). However, they do not perform well in complex datasets, such as MSCOCO (M. Zhu et al., 2019), containing multiple objects. Moreover, the GAN-based approach has low text utilisation and loses important fine-grained semantic information. A major challenge in synthesising complex images is to improve the accuracy of identifying relationships between objects. Hong et al. adopt semantic layers to analyse the connections between objects before generating images (Hong, Yang, Choi & Lee, 2018), having two phases, i.e., semantic layer and GAN-based image generation. However, the text message has been encoded into a single text vector, ignoring the fine-grained text information. The resulting images do not have enough details to support the generated results. Similarly, Li et al. design a two-phase model to synthesise images, where an object-driven GAN neural network is introduced by using part of the fine-grained information (W. Li et al., 2019a). However, the improved model misses some important information when generating images, as demonstrated in Figure 4.1 This reveals that their model still suffers from low text utilisation. I propose a new object-driven self-attention framework to improve the utilisation of fine-grained content.

The attention mechanism is much like the logic of seeing a picture, where people's attention is always focused on the important part of the image. This allows the attention mechanism to conserve resources and quickly obtain the most valuable information (F. Wang & Tax, 2016). As the core theory of the most popular deep framework, i.e., Transformer, the self-attention mechanism turns out to be a very effective way to model context, which improves the attention mechanism, reduces the dependence on external information and is better at capturing the relevance within the data (H. Zhang, Goodfellow et al., 2019).

In summary, the classic GAN model maps textual information into a single text vector, ignoring word-level information. Both Hong and Li aim to improve the text utilisation of the model by introducing a semantic layer before image generation to achieve significant results in synthesising multi-object images. Unfortunately, they used an LSTM model based on the RNN architecture for the semantic layer, so some fine-grained information is still missed when parsing the semantic information. However, attention mechanisms, in particular self-attention, can focus limited resources on the detailed information of an object to fully discover hidden connections.

2.2 Diffusion models

Text-to-Image synthesis is a typical application of multimodal and cross-modal comparative learning. In the field of image generation, most models mainly fall into two categories, i.e., the GAN-based generation models (M. Zhu et al., 2019; B. Zhu & Ngo, 2020; H. Zhang, Xu et al., 2019; Xia, Yang, Xue & Wu, 2021; Crowson et al., 2022; Cheng, Wu, Tian, Wang & Tao, 2020) and the diffusion-based models (Ho, Jain & Abbeel, 2020; Ho et al., 2022; Nichol et al., 2022; Rombach, Blattmann, Lorenz, Esser & Ommer, 2022; Song, Meng & Ermon, 2021). The former has been developed over the last few years and widely used in many scenarios, such as medical and image

restoration. The latter has demonstrated outstanding performance over the GAN models, acknowledged as state-of-the-art deep generative models (Saharia et al., 2022; Dhariwal & Nichol, 2021; L. Yang et al., 2022).

Diffusion models and GAN generative models are essentially comparable, both being a process of gradually removing noise. However, in contrast to GAN, the diffusion models do not suffer from training instability and model collapse. The diffusion model transforms the data distribution into random noise and reconstructs data samples with the same distribution (Saharia et al., 2022; H. K. Cao et al., 2022). The diffusion model demonstrates outstanding performance for a number of tasks, such as multimodal modelling. Many contemporary text-to-image synthesis models, e.g., DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022) and GLIDE (Nichol et al., 2022), are constructed based on the diffusion model. They cascade multiple diffusion models to improve the image generation quality step by step. DALL-E 2 uses a priori diffusion model and CLIP Latents to process the text. In contrast, Imagen discards the priori model and replaces it with a large pre-trained text encoder, i.e., T5. Although the T5 model leveraged in the Imagen model improves the understanding of the text, it does not ensure that the model understands the semantic layout of the text, especially in complex sentences containing multiple objects and relationships. As a result, the model will not be able to reproduce some entities or will lose some entity relationships. Therefore, I attempt to model the global semantic layout by adding a scene graph in the text processing. Furthermore, Imagen builds its diffusion model based on Efficient-Unet. Efficient-Unet is not the best choice in image generation tasks, because it contains multiple CNN blocks and leads to a limited view within the CNN kernel window.

2.3 Scene graph and Graph representation learning

A sentence's nature is a linear data structure, where one word follows another (Johnson, Gupta & Fei-Fei, 2018). Usually, when a sentence is complex with multiple objects, it is time-consuming to analyse the sentence directly, and the accuracy of the text-image alignment is not guaranteed. Complex sentences often incorporate rich scene information. Mapping this information into a scene graph can provide an intuitive understanding of the relationships between objects in a sentence (Mittal, Agrawal, Agarwal, Mehta & Marwah, 2019). Previous studies reveal that the performance of multimodal models, such as text-to-image synthesis, is significantly dependent on mining visual relationships (G. Zhu et al., 2022). Scene graphs can provide a high level of understanding regarding scene information (Johnson et al., 2018). Therefore, the scene graph is recognised as a useful representation of images and text. Specifically, each node in a scene graph represents an object, such as a person or an event, and each object has multiple attributes, such as shape. The relationships between objects are denoted by the edges between nodes, which can be an action or a position (Chang et al., 2021). Recently, the scene graphs have been used extensively for tasks such as text-based image retrieval (Johnson et al., 2015; Schuster, Krishna, Chang, Fei-Fei & Manning, 2015), semantic segmentation (Taghanaki, Abhishek, Cohen, Cohen-Adad & Hamarneh, 2020; Jaritz, Vu, de Charette, Wirbel & Pérez, 2020), visual question answering (L. Li, Gan, Cheng & Liu, 2019), image captioning (Gao, Wang & Wang, 2018; X. Yang, Tang, Zhang & Cai, 2019; Zhong, Wang, Chen, Yu & Li, 2020; J. Gu et al., 2019) and image generation (Johnson et al., 2018; Mittal et al., 2019; Y. Li et al., 2019; B. Zhao, Meng, Yin & Sigal, 2019).

In addition, there is no way for an image generation model to manipulate graph-like data such as scene graphs directly, so scene graphs are usually used in conjunction with graph representation learning (Hamilton, 2020). The main objective of graph

representation learning is to extract node and edge contexts from the scene graph and map them to a set of embeddings. Graph representation learning methods can currently be classified into two types, i.e., machine learning based on Random-Walk and deep learning Graph Convolution-based methods (Hamilton, 2020). Node2vec (Grover & Leskovec, 2016) is a typical representative model of the former. It is based on SkipGram (Mikolov, Chen, Corrado & Dean, 2013) theory to learn the embedding of nodes on a graph and optimises the sampling method. It is proposed in related studies (Chen, Wang, Wang & Kuo, 2020) that two sampling methods, Breadth-First Search (BFS) and Depth First Search (DFS), are mainly included when sampling neighbouring nodes in a graph. BFS requires that each sampled node is a direct neighbour of that node. This sampling method results in a graph representation that is more concerned with local information. In contrast, DFS, where each node is sampled to increase the distance to the initial node as much as possible, produces a graph representation that focuses more on global information. Random-Walk-based representation learning (Hamilton, Ying & Leskovec, 2017) is composed of multiple stages, each with different optimisation goals, which is a typical non-end-to-end model. Graph convolution-based methods, e.g. graph convolution neural networks (Johnson et al., 2018), are able to learn both node feature information and structural information via an end-to-end way. It focuses on both local information and global structural features. Graph convolution is extremely applicable to nodes and graphs of topology. It is currently the best choice for graph data learning.

2.4 Unet

Unet is an encoder-decoder architecture, which is scalable in structure (Ronneberger, Fischer & Brox, 2015). The encoding stage of the Unet consists of four downsamples. Symmetrically, its decoding stage is also upsampled four times, restoring the result of the encoder to the resolution of the original image. In contrast to Fully Convolutional

Networks (FCN) (Shelhamer, Long & Darrell, 2015), Unet upsamples four times and uses a jump connection in the encoder and decoder of the corresponding convolution blocks. The jump connection ensures that the final recovered feature map incorporates more low-level semantic features and features at different scales are well fused, allowing for multi-scale prediction. In addition, the four times up-sampling also allows the segmentation map to recover information such as edges more finely. However, Unet also has some shortcomings. For example, Unet++ (Z. Zhou, Siddiquee, Tajbakhsh & Liang, 2018, 2020) argues that it is inappropriate to directly combine the shallow features from the encoder with the deeper features from the decoder in Unet. Direct fusion would potentially lead to semantic gaps. Furthermore, Unet 3+ (H. Huang et al., 2020) maximises the scope of model information fusion and circulation. Each decoder layer in the Unet 3+ fuses small-scale and same-scale feature maps from the encoder with larger-scale feature maps from the decoder, which capture both fine-grained and coarse-grained semantics at full scale.

Many researchers develop a set of Unet variants by improving and optimising the original Unet. For example, ResUnet (Z. Zhang, Liu & Wang, 2018) and DenseUnet (Cai et al., 2020) are inspired by Residual and Dense connections, respectively; each sub-module of the U-Net is replaced with a form having a Residual connection and a Dense connection. There are variants, e.g., MultiResU-Net (Ibtehaz & Rahman, 2020) and R2 U-Net (Alom, Hasan, Yakopcic, Taha & Asari, 2018). All of these models are constructed using multiple convolutional blocks. With the advent of the Transformer, researchers begin to develop the Unet base on the Transformer, such as Swin-Unet (H. Cao et al., 2021). While Swin-Unet mitigates the limitations of CNN convolutional operations, it is likely to suffer from training instability due to the use of the Swin-Transformer block. Swin-Transformer v2 (Liu et al., 2022) is an improvement on Swin-Transformer, which is effective in avoiding training instability and is easier to scale.

2.5 Summary

In this chapter, a detailed review of research related to GANs-based text-to-image synthesis and diffusion-based image generation is presented. The limitations of these models and research gaps are summarised as follows:

(1) GANs-based generation models perform quite well in simple datasets, showing positive business potential, but perform very terribly in data containing multiple objects or complex relationships.

(2) Many studies related to GANs focus on optimising the architecture of generation models, such as stacking multiple GANs. Almost no research attempts to increase text utilisation.

(3) Deep synthesis models seem to be a black box, it is very difficult to ensure that the deep generation models learn the semantic information.

(4) Most Unet architectures rely on CNNs, these models are very difficult to extract global and layout features. Very few studies adopt the Transformer to construct Unet models.

(5) Very few studies focus on enhancing the text semantic information understanding.

(6) There are no research attempts to integrate the scene graphs and diffusion models to optimise the generated results.

The above summarises the research gaps that need to be filled in the field of text-to-image synthesis recently. These limitations are covered by the proposed models in this thesis.

In particular, to address research gaps 1, 2, 3 and 5, I propose an Object-driven Self-Attention GAN model that uses self-attention mechanisms to improve text utilisation. The model generates a fine-grained semantic layout to guide image generation,

theoretically enabling the synthesis of complex images better than baselines. This is the first research work to build a GAN generation model based on a self-attention and semantic layer. In addition, based on research gaps 3, 4, 5 and 6, I propose a Swinv2-Imagen model that leverages scene graphs as auxiliary modules to help the model understand the text semantics more comprehensively. As a result, it effectively addresses the limitations of CNN convolution operations, theoretically enabling the synthesis of images better than baselines. To the best of my knowledge, this is the first research combining a scene graph with a diffusion model in the text-to-image generation field.

Chapter 3

Preliminary - Related techniques

This chapter provides an introduction to the deep learning techniques that will be used subsequently, both in terms of mathematical theory and model architecture. Deep learning is usually based on abstract models represented by neural network models. Once the model has been trained, the user only needs to focus on the inputs and outputs. The model can be thought of as a black box, ignoring the internal workings of the process. CNN, RNN, Self-Attention, GAN and diffusion models will provide a theoretical basis for our subsequent research.

3.1 Convolutional Neural Networks

Traditional multilayer feedforward neural networks adopt a fully connected approach to connect neurons between layers (Sainath, Vinyals, Senior & Sak, 2015; Schwing & Urtasun, 2015). While this connection method works well for low-dimensional data, when dealing with high-dimensional data, e.g., images and videos, the model parameters tend to increase explosively. For example, suppose a grey scale image of size 1000 * 1000 is fed into a fully connected feedforward neural network with only one hidden layer, which includes 100 neurons. The image, firstly, is flattened into

a 1,000,000-pixel vector. Then the number of weight parameters between the input layer and the hidden layer is approximately 100 million. In practice, the input image is usually multichannel, and the number of hidden layers and neurons is relatively large. This not only significantly affects the speed of forward and backward propagation, but also requires higher computational resources from the CPU, GPU and TPU.

Furthermore, the pixels in an image are connected to each other, and if a fully connected neural network is used to process the image data, there is no way to use the location information between the pixels and the flat structure of the image is lost (Ketkar, 2021; Khan, Sohail, Zahoora & Qureshi, 2020). To solve this problem, CNNs have been proposed to process data such as images more efficiently and rationally. CNN is unique in that it is a neural network with a convolution layer and a pooling layer.

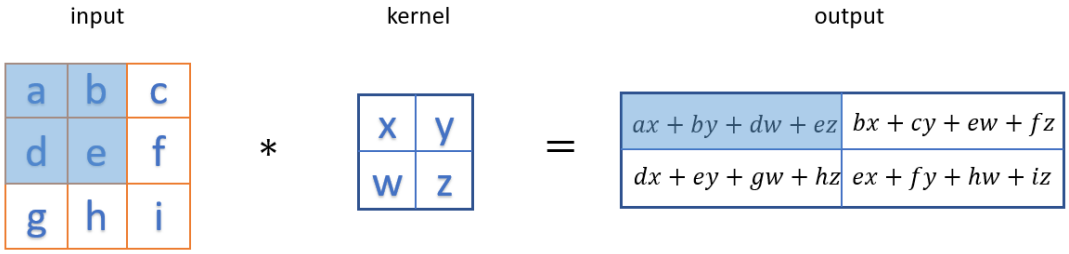

Figure 3.1: Cross-correlation

A convolutional layer is a network layer that contains cross-correlation operations. To perform the cross-correlation operation, a matrix of size k*k, called a convolutional kernel or filter, is given. The size of the convolution kernel determines the size of the region in which the cross-correlation operation is performed. Each convolution kernel focuses on only a local part of the image (Khan et al., 2020). The values in the convolution kernel are called network parameters, which are learnt. Figure 3.1 illustrates the computational process of a two-dimensional cross-correlation operation. It can be seen that the convolution operation is obtained by matrix multiplication of the convolution kernel with the local area corresponding to the input. For example, the

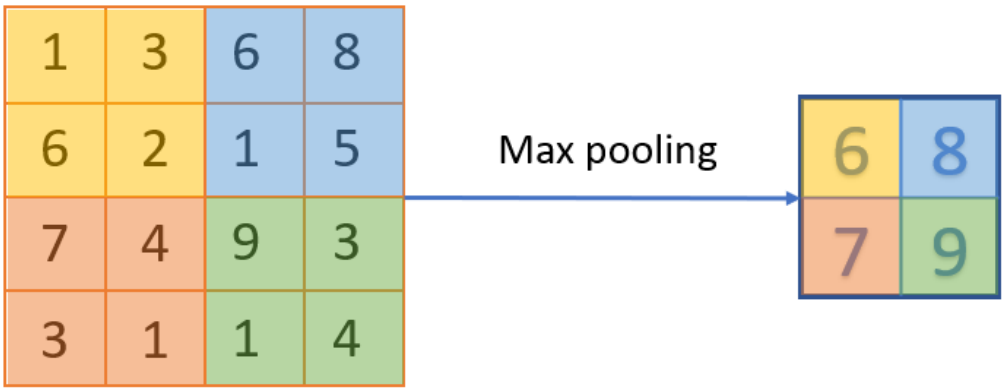

Figure 3.2: Max pooling

first value of the convolution output is $ax + by + dw + ez$. The value of the convolution kernel remains constant throughout the shift of the input matrix, which reflects the idea of sharing weights in the convolution operation (Ketkar, 2021).

A pooling layer is usually added after the convolution layer to reduce the dimension of the feature map generated by the convolution layer and to reduce the computational effort. Two common pooling algorithms are average pooling and maximum pooling. Figure 3.2 shows the two-dimensional maximum pooling layer. There are no learnable parameters in the pooling layer, but it removes some unimportant elements from the feature maps, alleviating the problem of the convolution layer being too sensitive to positional information (Ketkar, 2021; Kamilaris & Prenafeta-Boldú, 2018).

CNN limitations

- CNNs are skilled at image detection, not image understanding. The learning objective of filters in CNNs is to detect features no matter where they are in the image and no matter what changes they have undergone (Albawi, Mohammed & Al-Zawi, 2017). This process only requires that the filter recognise the feature, but it is not necessary that the model understand the feature.

- CNNs consider that a neuron does not need to consider the whole image, i.e. each filter is only responsible for recognising one feature in the image. Thus the image is divided into several different regions of the same size as the filter size (W. Luo, Li, Urtasun & Zemel, 2016). In other words, CNNs focus only on the relationships between pixels within this filter and ignore the global relationships of the image.

3.2 Self-Attention

Attention mechanisms are models that simulate the attention of the human brain (Vaswani et al., 2017). Human visual attention is able to quickly get a global view of things and focus on the most important parts. For example, when reading a text, the reader will first notice the highlighted colours. This way of using limited attention to quickly extract important information greatly improves the efficiency and accuracy of processing information. Self-attention is a variation of the attention mechanism, which aims to identify the critical information for the current task from the global information. It reduces the reliance on external information and is better at capturing the internal relevance of data or features (H. Zhang, Goodfellow et al., 2019). The application of the self-attention mechanism to text focuses on solving the long-distance dependency problem by computing the interactions between words (X. Wang, Tu, Wang & Shi, 2019). Figure 3.3 shows the calculation process for self-attention.

The input embeddings, **a** , when fed into the self-attention layer, are simultaneously computed with three different weight matrices, i.e., W_q, W_k and W_v, to obtain three vectors of query, key and value, i.e., **q**, **k** and **v**.

$$q_i = W_q a_i \tag{3.1}$$

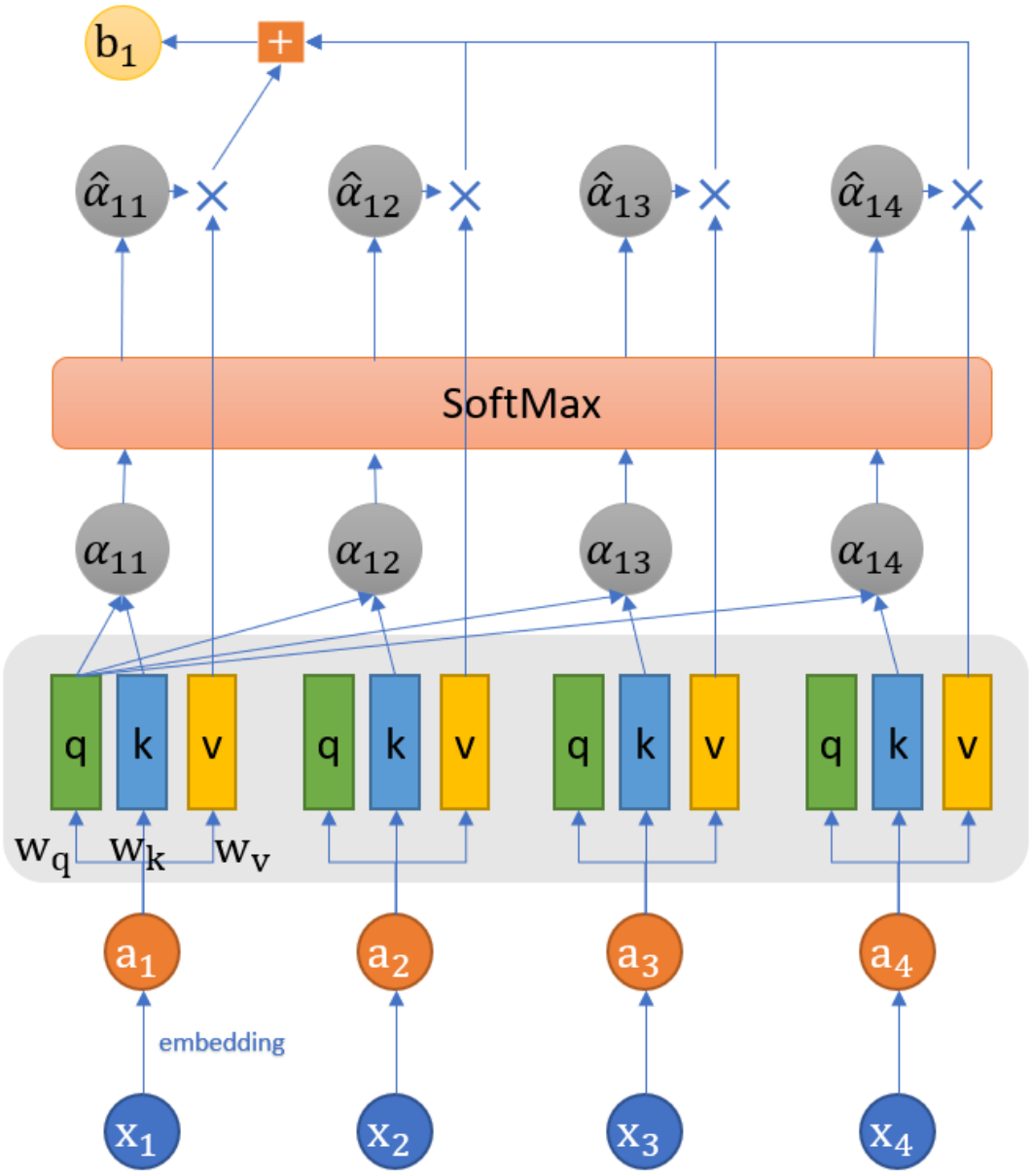

Figure 3.3: Self-Attention

$$k_i = W_k a_i \tag{3.2}$$

$$v_i = W_v a_i \tag{3.3}$$

Each **q** is calculated with all k to obtain an attention score, i.e., α. After SoftMax or other normalisation methods, the attention score vector is weighted and summed with the value vector, **v**, to obtain the corresponding output vector **b**.

$$b_i = \sum_j \hat{\alpha}_{i,j} v_j \quad , \tag{3.4}$$

where $\hat{\alpha}_{i,j}$ refers to the normalised attention score vector calculated form the q_i and the k_j.

3.3 Generative Adversarial Networks

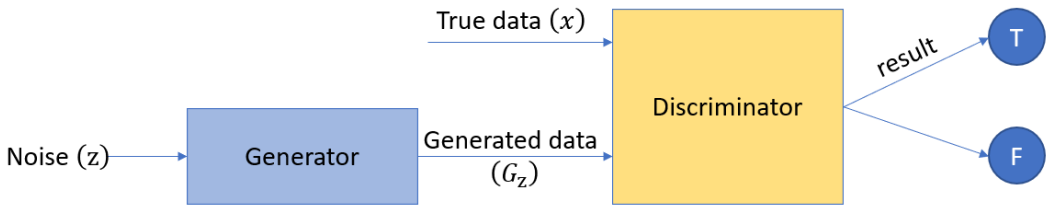

Figure 3.4: GAN

GAN utilises the idea of the zero-sum game (Russell & Norvig, 2010) in game theory to express the image generation problem by using the adversarial game of two networks, the generator and the discriminator. Compared with other generative models, the GAN models can generate more realistic samples (R. Zhou, Jiang & Xu, 2021b). Therefore, in the past few years, GAN-based models have become very popular in the field of image generation. The generator takes as input noisy data that follows a uniform or normal distribution and generates realistic data. A discriminator, on the other hand,

is a binary classifier that attempts to accurately identify whether the input samples are real or generated (Goodfellow et al., 2014).

The GAN model is trained in two stages. Firstly, the generator parameters are fixed and the discriminator is trained. The aim of this stage is to train the discriminator to assign a high score to the real samples and a low score to the generated samples. Mathematically, the purpose of this stage is to maximise Equation 3.5

$$\tilde{V} = \frac{1}{m}\sum_{i=1}^{m}\log D(x^i) + \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(\tilde{x}^i)) \quad , \tag{3.5}$$

where, m refers to the number of examples. D and G mean the discriminator and generator respectively. $\tilde{x}^i = G(z^i)$.

Next, the discriminator parameters are fixed and the generator parameters are updated so that the generated samples can fool the discriminator. Mathematically, the purpose of this stage is to maximise 3.6

$$\tilde{V} = \frac{1}{m}\sum_{i=1}^{m}\log D(G(z^i)) \quad . \tag{3.6}$$

3.4 Diffusion Models

The diffusion model is also a generation model, which can also be described as an Encoder-Decoder architecture. Figure 3.5 shows the architecture of the diffusion model. It first adds Gaussian noise (ϵ) to the original image $(x_0 \sim q(x_0))$ in an iterative manner, the number of iterations being T (T is timestep, usually $T = 1000$). When T tends to infinity, i.e., $(T \to \infty)$, the image is nearly a random Gaussian noise distribution x_T. This process is called forward diffusion and can be thought of as an encoder. The model then learns how to recover the noise distribution (x_T) to the original image $(x_0 \sim q(x_0))$ by gradually removing the noise from x_T. The process is called reverse

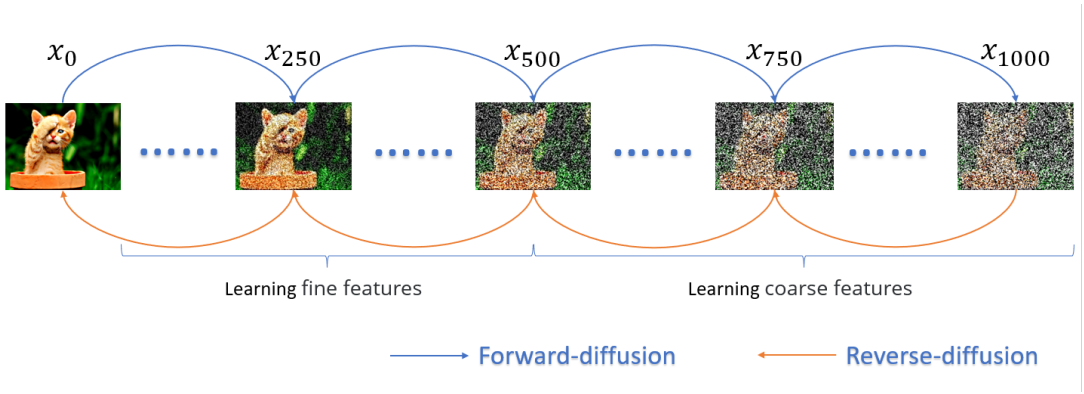

Figure 3.5: Diffusion model

diffusion and can be thought of as a decoder (L. Yang et al., 2022; Dhariwal & Nichol, 2021).

In the forward diffusion, the result at timestep t is mainly related to the outcome at moment $t-1$ and the added noise ϵ_t, i.e.,

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \quad q(x_t|x_{t-1}) \sim \mathcal{N}(\sqrt{1-\beta_t}, \beta_t), \tag{3.7}$$

$$q(x_{1:T}|x_0) = \prod_{i=1}^{T} q(x_t|x_{t-1}) \tag{3.8}$$

where β_t I prefer to understand as a linear weight value. At different timestep, x_{t-1} and ϵ_t have different effects on the result. When T is small, e.g., $t = 1$, x_{t-1} has a greater impact on the result and adds little noise. Conversely, when t is large, e.g. $t = 900$, more noise is added and the contribution to the result is larger than x_{t-1}.

The distribution of the noise added at each timestep in the forward process is identical, i.e., $\epsilon_1, \epsilon_2, \ldots\ldots \sim \mathcal{N}(0, \mathbf{I})$. Thus, we can compute the result at any timestep x_t directly from x_0, i.e.,

$$x_t = \sqrt{\bar{\alpha}_t}x_o + \sqrt{1-\bar{\alpha}_t}\epsilon_t \tag{3.9}$$

where $\alpha = 1-\beta$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The detailed calculation process is available in Appendix

A.

Reverse diffusion is an image-generation process. The Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ will be taken as input to infer and reconstruct the true sample by sampling from distribution $q(x_{t-1}|x_t)$, i.e.,

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta(x_t, t)) \tag{3.10}$$

where $f_\theta(x_t, t)$ is a function used to predict the noise ϵ added in the forward diffusion. This is mainly because it is difficult to infer the true distribution of the image directly from the random noise x_T. The detailed calculation process is available in Appendix A. In other words, the objective of the diffusion generation model is to evaluate the difference between the predicted noise data and the true added noise data, i.e.,

$$p(x_{t-1}|x_t) = ||\epsilon - f_\theta(x_t, t)|| \quad . \tag{3.11}$$

The generation process of the diffusion model is from coarse to fine. Coarse features, e.g., object outline, are learned at moments near x_T. Other fine features, e.g., object colour, are learned at moments near x_0. The model parameters are shared throughout the generation process at all times, and there is no way for the model alone to distinguish which features should be learned at the current time t. Therefore, a time embedding is usually added when training the model, which tells the model which features should be learned at this moment (Ho et al., 2020). The function of the time embedding is similar to the position embedding in the Transformer.

The above is a basic introduction to the diffusion model. Although the diffusion models have yielded very good results in a number of applications such as text-to-image synthesis, they are currently extremely slow in training and inference compared to the GAN-based models. In general, the number of inference steps of a diffusion model is

positively correlated with its time step T.

Chapter 4

Obj-SA-GAN: Object-Driven Text-to-Image Synthesis with Self-Attention Based Full Semantic Information Mining

4.1 Introduction

With the explosive growth of information and the development of social media, people are inundated with information nowadays. Image can deliver the core information in a more effective way to the users than text-based information (Agnese et al., 2020a). People also prefer to perceive image information rather than reading text. Hence, images play an increasingly indispensable role in the current information delivery process. However, most available high-quality images, such as cookbooks and movie posters, are created manually, turning out to be inefficient and expensive (Shamsolmoali et al., 2021). Motivated by this demand, it is significant to investigate how the machines can understand the semantic information in text and generate high-quality creative

images.

Text-to-image synthesis aims to address this problem. It is a technique that automatically generates images based on textual information. Text-to-image synthesis encompasses two key research areas, i.e., Computer Vision (CV) and Natural Language Processing (NLP) (Pavan Kumar & Jayagopal, 2021). The task of text-to-image synthesis typically includes two stages. First, the semantic sense is parsed from the text message, which directly determines whether the generated image satisfies the conditions given in the text message. Second, a generative model is utilised to synthesise a matched image from the parsed semantic sense (Frolov et al., 2021). There are a number of existing text-to-image synthesis models, and they have achieved remarkable success in many areas, such as medical image generation and computer-aided systems (Singh & Raza, 2021).

In the contemporary research field, there are a few dominant methods for the text-to-image task, including VAE, Deep Recurrent Attention Writer (DRAW), and approaches based on GAN (R. Zhou et al., 2021a). Specifically, VAE adopts statistical techniques to build the model and calculate the mean square error between the generated and genuine images (Pavan Kumar & Jayagopal, 2021). DRAW is developed based on CNN and attention mechanisms. However, the resolutions of the images generated by these models are not clear enough to attain the desired results (Frolov et al., 2021; Y. Zhou & Shimada, 2021). By contrast, GAN-based models can generally perform better (Agnese et al., 2020a; Frolov et al., 2021; Pavan Kumar & Jayagopal, 2021; R. Zhou et al., 2021a). The GAN model and its variants take simple text information as input and generate a high-quality image that matches it exceptionally well. However, such models are merely limited to simple datasets, which have only one object in each image, such as faces (Y. Zhou & Shimada, 2021), birds (M. Zhu et al., 2019) and flowers (Reed et al., 2016).

When textual information becomes more complex, having multiple objects in the

text message, the GAN based models are likely to miss pivotal fine-grained information in the generation process, e.g., word-level semantic information. This leads to significant quality degradation of the generated images and the produced results fail to match the given semantic conditions (W. Li et al., 2019a). For example, synthesising an image from the sentence "a woman is sitting on a chair at a table with a cup and cell phone" requires the generative model to achieve two objectives. First, it needs to identify all the objects, i.e., woman, chair, table, cup and cell phone. Second, it needs to rationalise the relationships between the objects, e.g., the woman sitting on the chair, the cup and the phone in her hands.

GAN models generally do not work well on complex images because they focus on learning the overall features of the images without paying attention to the corresponding objects. Taking a living room image as an example, GAN models fail to distinguish between the table and the bed in it but merely place some shapes and colours in a particular position of the synthesised images. In other words, after training, the model does not really understand the image but only remembers where to place some appropriate shapes or colours. This also explains the reason for lacking clear details when synthesising complex images (M. Zhu et al., 2019). Therefore, it is challenging to deal with the relationships between objects when synthesising complex graphs. To alleviate this problem, some researchers developed the idea of analysing the relationships between objects specifically through an additional semantic layer before generating the images, where the image synthesis phase is based on the result of the semantic layer (Hong et al., 2018; W. Li et al., 2019a). These models achieve improvements, but some important image features are missing. For example, when generating images from the sentence "a brown dog lying on bed with his banana toy", the banana toy was not synthesised.

In this chapter, I recognise the importance of semantic layout for complex image synthesis and propose a generative model, namely, Object-Driven Self-Attention GAN (Obj-SA-GAN). It leverages the self-attention mechanism to analyse text and then

Figure 4.1: Synthesised images using object-driven image synthesis models (W. Li et al., 2019a)

uses it to guide image synthesis. Self-attention has two outstanding advantages over other architectures, i.e., RNN. Self-attention extracts features from text sequences by treating the input x_i as key, value and query simultaneously, which can understand the elements in the sequence better (Vaswani et al., 2017). Furthermore, the longest path for self-attention is $O(1)$, implying that any pair of words in a text are directly connected by an individual calculating step. Thus, the distance between long-distance dependent features is dramatically minimised, promoting the effective utilisation of these features. It is understandable to explain the proposed model using a metaphor. Suppose the entire model is likened to a person who intends to draw a picture based on a text description. He or she first analyses which objects are contained in the text and where they should be placed. This function is similar to the box generator. Then, paying their attention to each object and consider full details, such as shape or colour (shape generator). Once this information has been considered, the final step is to draw the picture (image generator). The whole process inherits from the attention architecture, i.e., when thinking about an object, pay most of the attention to the object.

The main contributions of this chapter are shown below.

- Obj-SA-GAN model for the task of text-to-image generation is proposed, which adopts Self-Attention to enhance generated semantic layout.

- The model is evaluated on the complex dataset MSCOCO, and the experimental

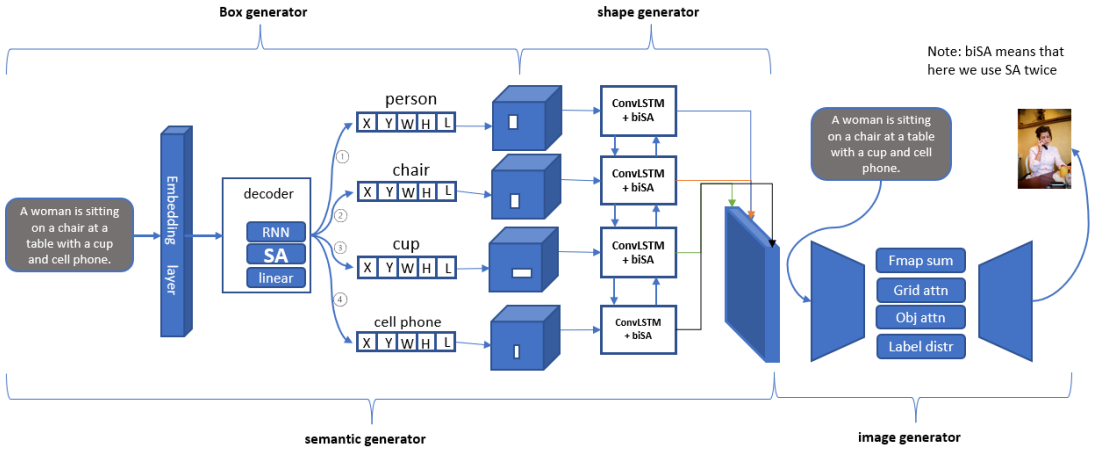

Figure 4.2: The overall architecture of the proposed Obj-SA-GAN model.

results show that the proposed model outperforms the current popular generative models in terms of FID metrics and reaches a new milestone.

- The proposed model also addresses the performance issues of GAN models when being utilised in complex scenes.

The rest of the chapter is organised as follows. In subsection 4.2, I elaborate on the proposed Obj-SA-GAN model. In subsection 4.3, extensive experiments are conducted to evaluate the performance of the proposed model and perform an ablation study to evaluate the contributions of each key component of our method. Finally, the research work is concluded in subsection 4.4, and the directions for future research are explained.

4.2 Object-driven Self-Attention Generative Adversarial Network

The architecture of the proposed Obj-SA-GAN model is presented in Figure 4.2 It takes a text description as input and extracts text information of different granularity at different stages, from coarse to fine. High-quality semantic layers are formed gradually and used to guide the downstream image synthesis task. The semantic generator includes

two sub-generators: box and shape generator. The box generator parses the coarse features, e.g., location and class data of the entity objects and determines global layout of the generated images. The shape generator further refines the generated box sequence, outlining the general contour of each object. The image generator takes the text vector and the hidden feature map (hmap) generated by the semantic layer as inputs. In this stage, the semantic layer information is converted into pixels to form an image that conforms to the text semantics. This process is generally consistent with the original paper (W. Li et al., 2019a). However, the difference is that I introduce the self-attention mechanism in the semantic generation part, making the generated semantic layer more accurate and detailed.

4.2.1 Box generator

The box generator defines a mapping from a text vector (s) to a sequence of boxes, namely,

$$B_{(1:t)} = B_1, B_2, \cdots, B_t \sim G_{box}(s) \tag{4.1}$$

It defines what kind of objects should be included in the picture and where to place these objects. The t^{th} box annotation can be represented as $B_t = (b_t, l_t)$, where b_t refers to the coordinates of the top left corner of each object box (x, y) and the width and height of the box (w, h). l_t denotes the label information of the object. Figure 4.3 demonstrates the architecture of the box generator. Box generator is a seq2seq model based on the encoder-decoder architecture. A given sequence of text is first mapped to an embedding intermediate vector through an embedding layer, which generates an embedding vector for each text sequence. The embedding vector is then fed into a Self-Attention module. In the Self-Attention module, the model extracts key information for each object and generates a new set of vectors C_N, where N denotes the number of objects in the text. The Self-Attention module pays attention to each object and extracts the corresponding

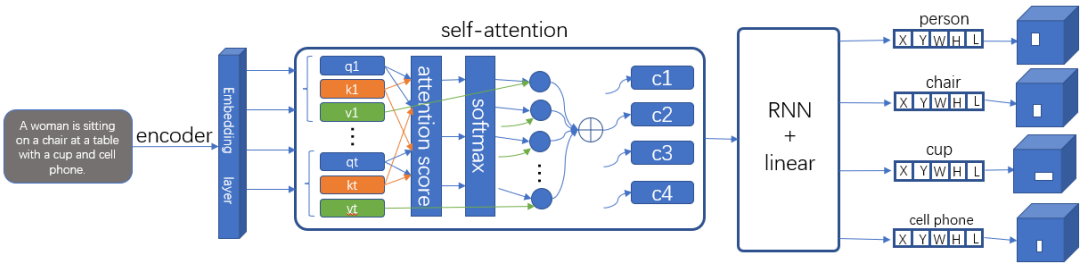

Figure 4.3: The architecture of the box generator.

core information. It also allows to generate more accurate box sequences for each object.

In order to train the box generator, I use Equation 4.2 as the loss.

$$L_{(box)} = -\lambda_l \frac{1}{T} \sum_{t=1}^{T} l_t^* logp(l_t) - \lambda_b \frac{1}{T} \sum_{t=1}^{T} logp(b_t^*) \tag{4.2}$$

In Equation 4.2, T indicates the number of objects in the text, l_t^* refers to the true label of the box, l_t indicates the predicted label, and b_t^* describes the true bounding box (x, y, w, h). L_{box} measures the error between the generated box coordinates and the actual box coordinates. The box generator only needs to detect the objects and the corresponding positions. It does not need to detect if the generated bounding box is consistent with the actual image. Thus, the predicted b_t is not involved in the Equation 4.2. The loss function considers both label loss and bounding box loss. The former describes a Negative Log Likelihood Loss (NLLLoss) to estimate the error related to the label, while the latter can be recognised as Squared Loss to estimate the error with the object box. In the current setting, I set λ_l= λ_b=1.0.

4.2.2 Shape generator

The shape generator is a further refinement of the box generator, which aims to predict the shape of an object in a given sequence of object box. Mathematically, $M_{(1:T)}$ =

$G_{shape}(B_{1:T}, Z_{1:T})$, where Z_T denotes a random noise vector. The shape generator is restricted by an instance constraint and a global constraint. The instance constraint ensures that the generated shape keeps consistent with the position of the previously generated box. The global constraint guarantees that generated shape fits the elements around it. The core component of the shape generator is a bidirectional convolutional LSTM (bi-convLSTM) model. The input is a feature map extracted from the box generator, followed by a bi-convLSTM block. I perform a self-attention operation before the forward and backward LSTMs. The hidden states in all steps are weighted to pay attention to the more important hidden state information in the entire text. This gives a better performance than using the bi-convLSTM alone.

A training strategy is employed to train the shape generator based on the GAN architecture. It consists of two components, instance-constrained discriminator (D_{inst}) and globally constrained discriminator (D_{global}), respectively. The loss function is formulated in Equation 4.3

$$l_{shape} = \lambda_i l_{inst} + \lambda_g l_{global} + \lambda_r l_{rec} \tag{4.3}$$

where l_{inst} and l_{global} denote the loss functions used by the two discriminators mentioned above. Both adopt the Binary Cross Entropy Loss (BCELoss) to measure the distance between the generated fake hamps and the real hmaps. l_{rec} refers to a perceptual loss, which measures the distance between the actual image and the generated image. In the current setting, I give λ_i=1.0 λ_g=1.0 and λ_r=10.0.

4.2.3 Image generator

The study conducted by Li et al. (W. Li et al., 2019a) focuses on the image generator part and makes great progress. More importantly, an attention mechanism for image generation is introduced, which is in line with our research. As shown in Figure 4.4, the

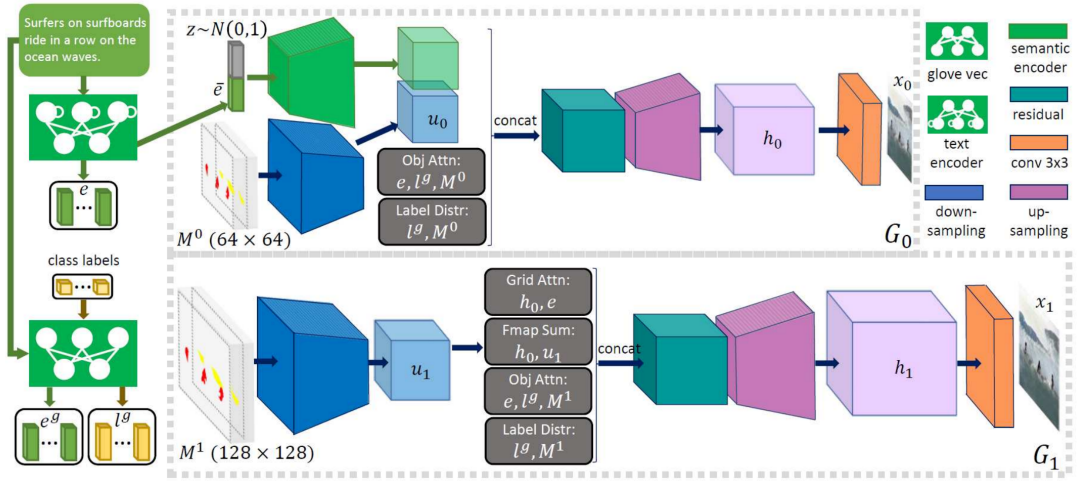

Figure 4.4: The Object-driven Self-Attention image generator (W. Li et al., 2019a).

image generator is presented as a multi-stage GAN, incorporating two generators, i.e., G0 and G1. Specifically, G0 concatenates features generated by text embedding and shape, extracts h0 features and generates a fake image x0. G1 magnifies the features synthesised by the shape generator and uses grid attention to extract h0 again. Finally, it connects the results with obj-attention to generate a new fake image x1.

4.3 Experiments

In this section, extensive experiments are performed to evaluate the proposed Obj-SA-GAN model by using the MSCOCO dataset. Firstly, a brief description of the datasets is given. Secondly, I compare the performance of the Obj-SA-GAN model with state-of-the-art generative models. Thirdly, I perform ablation experiments to compare the contribution of each module of the model.

4.3.1 Setup

Datasets

The Microsoft Common Objects in Context 2014 (MS COCO-2014) dataset (Lin et al., 2014) and the ImageNet dataset (Deng et al., 2009) are utilised in this research.

- MS COCO [1] was released in 2014. It is a collection of 164K images, which have been partitioned into the training set (82K), validation set (41K) and testing set (41K). The dataset is complex because most of the images possess at least two objects.

- ImageNet [2] was released in 2009. It consists of 14 million images, covering most of the categories of images seen in life. ImageNet has more than 20K classifications, and each image is manually categorised.

Evaluation metrics

I adopt IS and FID as evaluation metrics (Hong et al., 2018; W. Li et al., 2019a). Both are acknowledged as standard metrics for evaluating the GAN-based generation model. Specifically, IS examined both the clarity and diversity of the resulting images. The higher the IS, the better the quality of the generated images. FID calculates the difference between the generated image and the original image. The smaller the difference, the better the generated image is.

Baselines

The baselines utilised as the counterparts of the proposed model are as follows.

- Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis (Infer) (Hong

[1] https://paperswithcode.com/dataset/coco
[2] https://www.image-net.org/

et al., 2018) is a text-to-image synthesis model which integrates a semantic layer model with a generative model.

- Obj-GAN (W. Li et al., 2019a) is an improved model of Infer, which focuses on enhancing the image generator module of Infer. Object-driven attention is adopted in the GAN to synthesise images.

- StackGAN (H. Zhang et al., 2017) stacks two GAN models, which is a typical multi-stage generation model. The goal of stage 1 is to generate a rough sketch; stage 2 adds detail to the sketch to produce a realistic and high-resolution image.

- AttnGAN (Xu et al., 2018) is also a multi-stage generation model. It not only encodes the text description into a global embedding, but also performs word-level feature extraction. The text is fully mined from both global and local aspects.

4.3.2 Experimental Results

In this subsection, I evaluate the proposed model by comparing it against a few state-of-the-art generative models quantitatively and qualitatively.

Performance evaluation

Table 4.1 demonstrates the results of the quantitative comparison. It can be seen from the table that the proposed model outperforms all the baselines. In terms of FID, our Obj-SA-GAN model yields outstanding performance compared with the existing generative models. Regarding IS, the Obj-SA-GAN model also performs best, reaching approximately 32.26, almost twice the Infer baseline. According to the results, I can conclude that involving Self-Attention in the semantic layer can produce a significantly positive effect on the deep mining of relationships between objects because it fully utilises limited textual information.

Table 4.1: Experimental results of varied models for Text-To-Image synthesis. Symbols ↑ and ↓ indicate the higher the best and the lower the best, respectively. n/a means that the indicator is not used in the article. I utilise **bold** indicates the experimental results of our proposed model. * indicates the best performance. The value that follows ± is the standard deviation.

Models	**Inception ↑**	**FID ↓**
Obj-SA-GAN	**32.26 ± 0.02 ***	**18.20 ***
Obj-GAN (baseline) (W. Li et al., 2019a)	29.89 ± 0.22	21.21
Infer (baseline) (Hong et al., 2018)	12.40 ± 0.08	n/a
P-AttnGAN 0 (W. Li et al., 2019a)	18.84 ± 0.29	59.02
P-AttnGAN 1 (W. Li et al., 2019a)	19.32 ± 0.29	54.96
P-AttnGAN 2 (W. Li et al., 2019a)	20.81 ± 0.16	48.47
Reed et al. (Reed et al., 2016)	7.88 ± 0.07	n/a
StackGAN (H. Zhang et al., 2017)	8.45 ± 0.03	n/a
AttnGAN (Xu et al., 2018)	23.79 ± 0.32	28.76
vmGAN (S. Zhang et al., 2018)	9.94 ± 0.12	n/a

Text-input: A brown dog lying on bed with his banana toy

Two kids standing outside flying a kite during the day

Figure 4.5: Generation results produced by our proposed model. The four subplots in each sample correspond to different epochs, ranging from 60 to 100.

Table 4.2: Ablation study of Obj-SA-GAN model

Models	Box attention	Shape attention	Inception	FID
Obj-SA-GAN	YES	YES	**32.26**	**18.25**
Obj-SA-GAN1	YES	NO	31.41	19.21
Obj-SA-GAN2	NO	YES	32.54	19.87

Qualitative analysis

In this subsection, qualitative analysis is conducted to visually and intuitively compare the results of each generated model. Figure 4.5 demonstrates the images generated by our model at different epochs. The input text is given as "a brown dog lying on bed with his banana toy". In Figure 4.1, I have shown the actual image of the sample and the images generated by the four existing generative models. However, none of the generated images has any traces of a banana toy. In contrast, by applying our Obj-SA-GAN model, the shape of the banana becomes more apparent with the epoch increases. The result explicitly reveals that adding Self-Attention to the semantic layer can promote the model to generate an accurate and reasonable semantic layout, effectively guiding the image synthesis.

4.3.3 Ablation study

In this subsection, two ablation experiments are conducted to investigate the effectiveness of the Self-Attention module and shape generator, respectively. In Table 4.2, I statistically present the performance of the models by eliminating the Self-Attention module in box and shape generator, respectively. It can be seen from the table that the FID of Obj-SA-GAN$_1$ appears close to Obj-SA-GAN$_2$. This reveals that the box and shape generator almost contribute equally to FID. As for IS, Obj-SA-GAN$_2$ reaches 32.54, nearly equal to the proposed model. This is because IS does not consider the semantic layout when evaluating the model.

4.4 Summary

In this chapter, I proposed a novel text-to-image synthesis model, i.e., Obj-SA-GAN, incorporating self-attention and semantic layout. The proposed model adopts self-attention in the box and shape generator, which enhances text utilisation and deeply parses complex text descriptions, from coarse to fine. The model gradually forms an accurate and fine-grained semantic layout to guide the global layout of the generated images. The proposed Obj-SA-GAN model can achieve excellent performance on the MSCOCO dataset, outperforming most existing generative models.

The key results of this chapter have been published in The 19th Pacific Rim International Conference on Artificial Intelligence [3] (PRICAI 2022).

[3] https://pricai.org/2022/

Chapter 5

Swinv2-Imagen: Hierarchical Vision Transformer Diffusion Models for Text-to-Image Generation

5.1 Introduction

People tend to describe rich and detailed pictures of scenes through language, and the ability to generate images from these descriptions can facilitate creative applications in various life contexts, including art design and multimedia content creation (Kim, Joo & Kim, 2020; R. Li, Wang, Feng, Zhang & Wang, 2020). This fact has inspired researchers to design models of text to image comparative learning to assist people with making decisions quickly in specific scenarios, such as presentation and advertising design (Mathesul, Bhutkar & Rambhad, 2021; Park, Azadi, Liu, Darrell & Rohrbach, 2021). In recent years, diffusion models have attracted the attention of many scholars due to their promising performance in image generation. Within this framework, DALL-E 2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022) have become successful generative models for image generation.

Imagen is currently one of the greatest image generation models. Its most significant distinguishing feature is its immensity, which is reflected, in particular, by its utilisation of a large text encoder, i.e., T5 (Raffel et al., 2020). T5 is pre-trained on a sizable plain text corpus. It turns out that T5 is very effective for enhancing image fidelity and image-text alignment (Saharia et al., 2022). However, using T5 alone to obtain text embeddings cannot guarantee that the model learns important text features, such as semantic layout. Besides visual elements, the semantic layout is recognised as an important factor in guiding text-to-image synthesis (W. Li et al., 2019b). Our experimental results provide evidence for this claim.

Furthermore, very few research works are dedicated to addressing the UNet issue of Imagen. The diffusion model of Imagen relies on the Efficient-UNet, which suffers from the limitations of CNN convolution operations. CNN are good at extracting the low-level features and elements of visual structure, such as colour, contour, texture and shape (Ganar, Gode & Jambhulkar, 2014). However, CNN focuses on the consistency of these low-level features under transformations, such as translation (Kauderer-Abrams, 2017) and rotation (Chidester, Do & Ma, 2018). This is also the main reason why CNNs are widely used in object detection (Z.-Q. Zhao, Zheng, Xu & Wu, 2019). In other words, while the convolutional filters are good at detecting key points, object boundaries and other basic units that constitute the visual elements, it fails to extract features efficiently in terms of global and layout. For text-to-image synthesis tasks, it is significant to consider how to accurately extract the complex relationships between objects from the limited text. The Transformer is more natural and efficient than CNN in processing this demand. This is mainly because the attention in the Transformer can effectively mine the relationships between text features, allowing the model not only focuses on local information but also has a diffusion mechanism to find expressions from the local to global layout (J. Li, Yan, Liao, Yang & Shao, 2021; Liang et al., 2022).

To solve the aforementioned drawbacks of Imagen, in this paper, a diffusion text-to-image generation model is proposed, called Swinv2-Imagen. The proposed model is based on a Hierarchical Visual Transformer and Scene Graph incorporating layout information. Specifically, the semantic layout is generated via semantic scene graphs, enabling Swinv2-Imagen to parse the layout information in the text description effectively. In this paper, Stanford Scene Graph Parser (Johnson et al., 2018) is adopted to obtain the Scene Graph from the text. Subsequently, the entity and relationship embeddings are extracted using a frozen Graph Convolution Network (GCN) (Johnson et al., 2018). The image generation process appears conditional on text, object and relationship embeddings. The layout representation with global semantic information ensures the realism of the generated images. In addition, the diffusion models are developed based on Swinv2-Unet, a variant of Swin Transformer v2 (Liu et al., 2022), which allows the model to learn features from local to global. Finally, the model is evaluated on the MSCOCO, CUB and MM-CelebA-HQ datasets. The results show that the proposed model outperforms the current best generative model, Imagen, on MSCOCO. The ablation experiments reveal that the addition of semantic layouts is effective in improving the semantic understanding of the model.

The key contributions of this chapter are summarised below.

1. I leverage scene graphs to extract entity and relational embeddings to improve local and layout information representation of text for a more accurate understanding of the text and realistic image generation;

2. I propose Swinv2-UNet as a novel diffusion model architecture. The model leverages attention to explore the relationship between features, allowing the diffusion model to focus on different granularities of features at different moments, from local to global;

3. I combine the scene graph with the Transformer to improve the effectiveness of

the model;

4. I achieve a new state-of-the-art FID result, (FID=7.21), on the MSCOCO dataset compared to the latest generative models. Better results are also obtained on both the CUB (FID=9.78) and MM CelebA-HQ (FID=10.31).

The rest of the chapter is organised as follows. In subsection 5.2, I elaborate on the proposed Swinv2-Imagen model. In subsection 5.3, I conduct extensive experiments to evaluate the performance of the proposed model and perform an ablation study to evaluate the contributions of each key component of our model. Finally, this chapter is concluded in subsection 5.4, and future research directions are discussed.

5.2 Swinv2-Imagen

The overall architecture of the proposed Swinv2-Imagen model is shown in Figure 5.1. It takes text descriptions as input and uses scene graphs to guide downstream image generation more accurately and efficiently. The upstream comprises two sub-modules: the text encoder, which maps the text input to a text embedding sequence and the scene graph generator sub-module. The scene graph generator includes a Scene Graph parser and a frozen Graph Neural Network, which aims to represent objects and relationships in a text with a graph structure. The downstream consists of a set of conditional diffusion models, integrating the intermediate embeddings in the upstream and generating high-fidelity images step by step.

The input of the model is a text-picture pair. Firstly, the text is encoded by T5 tokenizers and input to the embedding layer to get the initial text embedding. Next, it goes through the T5 encoder (n-layer T5 Block) to obtain Text Embeddings. Meanwhile, the scene graph parser extracts the scene graph from the text, and the frozen GCN (m-layer Graph Triple Convolution) obtains the corresponding Object and Relation

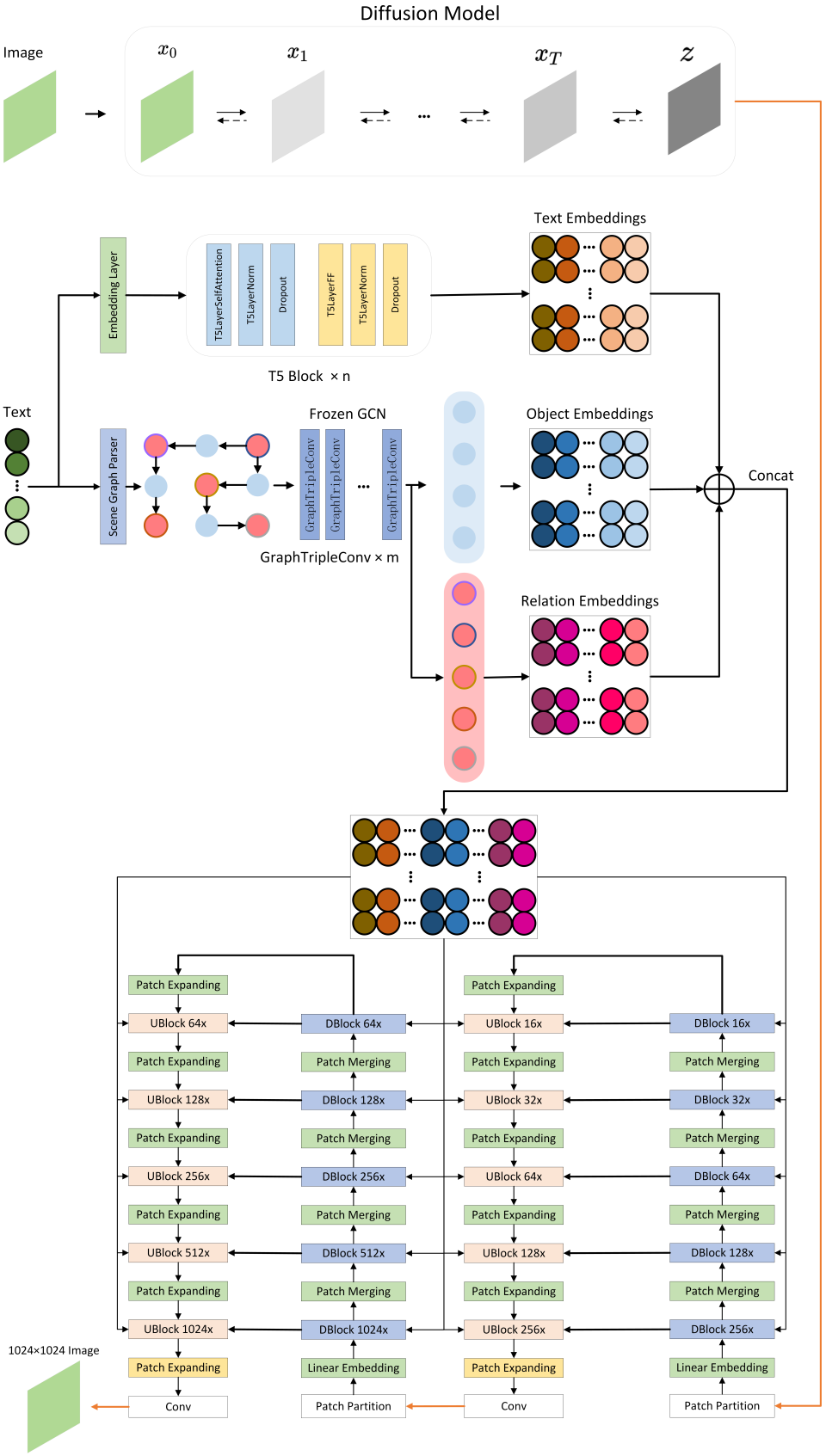

Figure 5.1: The overall architecture of Swinv2-Imagen.

embeddings. Finally, the Conditional embeddings are obtained by concatenating the Text embeddings, Object embeddings and Relation embeddings in this order. The Conditional embeddings are used as conditional input for subsequent super-resolution image generation. In the following subsections, the main components of Swinv2-Imagen are described in detail.

5.2.1 Pre-trained frozen text encoders

It is widely acknowledged that a robust semantic text encoder is essential for text-to-image synthesis models and plays a crucial role in analysing the complexity and composition of textual input (Saharia et al., 2022). Previously, language models were mainly built on RNN architectures. However, since the emergence of the Transformer, a number of transformer-based pre-trained language models have been developed, such as GPT (Radford & Narasimhan, 2018; Radford et al., 2019; Brock, Donahue & Simonyan, 2019), BERT (Devlin, Chang, Lee & Toutanova, 2019) and T5 (Raffel et al., 2020). The traditional Imagen model is compared against popular text encoders, BERT, CLIP and T5-XXX, by freezing parameters. The existing research results prove the promising performance of T5-XXX in terms of both image-text alignment and image fidelity (Saharia et al., 2022). Therefore, the T5 large language model is adopted for text encoding in the proposed model.

5.2.2 Scene Graph and Frozen Graph Convolutional Neural Network

This sub-module aims to extract entity and relationship features from the text to enhance the text understanding of the model. a Scene Graph parser is adopted to represent text as a scene graph, followed by a frozen GCN to extract the entity and relational embeddings for the image generation with diffusion models. Scene graphs with graph

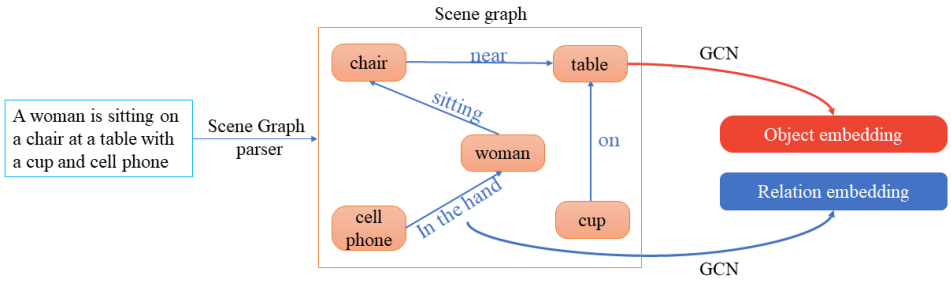

Figure 5.2: The process of Object and Relation embeddings extraction.

neural networks (Johnson et al., 2018) have been proven to be highly effective in extracting object relationships from the text. As shown in Figure 5.2, Swinv2-Imagen constructs a scene graph for the text and is followed by a graph neural network to extract the entities and relationships from the scene graph. For any given text description, the corresponding scene graph is represented as (O, E), where $O = (o_1, o_2, o_3, \cdots, o_n)$ denotes each object in the sentence i.e., subject and object, and E is a collection of edges of the form (o_i, r, o_j), where $r \in \mathcal{R}$, $\mathcal{R}$ refers to a collection of relationships, such as position and action. In the end, object and relation embeddings are constructed, which are used to assist the T5 model in analysing and understanding the text more comprehensively.

The input to the graph convolution is a scene graph, having each node and edge represented as a vector with dimension D_{in}, i.e., $\mathbf{v}_i, \mathbf{v}_r \in \mathbb{R}^{D_{in}}$. In the graph convolution sub-module, these vectors are adopted to compute output vectors with dimension D_{out} for each node and edge, i.e., $\mathbf{v}_i', \mathbf{v}_r' \in \mathbb{R}^{D_{out}}$. Three functions, g_s, g_o and g_p are used to calculate the object features vectors and relation vectors of output. They take a triplet as input, i.e., $(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j)$. In the scene graph, given an edge $\mathbf{v}_r$, the two associated objects, $\mathbf{v}_i$ and $\mathbf{v}_j$, are determined. Thus, the output relationship vector $\mathbf{v}_r'$ can be simply expressed as:

$$\mathbf{v}_r' = g_p\left(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j\right) \tag{5.1}$$

In contrast, the calculation of output object vectors $\mathbf{v}_i'$, is more complicated. Generally, an object is associated with two or more relations. Therefore, the output vector of an entity o_i is calculated by considering all the vectors directly connected to the object, i.e., $\mathbf{v}_j$, and the corresponding relationship vectors, $\mathbf{v}_r$. The function g_s in Equation 5.2 is used to compute all vectors starting at node o_i and function g_o in Equation 5.3 is used to compute all vectors ending at node o_i. Afterwards, these vectors are collected into lists V_i^s and V_i^o.

$$V_i^s = \{g_s(\mathbf{v}_i, \mathbf{v}_r, \mathbf{v}_j) : (o_i, r, o_j) \in E\} \tag{5.2}$$

$$V_i^o = \{g_o(\mathbf{v}_j, \mathbf{v}_r, \mathbf{v}_i) : (o_j, r, o_i) \in E\} \tag{5.3}$$

Then the output vector $\mathbf{v}_\mathbf{i}^{'}$ for the entity o_i is expressed as.

$$\mathbf{v}_i' = h(V_i^s \cup V_i^o) \quad , \tag{5.4}$$

where h denotes a function that pools all vectors in lists V_i^s and V_i^o to a single output vector (Johnson et al., 2018).

5.2.3 Image generator

The image generator is composed of three diffusion models located downstream. In the diffusion model, a hidden variable z is obtained by adding noise to the image for T times. After forward and backward diffusion, a basic 64 * 64 image can be learned. The basic image is input to the first Swinv2-Unet to generate a 256 * 256 image. Finally, the image goes to the second Swinv2-Unet super-resolution generation, producing a 1024 * 1024 high-definition image. In contrast to Imagen, I focus on improving super-resolution diffusion models. A new Unet variant is introduced to our super-resolution

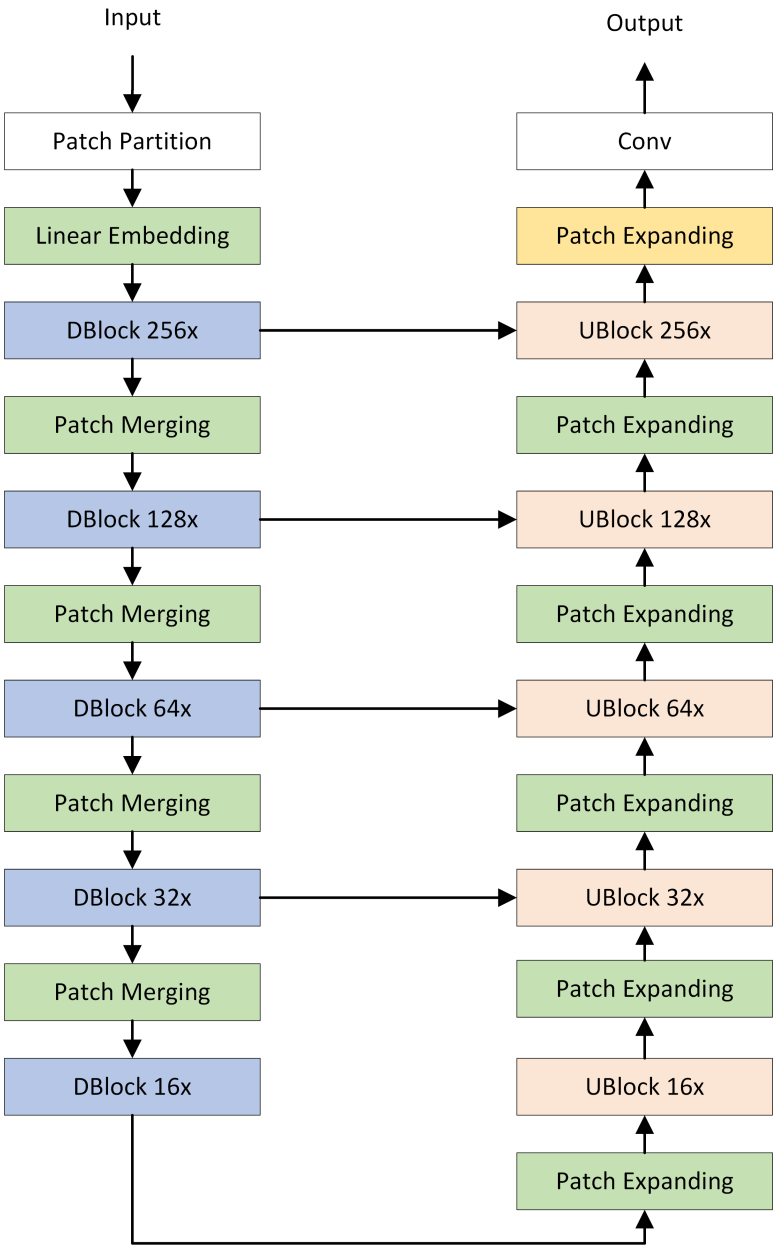

Figure 5.3: The Unet architecture of the super-resolution submodule.

diffusion model, called Swinv2-Unet. The Swin Transformer Block is replaced with the Swin Transformer v2 Block based on the original Swin-Unet (H. Cao et al., 2021), the complete structure of which is shown in Figure 5.3.

A distinctive feature of Swinv2-Unet compared to Swin-Unet is the replacement of the $dot(\mathbf{K}, \mathbf{Q})$ operation with cosine normalisation (C. Luo, Zhan, Wang & Yang, 2018) in the attention part, which makes the attention output more stable. Given two vectors, **Q** and **K**, the cosine normalisation could be expressed as:

$$Cosine(\mathbf{Q}, \mathbf{K}) = \frac{\sum_i (q^i k^i)}{\sqrt{\sum_i (q^i)^2}\sqrt{\sum_i (k^i)^2}}. \tag{5.5}$$

The DBlock and UBlock of Swinv2-UNet consist of the Swin Transformer v2 block, which comprises LayerNorm (LN) layers, multi-headed self-attention modules, Residual connections and a 2-layer MLP with Gaussian Error Linear Unit (GELU) non-linearity. The Swin Transformer v2 block could be represented as:

$$\hat{z}^{l+1} = LN(Attn(z^l)) + z^l \tag{5.6}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad , \tag{5.7}$$

where z^l and z^{l+1} denote the input and output of the Transformer v2 block, respectively. $\hat{z}^{l+1}$ is an intermediate variable. + denotes the residual connection or skip connection.

The attention of Swin-v2 is expressed as:

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = SoftMax\left(\frac{Cosine(\mathbf{Q}, \mathbf{K})}{\tau} + B\right) \quad , \tag{5.8}$$

where **Q**, **K**, and **V** denote the matrix of query, key and value, respectively. $Cosine()$ refers to a function that calculates the scaled cosine similarity of **Q** and **K**. τ denotes a learnable scalar, usually greater than 0.01. B is a matrix of relative position bias.

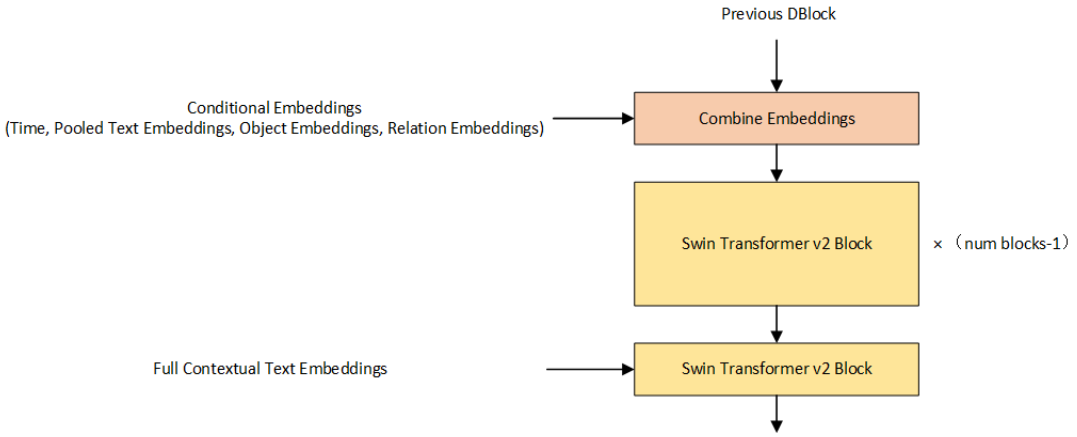

Figure 5.4: Swinv2-Unet DBlock

Figure 5.4 illustrates the network structure of the Swinv2-Unet DBlock, which is the basic component of the downsampling path under the encoding-decoding structure of Unet. Firstly, the DBlock combines the pooled text embeddings, object embeddings and relation embeddings into a conditional embedding input to the cross-attention layer. Next, it is followed by the Swinv2-Transformer v2 blocks for (num_block-1) times feature extraction.

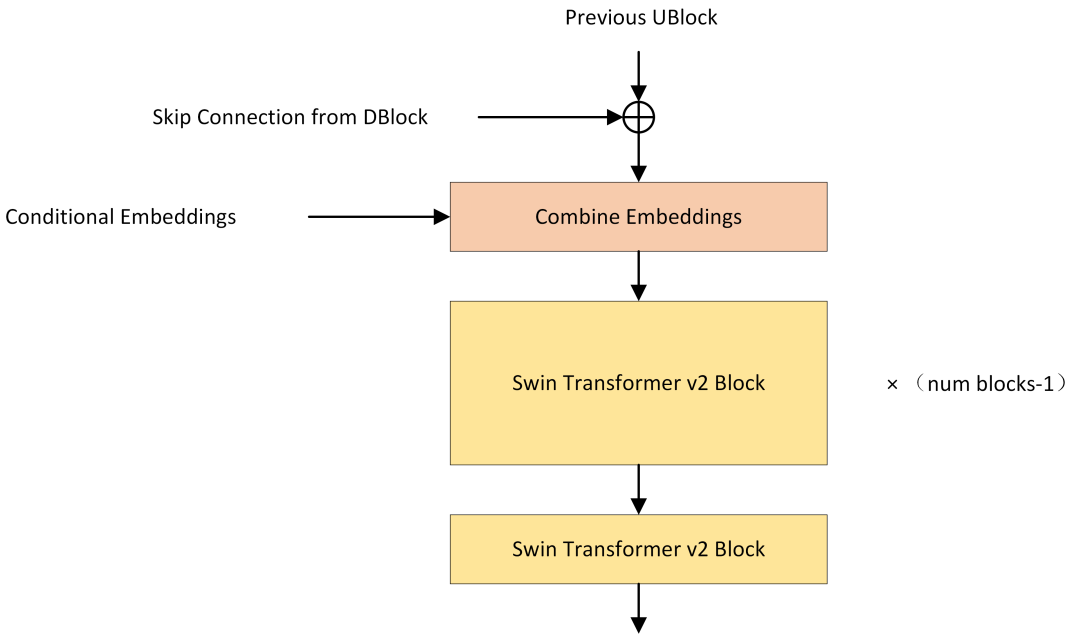

Figure 5.5: Swinv2-Unet UBlcok

Figure 5.5 shows the network structure of the Swinv2-Unet UBlock, which is the basic component of the upsampling path on the Unet encoder-decoder. The inputs to the UBlock include the output of the previous UBlock layer and the corresponding DBlock. The DBlock and UBlock are connected using skip connections (Z. Zhou et al., 2020). Subsequently, the conditional embedding inputs are also introduced to the cross-attention layer. Similar to the DBlock, this layer is followed by the Swinv2-Transformer v2 blocks for (num_block-1) times feature extraction.

The encoder is presented as a stacking of DBlocks and Patch Merging. In the encoder, images are fed into five consecutive DBlocks for learning, where the feature dimension and resolution are maintained. Meanwhile, Patch Merging performs Token Merging and increases the feature dimension to four times the original dimension. Next, a linear layer is applied to standardise the feature dimension to twice the original dimension. The process is repeated four times in the encoder.

Similar to Unet, skip connections are used to integrate the multi-scale features of the encoder with the upsampled features. The model connects shallow and deep features to minimise the loss of spatial information due to downsampling. The next layer is a linear layer where the dimensionality of the connected features is kept the same dimensionality as that of the upsampled features.

The decoder is a symmetric decoder corresponding to the encoder. For this reason, unlike the Patch Merging used in the encoder, Patch Expanding is used in the decoder to upsample the extracted features. The Patch Expanding reshapes the feature maps of adjacent dimensions into a higher resolution feature map ($2 \times$ upsampling) and accordingly reduces the number of feature dimensions to half the original dimensionality.

5.3 Experiments

In this section, I perform extensive experiments to evaluate the proposed Swinv2-Imagen model by using the MSCOCO, CUB and Multi-modalCelebA-HQ (MM CelebA-HQ) datasets. Firstly, a brief description of the datasets is given. Secondly, I compare the performance of the Swinv2-Imagen model with state-of-the-art generative models. Finally, I conduct ablation experiments to compare the contributions of each module.

5.3.1 Setup

Datasets

The Microsoft Common Objects in Context 2014 (MS COCO-2014) (Cho et al., 2014), the Caltech-UCSD Birds-200-2011 (CUB-200-2011) (Ho, 2022) and MM CelebA-HQ (Xia et al., 2021) datasets are utilised in this research. Three datasets cover both simple (CUB) and complex (MSCOCO) datasets. The use of the MM CelebA-HQ dataset is mainly because most generative models such as Cogview and Craiyon, produce distorted and less realistic faces.

- MSCOCO [1] was released in 2014. It is a collection of 164K images, which have been partitioned into the training set (82K), validation set (41K) and testing set (41K). The dataset is complex because most of the images possess at least two objects.

- CUB [2] contains 12K bird images of 200 subcategories, 6K for training and 6K for testing. It is a simple dataset, having only one object per image.

- MM CelebA-HQ [3] is a large-scale face image dataset. It is a collection of 30K

[1] https://cocodataset.org/
[2] https://deepai.org/dataset/cub-200-2011
[3] https://github.com/weihaox/Multi-Modal-CelebA-HQ-Dataset

high-resolution face images. The dataset is used widely to train and evaluate algorithms for text-image generation and text-guided image manipulation.

Evaluation metrics

I adopt FID (Heusel, Ramsauer, Unterthiner, Nessler & Hochreiter, 2017) and IS (W. Li et al., 2019a) as evaluation metrics. Both are acknowledged as standard metrics for evaluating the image generation model. Specifically, IS examines both the clarity and diversity of the resulting images. The higher the IS, the better the quality of the generated images. FID calculates the difference between the generated image and the original image. The smaller the difference, the better the generated image is.

Baselines

- PCCM-GAN (Qi, Sun, Qian, Xu & Zhan, 2021) (Photographic Text-to-Image Generation with Pyramid Contrastive Consistency Model) is a typical multi-stage generative model. Its main innovations include the introduction of stack attention and the lateral connection of the PCCM. The two modules enhance the generative model to simultaneously extract semantic information from both global and local aspects, ensuring that the generated images are semantically consistent.

- DM-GAN (M. Zhu et al., 2019) (Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis) is also a multi-stage generative model. It uses a memory module and a gate mechanism in the image refinement process. The aim is to re-extract important information from the image as an aid when the generated image is not as good as expected.

- SDGAN (H. Zhang, Koh, Baldridge, Lee & Yang, 2021) (Semantics Disentangling for Text-to-Image Generation) consists of two modules, i.e., Siamese

and semantic conditioned batch normalization, to extract high-level and low-level semantic features respectively.

- CogView (Ding et al., 2021) is based on the Transformer architecture. Its input is a text-image pair. The text and image features are combined and passed to the GPT language model for autoregressive training.

- GLIDE (Nichol et al., 2022) is a large-scale image generation model based on diffusion models with 3.5 billion model parameters.

- DALL-E 2 (Ramesh et al., 2022) is also based on diffusion models. One of its highlights is the use of a priori model built on the diffusion models. Its inputs are also text and corresponding images. The text is first passed through the priori model and a corresponding image vector is generated. The image is passed through the CLIP module which also generates an image vector to supervise the result of the priori model.

- LAFITE (Y. Zhou et al., 2022) is a variant of generative adversarial networks. It leverages the CLIP model to extract features from images and text, ensuring text-image consistency.

- Imagen (Saharia et al., 2022) is a text-to-image synthesising model based on the diffusion model. It passes text through a large pre-trained T5 language model and generates high-fidelity images through cascading diffusion model blocks.

Training parameters

I apply an Imagen-like training strategy, i.e., training the base model and then the super-resolution model twice. The Adam optimiser is adopted, having a learning rate of 1e-4. 10,000 linear warm-up steps are given with a batch size of 8 and training epochs of 1,000. The loss function is Mean Squared Error (MSE), formulated as follows.

$$MSE(I,K) = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - K(i,j)]^2 \quad , \tag{5.9}$$

where M and N denote the total number of pixels in the real image I and the generated image K, respectively. A smaller MSE implies that the generated image is closer to the real image.

5.3.2 Experimental Results

In this subsection, I evaluate the proposed model by comparing it against a few state-of-the-art generative models.

Performance evaluation

Model	MSCOCO		CUB		MM CelebA-HQ
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
PCCM-GAN (Qi et al., 2021)	33.59	26.52	22.15	4.65	-
DM-GAN (M. Zhu et al., 2019)	32.64	30.49	16.09	4.75	131.05
SDGAN (H. Zhang et al., 2021)	29.35	35.69	-	4.64	-
DALL-E (H. Zhang et al., 2021)	27.5	17.9	56.1	-	12.54
CogView (Ding et al., 2021)	13.9	18.2	-	-	-
GLIDE (Nichol et al., 2022)	12.24	-	-	-	-
DALL-E 2 (Ramesh et al., 2022)	10.39	-	-	-	-
LAFITE (Y. Zhou et al., 2022)	8.12	32.34	10.48	5.97	12.54
Make-A-Scene (Gafni et al., 2022)	7.55	-	-	-	-
Imagen (Saharia et al., 2022)	7.27	-	-	-	-
Swinv2-Imagen	**7.21**	**31.46**	**9.78**	**8.44**	**10.31**

Table 5.1: Experimental results of varied models for Text-To-Image synthesis. Symbols ↑ and ↓ indicate the higher the best and the lower the best, respectively. – means that the indicator is not used in the article.

Table 5.1 demonstrates the results of the quantitative comparison. The proposed model is compared against 10 popular generative models, including GAN and diffusion models. It is evident that the proposed Swinv2-Imagen model outperforms the baselines

on all three datasets. Particularly, on the MSCOCO dataset, Swinv2-Imagen significantly outperforms the GAN-based generative model and slightly surpasses the Imagen, achieving an FID of 7.21.

Qualitative analysis

Figures 5.6, 5.7 and 5.8 show examples of images generated by our proposed model on MSCOCO, CUB and MM CelebA-HQ, respectively. It can be seen that our model understands the text very well. For example, given the text input, 'Food cooks in a pot on a stove in a kitchen', the resulting picture not only contains the food, the stove and the pot, but also places these objects to the exact location. More importantly, based on the word 'kitchen', the model also generates other common kitchen objects, such as spoons and storage shelves. This shows that our model understands the text accurately and comprehensively. More images generated by Swinv2-Imagen are shown in Appendix B.

Figure 5.9 illustrates the qualitative comparison of the proposed model and the GAN-based, diffusion-based generative models, i.e., DM-GAN (M. Zhu et al., 2019), DF-GAN (Tao et al., 2022), VQ-Diffusion (S. Gu et al., 2022). Compared to the diffusion-based models, the GAN-based models lost many detailed features in the generated results. For example, the bird's eyes are very blurred in the third image in the first row and the second image in the second row. Compared to VQ-Diffusion, which is a diffusion-based model, our results are more realistic and contain more fine-grained features. Particularly, the blue birds in the third column, obviously, our result is better than VQ-Diffusion. In addition, our model also outperforms other generation models in terms of text-image alignment. The text description of the first column requires the bird's breast to be white, but this feature seems to be gray in the results of other models, especially DM-GAN. In summary, by comparing with other GAN-based and diffusion-based generation models, it can be seen that our model synthesises fine-grained and detailed images on CUB.

Figure 5.10 presents the qualitative comparison between our model and LAFITE (Y. Zhou et al., 2022) on MSCOCO. Intuitively, our results are more colourful and saturated. For example, in the first and fourth columns, our bus and city street include more colours and the images are brighter. Furthermore, our model is also better for text understanding. In the third column, the room should include two colours, white and beige, however, in the LAFITE result, there are just white walls and a white cupboard. There is not any trace of the beige features. In contrast, our generated room contains the two colours required by the text, and the overall layout is more realistic. Finally, our model is also better regarding image quality. The tops of the bus and room generated by LAFITE are distorted and the results are generally blurred. Our model has a significant advantage over LAFITE in generating objects such as buildings, buses, trees, etc. Although the two models are very close in terms of FID and IS in the quantitative analysis in Table 5.1, our model is superior in terms of the quality of the generated images.

5.3.3 Ablation study

Model	Scene Graph	Swinv2-Unet	FID
Imagen			7.27
Imagen_sg	YES		7.24
Swinv2-Imagen_su		YES	7.23
Swinv2-Imagen	YES	YES	7.21

Table 5.2: Ablation study of Swinv2-Imagen model

In order to improve the performance of the generation models, I introduce two new modules to Imagen, i.e., scene graph and Swinv2-Unet. These are the main innovations of the article. In this subsection, two ablation experiments are conducted on MSCOCO to investigate the contributions of the scene graph module and Swinv2-Unet, respectively. The choice to experiment on MSCOCO is based on two considerations.

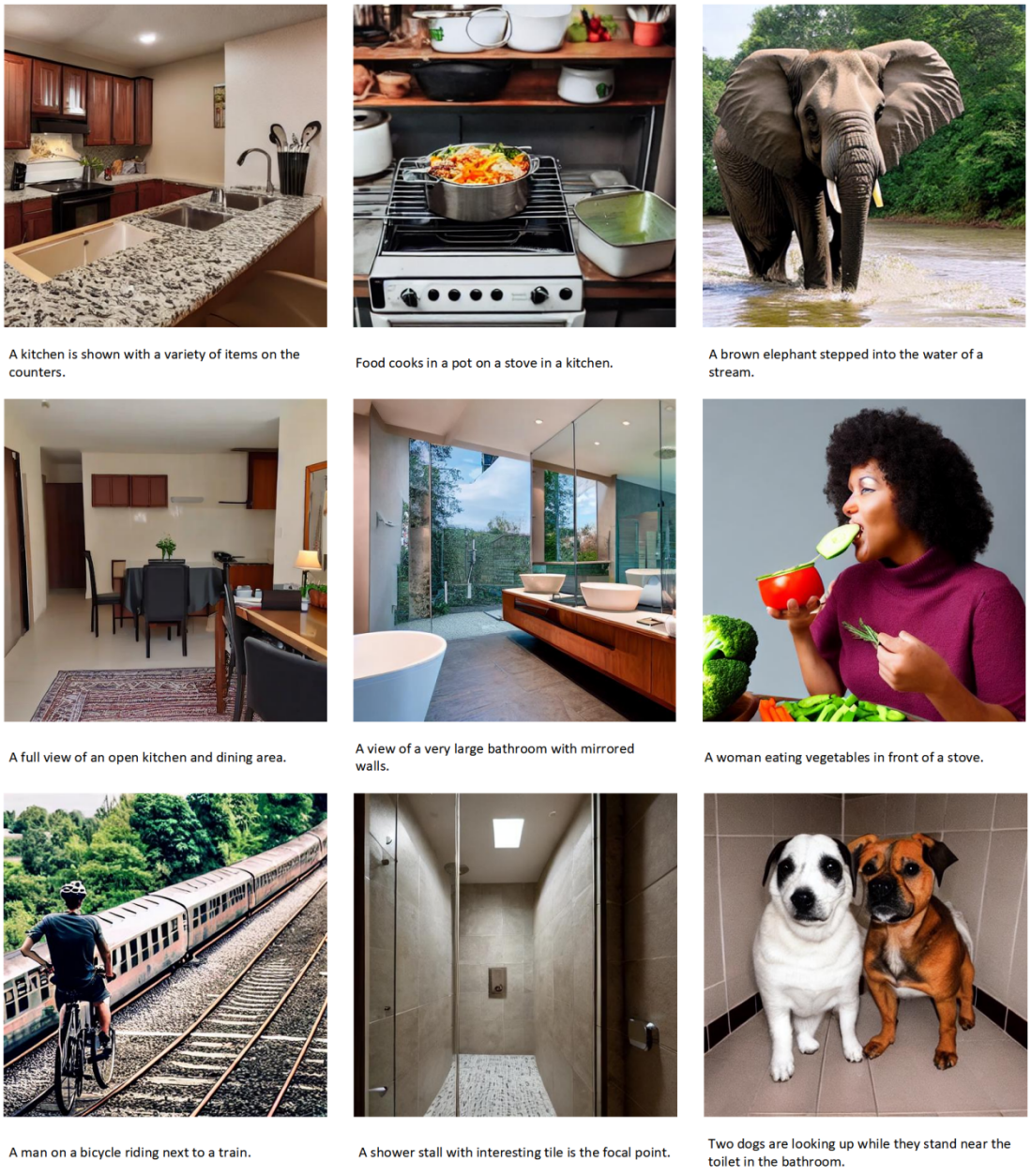

Figure 5.6: Generated examples on MSCOCO

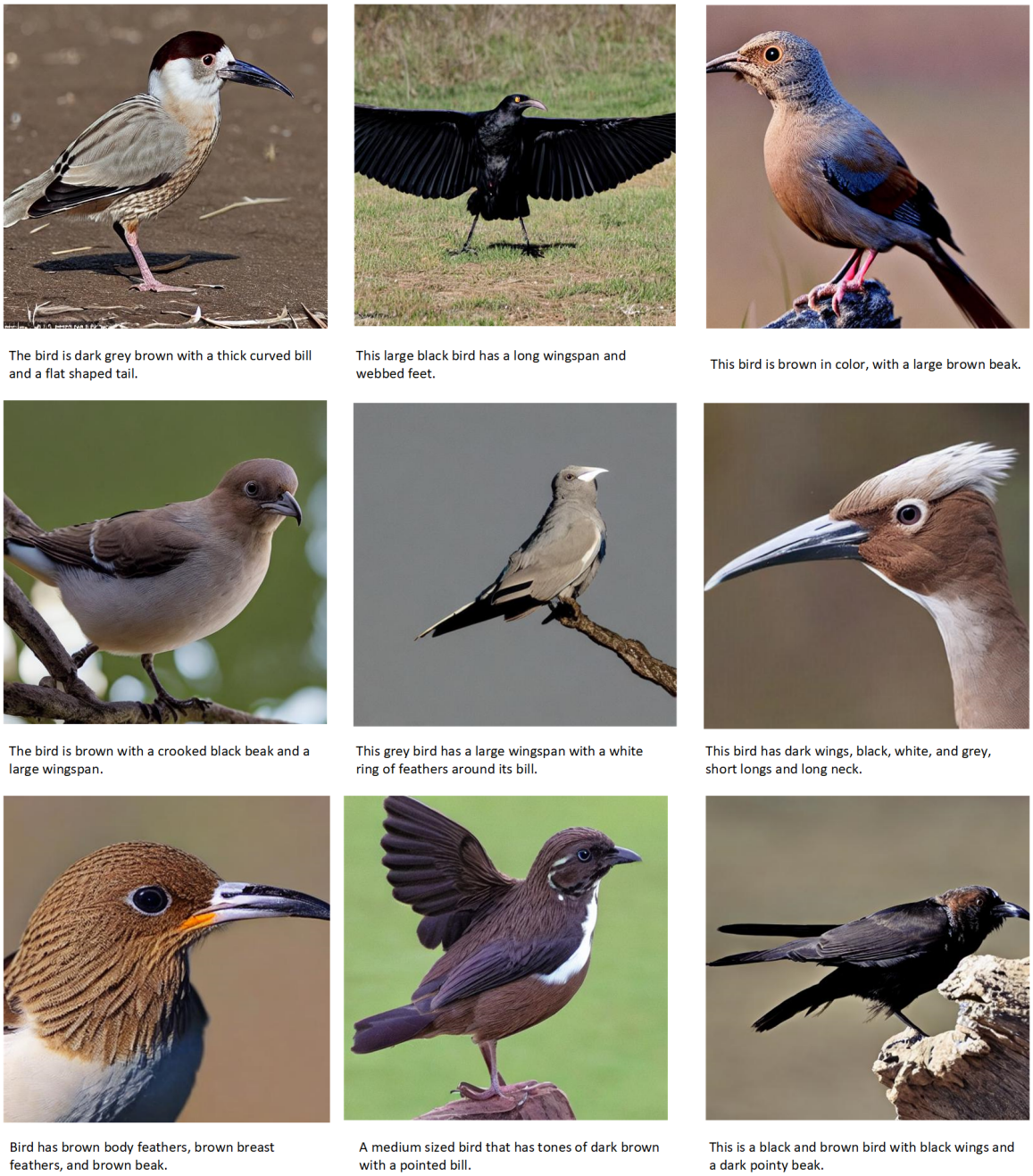

Figure 5.7: Generated examples on CUB

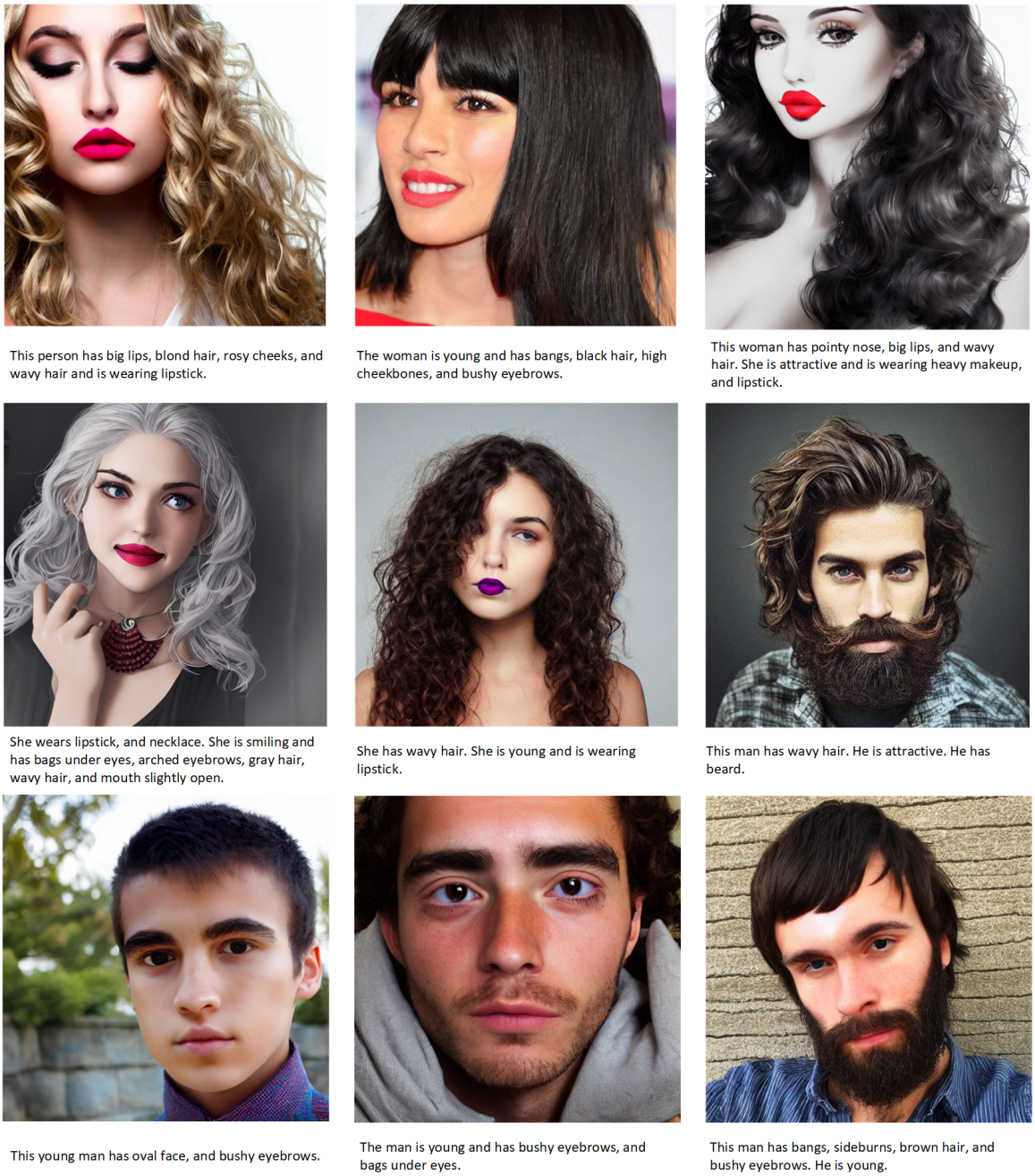

Figure 5.8: Generated examples on MM CelebA-HQ

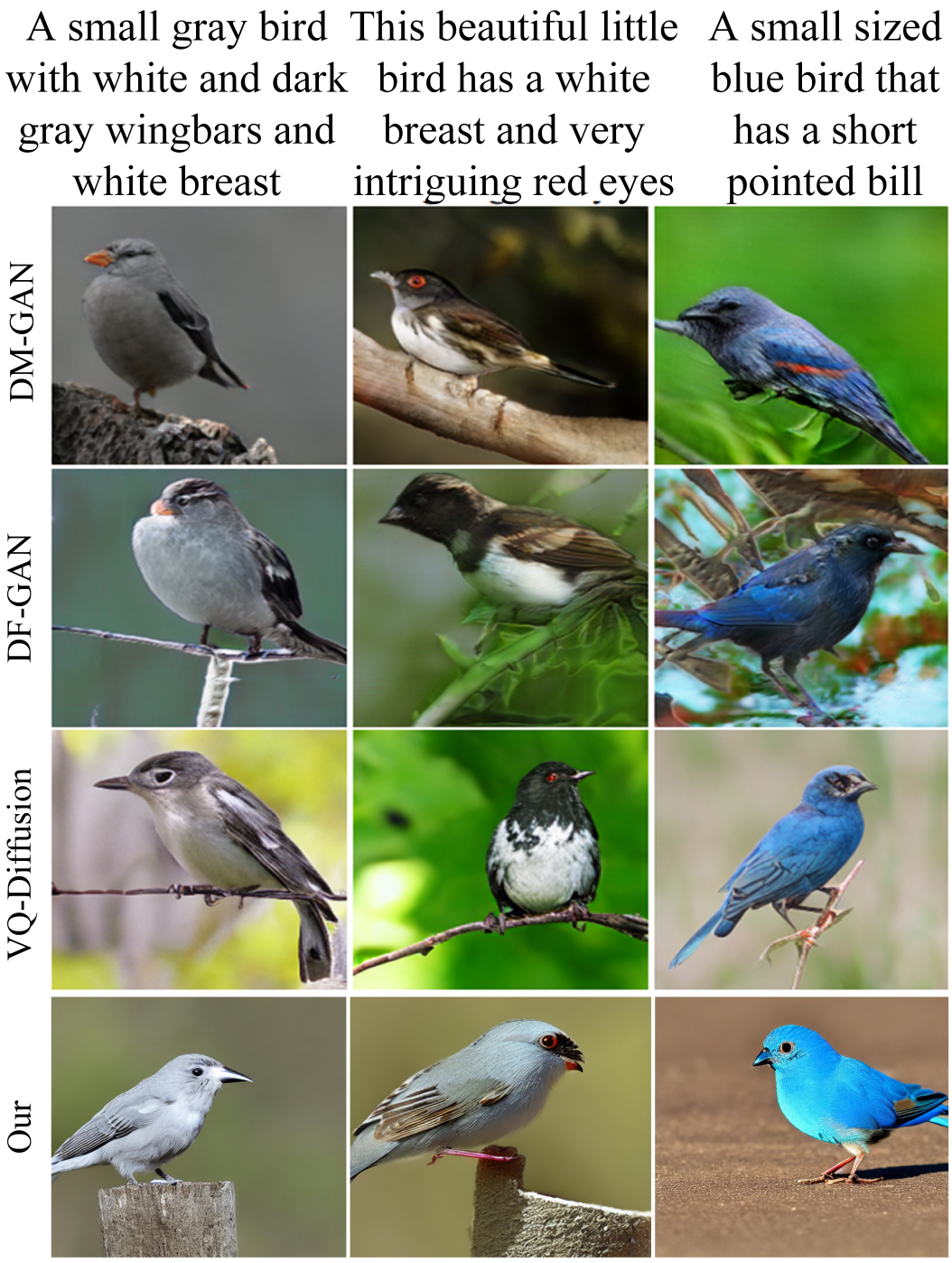

Figure 5.9: Comparison with GAN-based and diffusion models on CUB-200 dataset

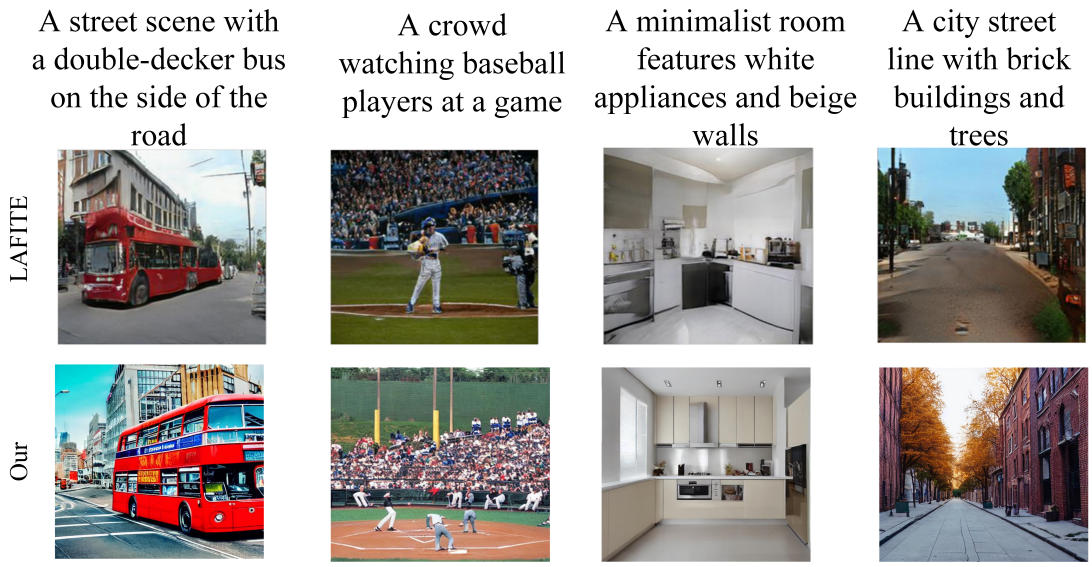

Figure 5.10: Comparison with LAFITE on MSCOCO dataset

Firstly, each image in MSCOCO contains multiple objects, which is more complex than CUB dataset. Theoretically, it allows a better evaluation on the effect of each module. Secondly, the main baseline I referenced, Imagen, is only experimented on MSCOCO. The aim of Experiment 1 is to evaluate the contribution of the scene graph module. I add only the scene graph, and the diffusion model is still built using Efficient-Unet, which is called Imagen_sg. Experiment 2 is designed to evaluate the performance of the Swinv2-Unet. I constructed a new diffusion model using our improved Swinv2-Unet and replace Imagen's super-resolution diffusion models with it, which is called Swinv2-Imagen_su. The result of Experiment 1 supports our conjecture that merely using a T5 encoder does not sufficiently learn the semantic information of the text, as mentioned in the introduction. Experiment 2 shows that the diffusion model constructed with the Transformer outperforms the CNN-constructed diffusion model in the image generation task. It also can be seen from Table 5.2 that the FIDs of the Imagen_sg and Swinv2-Image_su are very close. This intuitively reveals that the two submodules almost contribute equally to the FID.

5.4 Summary

In this chapter, I propose a novel text-to-image synthesis model based on Imagen, i.e., Swinv2-Imagen, which integrates the Transformer and Scene Graph. The improved sliding window-based hierarchical visual Transformer (Swin Transformer v2) avoids the local view of CNN convolution operations. It improves the efficiency and effectiveness of the Transformer applied to image generation. In addition, I introduce a Scene Graph in the text processing stage. Feature vectors of entities and relationships are extracted from the Scene Graph and incorporated into the diffusion model. These additional feature vectors improve the quality of generated images. Swinv2-Imagen produces 1024 × 1024 samples with unprecedented fidelity with these novel components.

The key results of this chapter have been published as a pre-print [4] and submitted to Neural Computing and Applications journal [5] (NCAA) for review.

[4] https://arxiv.org/abs/2210.09549?context=cs
[5] https://www.springer.com/journal/521

Chapter 6

Conclusion and Future works

6.1 Introduction

In this thesis, I propose and developed two novel text-to-image generation models, Obj-SA-GNA and Swinv2-Imagen, based on GANs and Diffusion models, respectively. I argue that semantic information in the text is a crucial factor for the task of text-to-image synthesis throughout this article, especially when the text description is complex. In this chapter, the main findings based on these two models are listed. The discussions about the limitations and future research directions are presented.

6.2 Research Contributions

The main research contributions of this thesis to the field of text-to-image generation are summarised in the following two aspects.

6.2.1 Obj-SA-GAN

- I proposed a non-end-to-end generation model, i.e., Obj-SA-GAN, which extracts semantic information with different granularity at different stages using the self-attention mechanism. The model fully analyses the text semantically from coarse to fine. It first generates a fine-grained semantic layout in the semantic generation stage. Then, the image generation is based on the semantic layout.

- The proposed Obj-SA-GAN model yields excellent performance on a complex dataset, e.g., MSCOCO, outperforming the current popular GANs-based generation models. The model addresses the performance issues of GANs-based models when being utilised in complex scenes.

6.2.2 Swinv2-Imagen

- I proposed a novel Unet architecture, i.e., Swinv2-Unet, which is based on the Transformer. This model replaces the CNN convolutional block with a Transformer block, which is more efficient and better at extracting global and layout features.

- I proposed a novel diffusion-based text-to-image generation model, i.e., Swinv2-Imagen, which parses the semantic information in the text into a scene graph to represent complex object relationships visually.

- I combined the scene graphs with the diffusion models. To the best of my knowledge, this is the first research combining a scene graph with a diffusion model in the text-to-image generation field.

- I achieved a new state-of-the-art FID result, 7.21, on the MSCOCO dataset compared to the latest generation models. Better results are also obtained on both the CUB (FID=9.78) and MM CelebA-HQ (FID=10.31).

6.3 Limitation and Future Research Directions

In this thesis, I proposed two novel text-to-image generation models, i.e., Obj-SA-GAN and Swinv2-Imagen. While these models have been shown to perform very well on both simple and complex data through both quantitative and qualitative experiments, they still have limitations that could be significantly improved in subsequent studies. In this subsection, I will explain the limitations of each model and present possible research directions.

Firstly, Obj-SA-GAN is a non-end-to-end generation model. The downstream image generation task is excessively dependent on the layout generated by the upstream semantic layer. The objective of each stage is inconsistent, and errors in the previous stage are accumulated in the later stages, which makes it difficult for the model to perform optimally at the end. Based on this problem, future research work could be to model an end-to-end generation model based on the Transformer, where the text semantic information is extracted by the network layers close to the input and the other network layers perform the image generation. This idea may need more data to support it.

Secondly, in the quantitative evaluation phase of the generation models, the Inception Score is completely useless for evaluating the semantic layout of the generated images and does not give a reasonable indication of the performance of our model. Based on the issue, future work is planned to design a new quantitative assessment metric that focuses on evaluating the model semantically and complements the Inception score and other metrics.

Thirdly, Swinv2-Imagen is a diffusion model, which is exceptionally slow in both the training and inference phases. This provides an excellent direction for our future research. In the future, I will investigate how diffusion models are able to make inferences as fast as GANs.

Finally, it has also recently been noted that autoregressive models can produce diverse and high-quality images from text. Thus, I plan to consider combining autoregressive and diffusion models for image generation and determine the best opportunities to combine their strengths.

6.4 Ethical implications of Text-to-Image synthesis

Text-to-image synthesis has a multitude of useful applications, such as enabling artists to create more lifelike digital art and filmmakers to produce more realistic special effects. However, the development of this technique also raises several concerns such as:

- Ethical challenges in media: Text-to-image synthesis can be misused to create false news or misleading information, leading to ethical dilemmas and casting doubt on the credibility and authenticity of news sources.
- Privacy risks: Text-to-image synthesis technologies can be used to create fake photos and videos that may infringe on personal privacy, causing concerns around privacy and raising suspicions about the use of social media and other online platforms.
- Legal challenges: Text-to-image synthesis technologies can be used for fraudulent purposes, which can lead to legal challenges and liability. For instance, false evidence presented in court or fake images used in commercial transactions can result in severe legal consequences.
- Bias issues: Text-to-image synthesis requires a large amount of training data to produce images. If the training dataset has particular biases, such as gender, race, age, etc., then the resulting images may also reflect the same biases.

In summary, the emergence and development of text-to-image synthesis technologies may have various social implications, including media ethics, privacy risks, legal challenges, and some bias issues. To avoid these issues, it is necessary to use as diverse a training dataset as possible and to manually review the input text to avoid any bias or stereotypes. The generated images also need to be manually vetted to ensure that they do not contain any inappropriate content.

References

Agnese, J., Herrera, J., Tao, H. & Zhu, X. (2020a). A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(4), e1345.

Agnese, J., Herrera, J., Tao, H. & Zhu, X. (2020b). A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*.

Albawi, S., Mohammed, T. A. & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*, 1-6.

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M. & Asari, V. K. (2018). Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *ArXiv*, *abs/1802.06955*.

Arjovsky, M. & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. *ArXiv*, *abs/1701.04862*.

Bai, S. & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304.

Barratt, S. T. & Sharma, R. (2018). A note on the inception score. *ArXiv*, *abs/1801.01973*.

Borji, A. (2022). Pros and cons of gan evaluation measures: New developments. *Comput. Vis. Image Underst.*, *215*, 103329.

Brock, A., Donahue, J. & Simonyan, K. (2019). Large scale gan training for high fidelity natural image synthesis. *ArXiv*, *abs/1809.11096*.

Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y. & Chen, G. (2020). Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, *10 6*, 1275-1285.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *ArXiv*, *abs/2105.05537*.

Cao, H. K., Tan, C., Gao, Z., Chen, G., Heng, P.-A. & Li, S. Z. (2022). A survey on generative diffusion model. *ArXiv*, *abs/2209.02646*.

Chang, X., Ren, P., Xu, P., Li, Z., Chen, X. & Hauptmann, A. G. (2021). A comprehensive survey of scene graphs: Generation and application. *IEEE transactions on pattern analysis and machine intelligence*, *PP*.

Chen, F., Wang, Y. C., Wang, B. & Kuo, C.-C. J. (2020). Graph representation learning:

a survey. *APSIPA Transactions on Signal and Information Processing*, *9*.

Cheng, J., Wu, F., Tian, Y., Wang, L. & Tao, D. (2020). Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10908-10917.

Chidester, B., Do, M. N. & Ma, J. (2018). Rotation equivariance and invariance in convolutional neural networks. *arXiv preprint arXiv:1805.12301*.

Cho, K., van Merrienboer, B., Çaglar Gülçehre, Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Emnlp*.

Crowson, K., Biderman, S. R., Kornis, D., Stander, D., Hallahan, E., Castricato, L. & Raff, E. (2022). Vqgan-clip: Open domain image generation and editing with natural language guidance. *ArXiv*, *abs/2204.08583*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Naacl*.

Dhariwal, P. & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *ArXiv*, *abs/2105.05233*.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., ... Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. In *Neurips*.

Efimova, V., Jarsky, I., Bizyaev, I. & Filchenkov, A. (2022). Conditional vector graphics generation for music cover images. *arXiv preprint arXiv:2205.07301*.

Esfahani, S. N. & Latifi, S. (2019, 10). Image generation with gans-based techniques: A survey. *International Journal of Computer Science and Information Technology*, *11*, 33-50. doi: 10.5121/ijcsit.2019.11503

Frolov, S., Hinz, T., Raue, F., Hees, J. & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. *Neural Networks*, *144*, 187–209.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D. & Taigman, Y. (2022). Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv*, *abs/2203.13131*.

Ganar, A. N., Gode, C. & Jambhulkar, S. M. (2014). Enhancement of image retrieval by using colour, texture and shape features. In *2014 international conference on electronic systems, signal processing and computing technologies* (pp. 251–255).

Gao, L., Wang, B. & Wang, W. (2018). Image captioning with scene-graph based semantic concepts. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*.

Ghosh, B., Dutta, I. K., Totaro, M. & Bayoumi, M. (2020). A survey on the progression and performance of generative adversarial networks. In *2020 11th international conference on computing, communication and networking technologies (icccnt)* (pp. 1–8).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Nips*.

Grover, A. & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Gu, J., Joty, S. R., Cai, J., Zhao, H., Yang, X. & Wang, G. (2019). Unpaired image captioning via scene graph alignments. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10322-10331.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., ... Guo, B. (2022, June). Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 10696-10706).

Hamilton, W. L. (2020). Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*.

Hamilton, W. L., Ying, R. & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *ArXiv*, *abs/1709.05584*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Nips*.

Ho, J. (2022). Classifier-free diffusion guidance. *ArXiv*, *abs/2207.12598*.

Ho, J., Jain, A. & Abbeel, P. (2020). Denoising diffusion probabilistic models. *ArXiv*, *abs/2006.11239*.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M. & Salimans, T. (2022). Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, *23*, 47:1-47:33.

Hong, S., Yang, D., Choi, J. & Lee, H. (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7986–7994).

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., ... Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1055-1059.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J. & Belongie, S. (2017). Stacked generative adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5077–5086).

Ibtehaz, N. & Rahman, M. S. (2020). Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks : the official journal of the International Neural Network Society*, *121*, 74-87.

Jain, A., Modi, D., Jikadra, R. & Chachra, S. D. (2019). Text to image generation of fashion clothing. *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, 355-358.

Jaritz, M., Vu, T.-H., de Charette, R., Wirbel, É. & Pérez, P. (2020). xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12602-12611.

Johnson, J., Gupta, A. & Fei-Fei, L. (2018). Image generation from scene graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1219-1228.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S. & Fei-Fei, L. (2015). Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668-3678.

Kamilaris, A. & Prenafeta-Boldú, F. X. (2018). A review of the use of convolutional neural networks in agriculture. *The Journal of Agricultural Science*, *156*, 312 - 322.

Karpathy, A. & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3128–3137).

Kauderer-Abrams, E. (2017). Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*.

Ketkar, N. S. (2021). Convolutional neural networks. *Deep Learning with Python*.

Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 1 - 62.

Kim, D., Joo, D. & Kim, J. (2020). Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, *8*, 153113–153122.

Lee, H., Ullah, U., Lee, J.-S., Jeong, B. & Choi, H.-C. (2021). A brief survey of text driven image generation and maniulation. In *2021 ieee international conference on consumer electronics-asia (icce-asia)* (pp. 1–4).

Li, J., Yan, Y., Liao, S., Yang, X. & Shao, L. (2021). Local-to-global self-attention in vision transformers. *arXiv preprint arXiv:2107.04735*.

Li, L., Gan, Z., Cheng, Y. & Liu, J. (2019). Relation-aware graph attention network for visual question answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10312-10321.

Li, R., Li, W., Yang, Y. & Bai, Q. (2022). Obj-sa-gan: Object-driven text-to-image synthesis with self-attention based full semantic information mining. In *Pricai 2022: Trends in artificial intelligence* (pp. 339–350). Springer Nature Switzerland.

Li, R., Li, W., Yang, Y., Wei, H., Jiang, J. & Bai, Q. (2022). Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *ArXiv*, *abs/2210.09549*.

Li, R., Wang, N., Feng, F., Zhang, G. & Wang, X. (2020). Exploring global and local linguistic representations for text-to-image synthesis. *IEEE Transactions on Multimedia*, *22*(12), 3075–3087.

Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S. & Gao, J. (2019a). Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12174–12182).

Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S. & Gao, J. (2019b). Object-driven text-to-image synthesis via adversarial training. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12166-12174.

Li, Y., Ma, T., Bai, Y., Duan, N., Wei, S. & Wang, X. (2019). Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems*, *32*.

Liang, C., Wang, W., Zhou, T., Miao, J., Luo, Y. & Yang, Y. (2022). Local-global context aware transformer for language-guided video segmentation. *arXiv preprint arXiv:2203.09773*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11999-12009.

Luo, C., Zhan, J., Wang, L. & Yang, Q. (2018). Cosine normalization: Using cosine similarity instead of dot product in neural networks. *ArXiv*, *abs/1702.05870*.

Luo, W., Li, Y., Urtasun, R. & Zemel, R. S. (2016). Understanding the effective receptive field in deep convolutional neural networks. *ArXiv*, *abs/1701.04128*.

Mathesul, S., Bhutkar, G. & Rambhad, A. (2021). Attngan: realistic text-to-image synthesis with attentional generative adversarial networks. In *Ifip conference on human-computer interaction* (pp. 397–403).

Mikolov, T., Chen, K., Corrado, G. S. & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Iclr.*

Mirza, M. & Osindero, S. (2014). Conditional generative adversarial nets. *ArXiv*, *abs/1411.1784*.

Mittal, G., Agrawal, S., Agarwal, A., Mehta, S. & Marwah, T. (2019). Interactive image generation using scene graphs. *arXiv preprint arXiv:1905.03743*.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... Chen, M. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Icml.*

Ning, X., Nan, F., Xu, S., Yu, L. & Zhang, L. (2020). Multi-view frontal face image generation: a survey. *Concurrency and Computation: Practice and Experience*, e6147.

Odena, A., Olah, C. & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *Icml.*

Park, D. H., Azadi, S., Liu, X., Darrell, T. & Rohrbach, A. (2021). Benchmark for compositional text-to-image synthesis. In *Neurips datasets and benchmarks.*

Pavan Kumar, M. & Jayagopal, P. (2021). Generative adversarial networks: a survey on applications and challenges. *International Journal of Multimedia Information Retrieval*, *10*(1), 1–24.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77. Retrieved from `https://doi.org/10.2753/MIS0742-1222240302` doi: 10.2753/MIS0742-1222240302

Qi, Z., Sun, J., Qian, J., Xu, J. & Zhan, S. (2021). Pccm-gan: Photographic text-to-image generation with pyramid contrastive consistency model. *Neurocomputing*,

449, 330-341.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Icml.*

Radford, A. & Narasimhan, K. (2018). Improving language understanding by generative pre-training..

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners..

Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, *abs/1910.10683*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *ArXiv*, *abs/2204.06125*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. & Lee, H. (2016). Generative adversarial text to image synthesis. In *International conference on machine learning* (pp. 1060–1069).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674-10685.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, *abs/1505.04597*.

Russell, S. J. & Norvig, P. (2010). Artificial intelligence - a modern approach, third international edition..

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, *abs/2205.11487*.

Sainath, T. N., Vinyals, O., Senior, A. W. & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4580-4584.

Schuster, S., Krishna, R., Chang, A. X., Fei-Fei, L. & Manning, C. D. (2015). Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Vl@emnlp.*

Schwing, A. G. & Urtasun, R. (2015). Fully connected deep structured networks. *ArXiv*, *abs/1503.02351*.

Shamsolmoali, P., Zareapoor, M., Granger, E., Zhou, H., Wang, R., Celebi, M. E. & Yang, J. (2021). Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, *72*, 126–146.

Shelhamer, E., Long, J. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431-3440.

Singh, N. K. & Raza, K. (2021). Medical image generation using generative adversarial networks: a review. *Health Informatics: A Computational Perspective in Healthcare*, 77–96.

Song, J., Meng, C. & Ermon, S. (2021). Denoising diffusion implicit models. *ArXiv*, *abs/2010.02502*.

Taghanaki, S. A., Abhishek, K., Cohen, J. P., Cohen-Adad, J. & Hamarneh, G. (2020). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, *54*, 137-178.

Tan, H., Liu, X., Li, X., Zhang, Y. & Yin, B. (2019). Semantics-enhanced adversarial nets for text-to-image synthesis. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10501–10510).

Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K. & Xu, C. (2022, June). Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 16515-16525).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, F. & Tax, D. M. (2016). Survey on the attention based rnn model and its applications in computer vision. *arXiv preprint arXiv:1601.06823*.

Wang, X., Tu, Z., Wang, L. & Shi, S. (2019). Self-attention with structural position representations. In *Emnlp*.

Wu, X., Xu, K. & Hall, P. (2017). A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology*, *22*(6), 660–674.

Xia, W., Yang, Y., Xue, J. & Wu, B. (2021). Tedigan: Text-guided diverse face image generation and manipulation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2256-2265.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1316–1324).

Yang, L., Zhang, Z., Hong, S., Xu, R., Zhao, Y., Shao, Y., ... Cui, B. (2022). Diffusion models: A comprehensive survey of methods and applications. *ArXiv*, *abs/2209.00796*.

Yang, X., Tang, K., Zhang, H. & Cai, J. (2019). Auto-encoding scene graphs for image captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10677-10686.

Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354–7363).

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H. & Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 833-842.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. & Metaxas, D. N. (2017).

Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 5907–5915).

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. & Metaxas, D. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, *41*(8), 1947–1962.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. & Metaxas, D. N. (2019). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*, 1947-1962.

Zhang, S., Dong, H., Hu, W., Guo, Y., Wu, C., Xie, D. & Wu, F. (2018). Text-to-image synthesis via visual-memory creative adversarial network. In *Pacific rim conference on multimedia* (pp. 417–427).

Zhang, Z., Liu, Q. & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, *15*, 749-753.

Zhao, B., Meng, L., Yin, W. & Sigal, L. (2019). Image generation from layout. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8584–8593).

Zhao, Z.-Q., Zheng, P., Xu, S.-t. & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, *30*(11), 3212–3232.

Zhong, Y., Wang, L., Chen, J., Yu, D. & Li, Y. (2020). Comprehensive image captioning via scene graph decomposition. *ArXiv*, *abs/2007.11731*.

Zhou, R., Jiang, C. & Xu, Q. (2021a). A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, *451*, 316–336.

Zhou, R., Jiang, C. & Xu, Q. (2021b). A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, *451*, 316-336.

Zhou, Y. & Shimada, N. (2021). Generative adversarial network for text-to-face synthesis and manipulation with pretrained bert model. In *2021 16th ieee international conference on automatic face and gesture recognition (fg 2021)* (pp. 01–08).

Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., ... Sun, T. (2022). Towards language-free training for text-to-image generation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17886-17896.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...*, *11045*, 3-11.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. (2020). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*, 1856-1867.

Zhu, B. & Ngo, C.-W. (2020). Cookgan: Causality based text-to-image synthesis. In

Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 5519–5527).

Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., ... Bennamoun (2022). Scene graph generation: A comprehensive survey. *ArXiv*, *abs/2201.00443*.

Zhu, M., Pan, P., Chen, W. & Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5795-5803.

Appendix A

Diffusion model calculation process

A.1 Forward diffusion calculation

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_1 \tag{A.1}$$

$$x_{t-1} = \sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_2 \tag{A.2}$$

Combining Equations A.1 and A.2:

$$\begin{aligned} x_t &= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon_2\right) + \sqrt{1-\alpha_t}\epsilon_1 \\ &\quad (\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \mathbf{I})) \\ &= \sqrt{\alpha_t\alpha_{t-1}} x_{t-2} + \left(\sqrt{\alpha_t(1-\alpha_{t-1})}\epsilon_2\right) + \sqrt{1-\alpha_t}\epsilon_1 \\ &= \sqrt{\alpha_t\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon_2 \\ &\longrightarrow \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \end{aligned} \tag{A.3}$$

A.2 Reverse diffusion calculation

$$
\begin{aligned}
q(x_{t-1}|x_t) &= q(x_{t-1}|x_t, x_0) \\
&= q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \\
&= e^{-\frac{1}{2}(\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{\beta_t}+\frac{(x_{t-1}\sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(x_t-\sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t})} \\
&= e^{-\frac{1}{2}((\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}})x_{t-1}^2-(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t+\frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0)x_{t-1})+C(x_t,x_0)}
\end{aligned}
\tag{A.4}
$$

Standard Gaussian distribution is:

$$
e^{-\frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{1}{2}(\frac{1}{\sigma^2}x^2-\frac{2\mu}{\sigma^2}x+\frac{\mu^2}{\sigma^2})} \tag{A.5}
$$

Combining Equations A.4 and A.5:

$$
\begin{aligned}
\mu_t(x_t, x_0) &= \frac{\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0}{\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}} \\
&= (\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \\
&= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0
\end{aligned}
\tag{A.6}
$$

x_0 could be represented by x_t, the connection between x_0 and x_t could refer to Equation A.3:

$$
x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \tag{A.7}
$$

Combining Equations A.6 and A.7:

$$
\begin{aligned}
\mu_t &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \\
&= \frac{\sqrt{\alpha_t}\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \\
&= \frac{\alpha_t - \bar{\alpha}_{t-1}}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}x_t + \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t) \\
&= \frac{1-\bar{\alpha}_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}x_t - \frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}(\sqrt{1-\bar{\alpha}_t}\epsilon_t) \\
&= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{\beta_t}{\sqrt{(1-\bar{\alpha}_t)}\sqrt{\alpha_t}}\epsilon_t \\
&= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{(1-\bar{\alpha}_t)}}\epsilon_t)
\end{aligned}
\tag{A.8}
$$

Appendix B

More generated images by Swinv2-Imagen

Figure B.1: A kitchen is shown with a variety of items on the counters.

Figure B.2: A full view of an open kitchen and dining area.

Figure B.3: Two dogs are looking up while they stand near the toilet in the bathroom.

Figure B.4: A view of a very large bathroom with mirrored walls.

Figure B.5: A colourful bird.

Appendix C

Training and Testing environment

dependencies:

- absl-py==1.2.0
- accelerate==0.12.0
- aiohttp==3.8.1
- aiosignal==1.2.0
- async-timeout==4.0.2
- attrs==22.1.0
- cachetools==5.1.0
- click==8.1.3
- cycler==0.11.0
- diffusers==0.3.0
- einops==0.4.1

- einops-exts==0.0.3
- ema-pytorch==0.0.10
- filelock==3.8.0
- fonttools==4.37.1
- frozenlist==1.3.1
- fsspec==2022.7.1
- google-auth==2.11.0
- google-auth-oauthlib==0.4.6
- grpcio==1.47.0
- huggingface-hub==0.9.1
- imagen-pytorch==1.11.0
- importlib-metadata==4.12.0
- kiwisolver==1.4.4
- kornia==0.6.6
- markdown==3.4.1
- markupsafe==2.1.1
- matplotlib==3.5.3
- multidict==6.0.2
- oauthlib==3.2.0

- packaging==21.0
- protobuf==3.19.4
- psutil==5.9.1
- pyasn1==0.4.8
- pyasn1-modules==0.2.8
- pydantic==1.9.2
- pydeprecate==0.3.2
- pyparsing==3.0.9
- python-dateutil==2.8.1
- pytorch-lightning==1.7.2
- pytorch-warmup==0.0.4
- regex==2022.7.25
- requests-oauthlib==1.3.1
- resize-right==0.0.2
- rsa==4.9
- sentencepiece==0.1.97
- tensorboard==2.10.0
- tensorboard-data-server==0.6.1
- tensorboard-plugin-wit==1.8.1

- timm==0.5.4
- tokenizers==0.12.1
- torchmetrics==0.9.3
- tqdm==4.63.1
- transformers==4.21.1
- werkzeug==2.2.2
- yarl==1.8.1
- zipp==3.8.1

Appendix D

Glossary

BCELoss	Binary Cross Entropy Loss
BERT	Bidirectional Encoder Representations from Transformers
BFS	Breadth-First Search
bi-convLSTM	bidirectional convolutional LSTM
CGAN	Conditional Generative Adversarial Networks
CLIP	Contrastive Language-Image Pre-Training
CNNs	Convolutional Neural Networks
CUB	Caltech-UCSD Birds-200-2011
CV	Computer Vision
DFS	Depth First Search
DMGAN	Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis
DRAW	Deep Recurrent Attention Writer

FCN	Fully Convolutional Networks
FID	Fréchet Inception Distance
GANs	Generative Adversarial Networks
GCN	Graph Convolution Network
GELU	Gaussian Error Linear Unit
GLID	Guided Language-to-Image Diffusion for Generation and Editing
GPT	Generative Pre-trained Transformer
IS	Inception Score
LN	LayerNorm
MLP	Multilayer Perceptron
MM CelebA-HQ	Multi-modalCelebA-HQ
MSCOCO	Microsoft Common Objects in Context
MSE	Mean Squared Error
NLLLoss	Negative Log Likelihood Loss
NLP	Natural Language Processing
Obj-SA-GAN	Object-driven Self-Attention Generative Adversarial Network
PCCM-GAN	Photographic Text-to-Image Generation with Pyramid Contrastive Consistency Model
RNNs	Recurrent Neural Networks
SDGAN	Semantics Disentangling for Text-to-Image Generation

Swinv2-Imagen	Hierarchical Vision Transformer Diffusion Models for Text-to-Image Generation
T5	Text-to-Text Transfer Transformer
VAE	Variational Auto-Encoder