Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction

SEMINAR REPORT

Submitted by

NANDAKRISHNAN O

(MZW21CS013)

to

the A P J Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MOUNT ZION INSTITUTE OF SCIENCE AND TECHNOLOGY

KOZHUVALLOOR, CHENGANNUR

OCTOBER 2024

Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction

SEMINAR REPORT

Submitted by

NANDAKRISHNAN O

(MZW21CS013)

to

the A P J Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MOUNT ZION INSTITUTE OF SCIENCE AND TECHNOLOGY

KOZHUVALLOOR, CHENGANNUR

OCTOBER 2024

DECLARATION

I, Nandakrishnan O hereby declare that the seminar report **Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction**, submitted for partial fulfillment of the requirements for the award of the degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Ms. Chitra Shaji Thomas, Assistant Professor, Department of Computer Engineering, Mount Zion Insitute of Science and Technology,Chengannur .

This submission represents ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources.

I also declare that I have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and / or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Chengannur Nandakrishnan O

26-10-2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MOUNT ZION INSTITUTE OF SCIENCE AND TECHNOLOGY
KOZHUVALLOOR

CERTIFICATE

This is to certify the seminar report entitled **" Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction"** is a bonafide record of the work done by **NANDAKRISHNAN O (MZW21CS013)** under our guidance and supervision towards partial fulfilment of the requirements for the award of Bachelor of Technology Degree in Computer science Engineering, of APJ Abdul Kalam Kerala Technological University during the year 2021-2025 and this work has not been submitted else where the award of any degree.

Seminar Coordinator	Head of the Department	Seminar Guide
Ms.Nandana S Kumar	Mr.Jacob Joseph	Ms.Chithra Shaji Thomas
Assistant Professor	Assistant Professor	Assistant Professor
Dept. Of Computer Science & Engg	Dept. Of Computer Science & Engg	Dept. Of Computer Science & Engg

ACKNOWLEDGEMENT

First of all, with prayers to God for his grace and blessings, for without which his unforeseen guidance,this seminar would have remained only in dream. I would like to express my profound gratitude to all the inspired and motivated me to make the semiar a sucess.

My seminar was only possible because of the encouragement I received from all quarters. I take this opportunity to thank our principal, **Prof. Dr. K Mathew** for granting permission to do my seminar work.

I would like to thank **Asst.Professor. Mr. Jacob Joseph** (Head of the computer science and engineering department), for her advices and support during the work of seminar.

I express my thanks to **Asst.Professor. Mrs. Nandana S Kumar** (co-ordinator) of the computer science and engineering department and the source of strength in completing this seminar.

I express my heart felt gratitude to **Asst.Professor. Mrs. Anpu Ann Thomas** computer science and engineering department, for her valuable guidance, timely advice and much appreciated corrections and support during every step of my work for preparing the seminar.

I would like to thank all the faculties of Computer Science and Engineering Department for their support and suggestions for this seminar.

I would like to extend my special thanks to my parents for their love, support and encouragement.

NANDAKRISHNAN O

ABSTRACT

Text-to-image synthesis, the process of turning words into images, opens up a world of creative possibilities, and meets the growing need for engaging visual experiences in a world that is becoming more image-based. As machine learning capabilities expanded, the area progressed from simple tools and systems to robust deep learning models that can automatically generate realistic images from textual inputs. Modern, large-scale text-to-image generation models have made significant progress in this direction, producing diversified and high-quality images from text description prompts. Although several methods exist, Generative Adversarial Networks (GANs) have long held a position of prominence. However, diffusion models have recently emerged, with results much beyond those achieved by GANs. This study offers a concise overview of text-to-image generative models by examining the existing body of literature and providing a deeper understanding of this topic. This will be accomplished by providing a concise summary of the development of text-to-image synthesis, previous tools and systems employed in this field, key types of generative models, as well as an exploration of the relevant research conducted on GANs and diffusion models. Additionally, the study provides an overview of common datasets utilized for training the text-to image model, compares the evaluation metrics used for evaluating the models, and addresses the challenges encountered in the field. Finally, concluding remarks are provided to summarize the findings and implications of the study and open issues for further research.

INDEX TERMS Deep learning, diffusion model, generative models, generative adversarial network, text to-image synthesis.

Contents

Contents

List of Figures

List of Tables

ABBREVIATIONS

No	Acronym	Meaning
1	GAN	Generative Adversarial Network
2	VAE	Variational Autoencoder
3	GACM	Guided Affine Combination Module
4	GDCM	Guided Detail Correction Module
5	FTM	Feature Transformation Module
6	LSTM	Long Short-Term Memory
7	CMPC	Crossmodal Progressive Comprehension
8	TGFE	Text-Guided Feature Exchange
9	MS-COCO	Microsoft Common Objects in Context (Dataset)
10	VGG	Visual Geometry Group (a neural network model)

CHAPTER 1
INTRODUCTION

In recent years, **text-guided image manipulation** has gained significant attention in the field of **computer vision** and **generative models**. This technique, which involves altering images based on textual descriptions, offers a wide range of applications in areas such as **graphic design**, **content creation**, and **multimedia production**. As the demand for automated and user-friendly image editing tools has grown, traditional approaches to text-guided image manipulation have encountered several challenges. These challenges become especially apparent when handling **complex scenarios** with multiple objects in an image, where earlier models struggled to achieve accurate and meaningful manipulation. Typically, these models, trained on simpler datasets, underperform when confronted with diverse and intricate scenes.

To address these limitations, the **Text-Guided Image Manipulation via Generative Adversarial Networks (GAN) with Referring Image Segmentation-Based Guidance** proposes an innovative method. This approach capitalizes on the advantages of **segmentation guidance** by utilizing **referring image segmentation** to improve manipulation precision. The method generates a **segmentation mask** that effectively extracts the text-related regions from an input image, ensuring that image manipulation occurs only in relevant areas. Unlike previous approaches, the segmentation mask enables the manipulation of complex images with multiple objects while preserving the consistency between manipulated and non-manipulated regions.

Additionally, the proposed method leverages the **Contrastive Language-Image Pretraining (CLIP) model** as a loss function to incorporate fine-grained textual attributes, such as **color** and **texture**, into the manipulated image. This integration ensures that the visual attributes in the manipulated region align precisely with the text description, enhancing both the **visual and semantic consistency** of the final output. The ability to explicitly identify and preserve text-unrelated regions during manipulation represents a significant improvement over past methods, making it possible to handle **multiple-object scenes** with high precision.

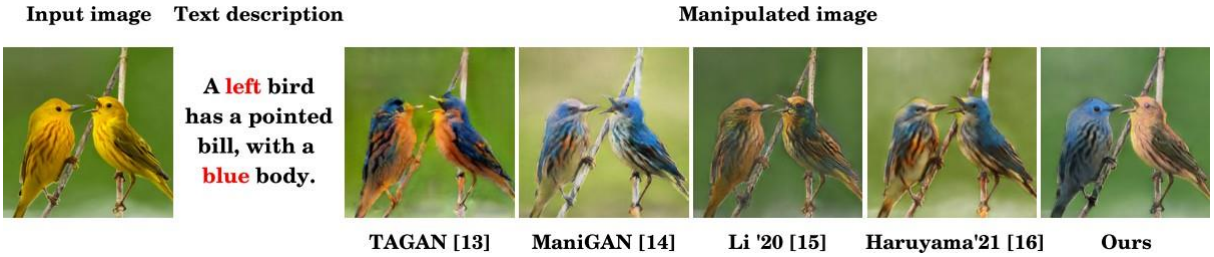

Figure 1.1: Examples of text-guided image manipulation.

CHAPTER 2

LITERATURE REVIEW

1. Image style transfer using convolutional neural networks[1]

It introduce a novel approach to image style transfer leveraging convolutional neural networks (CNNs). Style transfer involves the transformation of the content of an image to adopt the artistic style of another reference image. The authors propose a method that utilizes the representations learned by CNNs to separate and manipulate content and style features independently. By defining a suitable loss function that captures the content and style of the images, the authors demonstrate the effectiveness of their approach in generating visually appealing stylized images. The method not only provides a powerful tool for artistic expression but also contributes to the understanding of deep neural network representations in the context of image processing and aesthetics. The experimental results and comparisons with existing techniques highlight the capabilities and potential applications of the proposed CNN-based image style transfer method.

2. Auto-Encoding Variational Bayes[4]

Auto-Encoding Variational Bayes[4] represents an innovative approach to learning probabilistic representations of data using deep learning techniques. The proposed method, termed Variational Autoencoder (VAE), combines the power of autoencoders with probabilistic modeling, specifically within the Bayesian framework. The VAE framework enables the generation of new data samples by learning a probabilistic mapping from the observed data to a latent space. This latent space representation is structured to follow a probabilistic distribution, facilitating meaningful interpolation and generation of new samples. The training of the VAE involves optimizing a variational lower bound on the log-likelihood of the data. The authors demonstrate the

effectiveness of VAEs in generating realistic samples and performing tasks such as data reconstruction. The paper contributes to the field of generative models and provides a foundation for utilizing variational inference in the context of autoencoders.

1. Deep image prior [7]

Deep image prior [7] introduces the concept of a "deep image prior," which is a convolutional neural network (CNN) trained to map random noise to a natural-looking image. The authors demonstrate that this randomly-initialized CNN can be used as a powerful prior in a variety of image processing tasks, such as denoising, super-resolution, and inpainting. Unlike traditional priors, the deep image prior is data-dependent and adapts to the specific image being processed. The paper explores the effectiveness of this approach in various applications and highlights its potential for solving inverse problems in computer vision.

2. Deep Colorization [8]

Deep Colorization [8] introduce an approach to automatic image colorization using deep learning techniques. The authors propose a deep neural network architecture that leverages convolutional neural networks (CNNs) to learn and predict color information from grayscale images. By training on a large dataset of grayscale and corresponding color images, the model learns complex relationships between features in black-and-white images and their corresponding color representations. The results demonstrate the effectiveness of the proposed deep colorization method in generating realistic and visually appealing colorizations, showcasing the potential of deep learning for automatic image colorization tasks.

3. A neural algorithm of artistic style [9]

A neural algorithm of artistic style [9] introduced a novel approach using Convolutional Neural Networks (CNNs) to extract both content and style features from images. By defining a content loss function based on the difference between feature representations of content images and generated images, and a style loss function based

on the Gram matrices of feature maps for style images, the algorithm effectively generated visually appealing images that combine the content of one image with the artistic style of another. The method demonstrated the ability to produce diverse and aesthetically pleasing results, marking a significant advancement in the field of neural style transfer and contributing to the intersection of deep learning and artistic expression.

1. Politeness transfer: A tag and generate approach [10]

Politeness transfer: A tag and generate approach [10] introduces a novel method for transferring politeness levels in natural language. The proposed approach utilizes a "tag and generate" strategy, where politeness tags are added to input sentences, and a generative model is employed to produce polite versions. The authors leverage a large dataset for training, aiming to improve the politeness of text while preserving the original content and meaning. The research contributes to the broader field of natural language processing, addressing the important task of politeness transfer in human-computer interactions and communication systems.

2. Text-adaptive generative adversarial networks: Manipulating images with natural language [13]

Text-adaptive generative adversarial networks: Manipulating images with natural language [13] introduces a Text-adaptive Generative Adversarial Networks (Text-GANs), a framework that leverages textual descriptions to guide the generation of realistic and contextually relevant images. The model aims to bridge the gap between textual and visual domains, allowing users to provide natural language instructions to manipulate and generate images. The paper discusses the architecture and mechanisms of Text-GANs, demonstrating its efficacy in generating visually coherent results based on textual input. This work contributes to the intersection of natural language processing and image generation, offering a promising avenue for interactive and intuitive image synthesis.

3. Controllable text-to-image generation [22]

The Controllable text-to-image generation [22] propose a method to generate images from textual descriptions while providing users with control over specific attributes of the generated images. They introduce a novel framework that leverages disentangled representations, allowing for the manipulation of desired image characteristics through textual prompts. The approach is designed to be adaptable to diverse input texts, enabling users to influence and fine-tune visual aspects of the generated images. The proposed controllable text-to-image generation framework contributes to advancing the field of generative models by enhancing the interpretability and control of the image synthesis process based on textual input.

3. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks [23]

AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks [23] introduces a novel approach for generating high-quality images from textual descriptions. The proposed model utilizes attentional generative adversarial networks (GANs) to enhance the synthesis process by focusing on relevant parts of the text during image generation. This attention mechanism enables the model to capture fine grained details and improve the overall realism of the generated images. The paper demonstrates the effectiveness of AttnGAN through experiments and showcases its capability in producing visually compelling results in the challenging task of translating detailed textual descriptions into corresponding images.

4. Segmentation from natural language expressions [29]

The Segmentation from natural language expressions [29] proposes a approach that leverages both image and language modalities to achieve accurate segmentation. They introduce a model that incorporates a semantic segmentation network with a recurrent neural network (RNN) to parse and understand natural language expressions. By combining information from visual and textual cues, the system is designed to generate

precise segmentation masks based on the contextual understanding of the input image conveyed through the accompanying natural language descriptions. The results demonstrate the effectiveness of their approach in enhancing segmentation performance through the integration of linguistic context.

5. Protecting the trust and credibility of data by tracking forgery trace based on GANs [34]

Protecting the trust and credibility of data by tracking forgery trace based on GANs [34] proposes a method for tracing and detecting forgery in digital communication networks to ensure the reliability and trustworthiness of data. The paper introduces a framework that leverages GANs to identify and track potential forgery traces, contributing to the enhancement of data security in the context of digital communication. The study, published in December 2022 in the journal Digital Communications and Networks, aims to address the challenges associated with data credibility in the era of digital communication and offers a promising avenue for maintaining trust in digital information.

6. Very deep convolutional networks for large-scale image recognition [35]

Very deep convolutional networks for large-scale image recognition [35] introduces the VGGNet architecture, a deep convolutional neural network (CNN) designed for image classification tasks. The key innovation lies in the use of very small 3x3 convolutional filters throughout the network, leading to a deep architecture with 16 or 19 weight layers. This uniformity in filter size simplifies the network structure and aids in learning hierarchical features. The VGGNet achieved top performance in the 2014 ImageNet Large Scale Visual Recognition Challenge, demonstrating the effectiveness of deeper networks for image classification tasks. The straightforward design and consistent architecture of VGGNet have since served as a foundation for subsequent developments in deep learning for computer vision.

7. Going deeper with convolutions[36]

Going deeper with convolutions[36] introduces the Inception architecture, a deep

convolutional neural network (CNN) that aims to address the challenges of training very deep networks. The Inception architecture employs a novel module called the “Inception module,” which performs convolutions with multiple filter sizes concurrently and captures information at different spatial scales. This approach allows the network to efficiently learn hierarchical features and achieve state-of-the- art performance on image classification tasks.

CHAPTER 3

METHODOLOGY

The methodology section focuses on the various generative models employed in text-to-image synthesis, a complex area of machine learning and computer vision. This section delves into the technical foundations and innovations behind key generative approaches, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), flow-based models, and diffusion models. Each method brings unique strengths to the task of generating high-quality, realistic images from textual descriptions. GANs, for instance, leverage a dynamic adversarial process between generator and discriminator models, while VAEs utilize a probabilistic approach to compress and reconstruct data. Flow-based models focus on invertible transformations, ensuring both efficient encoding and decoding, and diffusion models represent a cutting-edge technique by iteratively adding noise to data and learning to reverse the process. These methodologies together represent the forefront of advancements in text-to-image synthesis, providing the foundation for both improved image realism and semantic alignment with input text.

3.1 GENERATIVE ADVERSARIAL NETWORKS

In 2014, Goodfellow et al. [2] introduced GANs, one of the well-known generating models. From that point forward, several additional models based on the concept of GANs were developed to address the previous shortcomings. GANs can be used in many different contexts, such as to make images of people's faces, to make realistic photos, to make cartoon characters, to age people's faces, to increase resolution, to translate between images and words, and so on [4]. GANs consist of two major sub-models: generator and discriminator. The generator is in charge of making new fake images by taking a noise vector as an input and putting out an image as an output. On the other hand, the discriminator's job is to tell the difference between real and fake images after being trained with real data. In other words, it serves as a classification network that is capable of classifying images by returning 0 for fake and 1 for real. Therefore, the generator's goal is to create convincing fakes in order to trick the discriminator, while the discriminator's goal is to recognize the difference [1]. Training improves both the discriminator's ability to distinguish between real or fake images, and the generator's ability to produce realistic-looking images. When the discriminator can no longer tell genuine images from fraudulent ones, equilibrium has been reached.

3.2 VARIATIONAL AUTOENCODER (VAE)

The utilization of a variational autoencoder (VAE) [17] provides a probabilistic framework for representing an observation inside a latent space. The input is subjected to encoding, which frequently involves compressing information into a latent space of reduced dimensionality. The primary objective of autoencoders is to effectively encode and represent the given data. The objective at hand involves the identification of a low-dimensional representation for a high dimensional input, which facilitates the reconstruction of the original input while minimizing the loss of content.

3.3FLOW-BASED GENERATIVE MODEL

Flow-based models are capable of learning distinct encoders and decoders. In a manner similar to the encoding phase observed in autoencoders, a transformation is employed to the data, with its parameters determined by a neural network [18]. Nevertheless, the decoder does not consist of a novel neural network that needs to autonomously acquire the decoding process; rather, it functions in direct opposition to its counterpart. In order to achieve the invertibility of a function ‘‘f’’ using neural networks, multiple strategies need to be employed.

3.4 DIFFUSION MODELS

As a subset of deep generative models, diffusion model shave recently been recognized as the cutting edge. The diffusion models have lately demonstrated significant results that have been proven to surpass GAN models [19] They have proven successful in a number of different areas, including the difficult task of image synthesis, where GAN shad previously dominated. Recently, diffusion models have become a hot topic in computer vision due to their impressive generative capabilities. The field of generative modelling has found many uses for diffusion models so far, including image generation, super-resolution, inpainting, editing, and translation between images [20]. The principles of nonequilibrium thermodynamics provide the basis for diffusion models. Before learning to rebuild desirable data examples from the noise, they generate a Markov chain of diffusion steps to gradually inject noise into data [20]. In order to learn, the diffusion model has two phases: one for forward diffusion and the other for backward diffusion. In the forward diffusion phase, Gaussian noise is progressively added to the input data at each level [21]. In the second

phase, called ‘‘reverse,’’ the model is charged to reverse the diffusion process so that the original input data can be recovered.

The architectures of generative model types are shown in Figure 2.

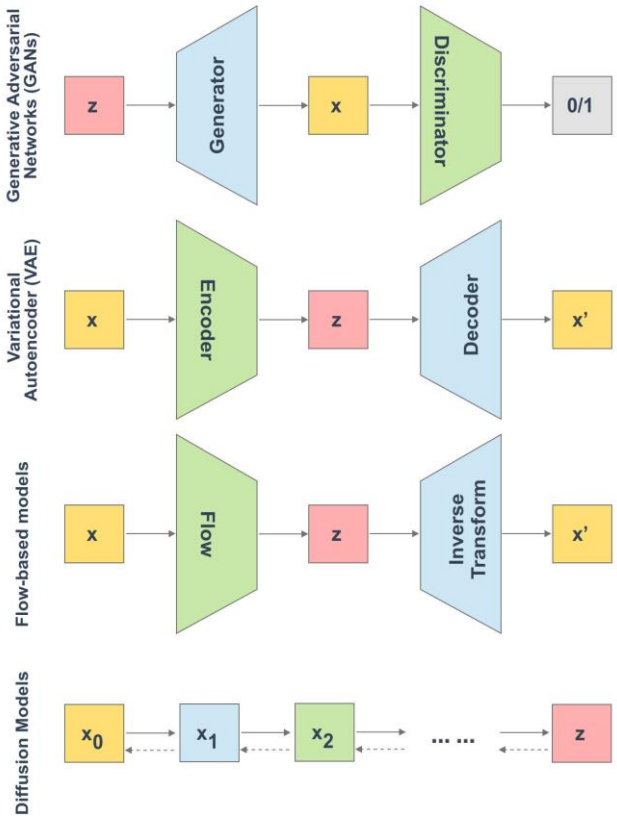

Figure 3.1: Types of generative models, reproduced from Weng [22].

3.5 DATASETS

Datasets play a crucial role in the development and evaluation of text-to-image generative models. In the realm of text-to image generative models, the utilization of diverse datasets is vital for achieving accurate and realistic visual outputs. This section will explore the various datasets frequently utilized in this research area. The most frequently used datasets by text-to-image synthesis models are:

A. MS COCO

Reference [31], known as the Microsoft Common Objects in Context, is a comprehensive compilation of images that is widely employed for the purpose of object detection and segmentation. The dataset comprises a collection of more than 330,000 images, with each image being accompanied by annotations for 80 object categories and 5 captions that provide descriptive information about the depicted scene. The COCO dataset is extensively utilized in the field of computer vision research and has been employed for the purposes of training and evaluating numerous cutting-edge models for object identification and segmentation.

B. CUB-200-2011

Caltech-UCSD Birds-200-2011 [32] is a popular dataset for fine-grained visual categorization. This dataset comprises 11,788 bird images from 200 subcategories. Images are divided into 5,994 training and 5,794 testing sets. Each image in the dataset has a subcategory, part location, binary attribute, and bounding box labels. Natural language descriptions supplemented these annotations to improve the CUB-200-2011 dataset. Each image received ten single sentence descriptions.

C. OXFORD 102 FLOWER

Reference [33] comprises a collection of 102 distinct categories of flowers, which can be effectively employed for image classification. The selected flowers were indigenous to the United Kingdom. The number of photos in each class ranges from 40 to 258. The images demonstrate significant variations in terms of size, pose, and lighting conditions. There exist categories that exhibit significant variations within their respective boundaries, as well as numerous categories that have notable similarities.

Figure 3 shows samples of images along with their captions from the MS COCO, Oxford 102 Flower, and CUB-200-2011 datasets.

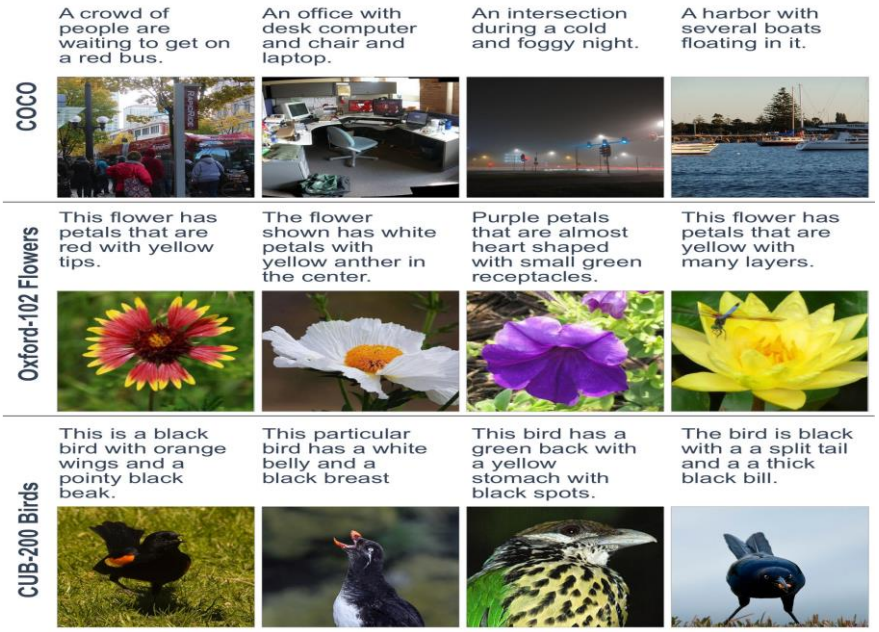

Figure 3.2: Sample images and their captions of common text-to-image datasets. Figure reproduced from Frolov et al.[1]

CHAPTER 4
RELATED WORK

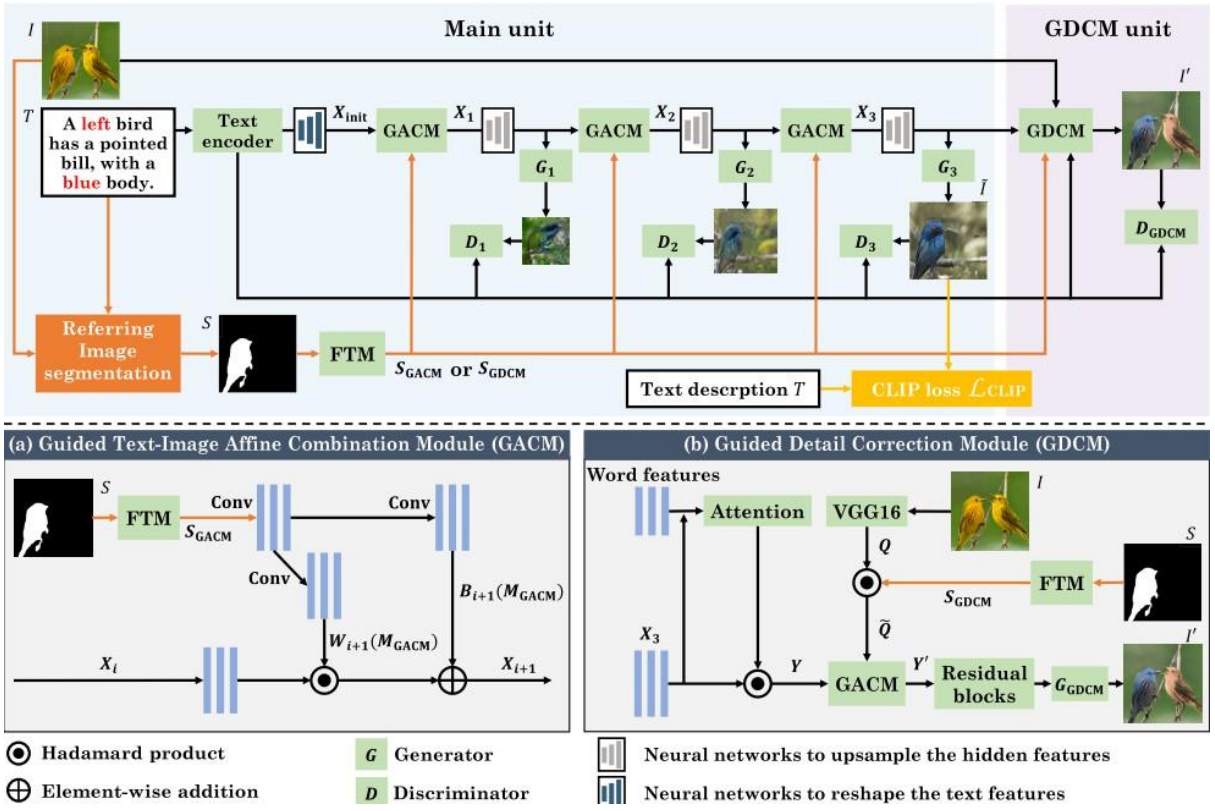

Figure 4.1: Structure of the proposed GAN for text-guided image manipulation.

4.1 APPLICATION OF IMAGE GENERATION

With the emergence of GANs [5], image generation research has made significant advances, and various image generation tasks, including text-to-image synthesis [20], [21], [22], [23], [24] and image-to-image translation [2], [6], [7], [8], [9], [10], have been proposed. Text-to-image synthesis is a cross-modal image generation task conditioned by a natural language text description [20], [21], [22], [23] and scene graphs [24]. Reed et al. [20], for the first time, tackled the task of generating a

64 × 64 natural image from text features based on a deep convolutional GAN architecture [25]. Zhang et al. proposed StackGAN [21] to progressively increase the resolution of generated images by stacking multiple GANs, and Xu et al. [23] and Li et al. [22] proposed an attention-driven generator and word-level discriminator for fine-grained text-to-image synthesis at a word level. In addition, to handle more complex cases, e.g., images with many objects and relationships, previous studies have proposed text- to-image synthesis methods conditioned by scene graphs [24]. Such methods generate a scene layout based on predicted segmentation masks and bounding boxes from input scene graphs and convert the layout into a realistic image using a GAN. By generating the scene layout according to the scene graphs, it is possible to explicitly infer multiple objects and their relationships in the images. The recent studies [26], [27], [28] have reported that image-to-image translation has practical applications such as image quality enhancement, image retrieval, and identity anonymization. With this trend, in the research field of image-to-image translation, various approaches represented by image inpainting [6], [7], image colorization [2], [8], and style transfer [9], [10], have been proposed. These models have demonstrated effectiveness in specific tasks such as filling holes in an image, reflecting the style of well-known artworks in real images, and predicting the color version of black-and-white images. However, some models [6], [7] also require segmentation masks that specify the region of holes in an image to users, and others [2], [8], [9], [10] colorize and translate the entire image regardless of the user's requirements, e.g., manipulation of specific regions or colors.

4.2 REFERRING IMAGE SEGMENTATION

Recently, several referring image segmentation approaches have been proposed to extract the object region in the input image corresponding to a text description as multimodal analyses of image and text representations. For example, Hu et al. [29] handled the referring segmentation task and generated a segmentation mask by directly 42536 concatenating multimodal features extracted using a convolutional

neural network and a long short-term memory (LSTM) network [30]. To analyze word-to-image interactions, a multimodal LSTM [31] sequentially integrates visual and text features in multiple time steps. Li et al. [32] proposed a method for integrating multilevel visual features that can recurrently refine the local details of the segmentation mask. In addition, Huang et al. [17] proposed the state-of- the-art CMPC-Refseg referring image segmentation method, which has demonstrated excellent results. The CMPC-Refseg method progressively perceives correct objects using a crossmodal progressive comprehension (CMPC) and text-guided feature exchange (TGFE) modules. The CMPC module first recognizes all objects assumed from nouns in the text description and then emphasizes the correct object by multimodal graph reasoning. Based on the concept of integrating multilevel visual features, the TGFE module enables the refinement of the results from the CMPC module. In text-guided image manipulation tasks, segmentation masks generated using CMPC-Refseg method may provide promising benefits in terms of focusing on the text-related region.

4.3 TEXT- GUIDED IMAGE MANIPULATION

By applying the concept of controlling image generation with natural language descriptions proposed in the text-to-image synthesis task, text-guided image manipulation [13], [14], [15], [16] achieves more user-friendly image manipulation. Such methods are designed to perform semantic manipulation according to text descriptions and preservation in text unrelated regions. For example, Dong et al. [33] adopted a novel structure primarily based on a GAN and tackled the task of generating the manipulated image conditioned by the given image and text description. In addition, Nam et al. [13] proposed a text adaptive discriminator that monitors the extent to which a text description is reflected in the image at the word level and obtains a generator that can generate finegrained visual attributes. Li et al. [14] adopted a multi-stage architecture of multiple GANs and enabled the generation of 256 × 256 images to produce high-quality manipulated images reflecting text descriptions. Haruyama et al. In addition, [16] focused on differences in representation ability between images and

text descriptions and achieved image manipulation that suppresses background manipulation. In recent years, the integration of vision and language has been emphasized in the computer vision field, and this multimodal analysis can realize

the development of user friendly image manipulation techniques. CLIP [18] is an image classification model pretrained on 400 million image–text pairs. As a result, the model provides generalized text and image representation capabilities without overfitting to a specific dataset and has achieved excellent results in the zero-shot task. Benefiting from the capabilities of this model, text guided image manipulation is expected to maximize the expressive capabilities of text descriptions and effectively reflect the expression in the corresponding manipulated images. While high quality image manipulation can contribute to several fields, how to ensure the trust and credibility of data is an urgent problem to be solved. For this problem, a novel approach [34] for forgery detection based on GANs analyzed the traces that the forgery method may leave on the tampered data and constructed a multiscale forgery trace generation system. Since data manipulation can have dangerous aspects depending on how it is used, the research on text guided image manipulation needs to be cooperatively conducted with the research mentioned above.

CHAPTER 5
PROPOSED IMAGE MANIPULATION METHOD

The model can manipulate input image I according to text description T while preserving text-unrelated regions to generate manipulated image I' . As depicted in Fig. 4.1, the proposed GAN includes a guided text-image affine combination module (GACM) and a guided detail correction module (GDCM). It extend the original ACM and DCM [14] by introducing segmentation guidance to distinguish the image between text-related and unrelated regions such that each module can focus on image manipulation rather than region selection. To generate manipulated image I', the proposed method refines multimodal features by fusing images and text descriptions by passing through two units: the main unit with the multi-stage architecture and the GDCM unit. Here, three generators Gi (i = 1, 2, 3) in the main unit take hidden features X i as inputs and generate images gradually in small-to-large scales, i.e., 64 × 64, 128 × 128, and 256 × 256 pixels. Note that i denotes the order of each stage in the main unit.

5.1 GAN ARCHITECTURE WITH REFERING IMAGE SEGMENTA- TION

Fig. 4.1 depicts the structure of the proposed GAN with the GACM and GDCM. To make the GACM and GDCM function with segmentation guidance, use a segmentation mask S obtained by referring image segmentation. To provide segmentation guidance for the hidden and image features in the GACM and GDCM, the segmentation mask S is transformed into feature maps for compatibility with these features. It use a feature transformation module (FTM) and generate feature maps by resizing the width and height of S and replicating the same along the channel dimension. It expect these

feature maps to provide segmentation guidance for the two modules and instruct the network about the target regions to manipulate or reconstruct. Specifically, in the GACM, segmentation guidance provides the module with information that can identify the region, where the text description T is to be embedded, and the module out- puts features useful for manipulating only the text-related region. Each output feature from the GACM is input into the corresponding generators Gi , which produce temporarily generated images containing text information in only the text- related regions. In addition, in the GDCM, image features of an input image are distinguished into text-related and unrelated regions according to the segmentation guidance. These features are used to prompt the network to reconstruct the contents of the input image I in text-unrelated regions while retaining new attributes acquired in the main unit. Finally, It acquire the manipulated image I from the generator GGDCM . In the following, it describe the segmentation guided modules and their objective functions.

5.2 MANIPULATION OF TEXT-RELATED REGIONS BASED ON GACM

To embed text information in the target region to be manipulated, the GACM performs the process according to the segmentation guidance. As depicted in Fig. 3.1 (a), the GACM takes two inputs, i.e., feature maps SGACM R256×17×17 using the FTM and hidden features X i R32×Hi ×Wi calculated from the previous stage of the GACM or X init R 32×64×64 . To obtain X init , it encode the text description T into global text features using a pretrained bi- directional LSTM [23] and reshape them with neural networks introduced in [14]. Note that H1 and W1 are 64 pixels, H2 and W2 are 128 pixels, and H3 and W3 are 256 pixels. To make the size of the feature maps SGACM equal to X i , Then process the map using two convolutional layers and acquire W i+1 (SGACM) encoding the text-related content and Bi+1 (SGACM) encoding text-unrelated contents. It calculate the hidden features X i+1 R32×Hi+1×Wi+1 , embedding the text features in the region to be manipulated according to the segmentation guidance as follows: X i+1 = X i W i+1 (SGACM) + Bi+1 (SGACM), (1) where denotes the Hadamard product. The main unit of the proposed method uses

a multi-stage architecture comprising GACMs to gradually expand the size of the hidden features X i+1 by upsampling blocks. In the GACM, it let segmentation guidance handle the attention to text related regions and allow the network to focus on gradually acquiring fine grained visual attributes in each stage. As a result, the proposed method realizes a model reflecting the text description T in only the text-related region.

5.3 RECONSTRUCTION OF TEXT-UNRELATED REGIONS BASED ON GDCM

The temporarily manipulated image produced by G3 in the main unit acquires rich visual attributes according to the text description in the text-related region. To further highlight the detailed content and recover text-unrelated regions lost in the main unit, the GDCM performs the process according to the segmentation guidance. Here, as depicted in Fig. 3.1 (b), it first apply the pretrained VGG16 network [35] and acquire the visual features Q R32×256×256 of the input image I from the ReLU layer in the second convolutional block.However, the visual features Q contain too many content details (e.g., color, texture, and edge information), which makes the generator simply reconstruct the input image and potentially lose rich visual attributes in the text- related region acquired by the main unit. To address this problem, it utilize inverted feature maps SGDCM R 32×256×256 by FTM and acquire distinguished features Q˜ R 32×256×256 by concatenating visual features Q with feature maps SGDCM

. By letting segmentation guidance handle the selection of regions that preserve the attributes acquired in the main unit, So that the model can focus exclusively on reconstructing the input image I in text-unrelated regions. To modify the hidden features X 3 R32×256×256 output from the main unit, use features Q˜ and the word features of the text description T . Here, using spatial and channel-wise attention modules [22], [23] (i.e., “Attention” in Fig. 3.1 (b)) based on words, it generate intermediate features Y R 32×256×256 by multiplying the hidden features X 3 with the spatial and channel-wise attention features of the same size as X 3 , following [14].

By reusing the capabilities of the GACM, it associate Y with Q˜ in the same manner as Eq. (1), where Y and Q˜ correspond to X i and SGACM , respectively, and generate features Y R32×256×256 . To refine the features, it pass Y through a residual block [14] containing multiple convolutional layers. The residual block is beneficial for stabilizing the learning of networks with deep layers. Finally, the generator GGDCM in the GDCM generates the final manipulated image I from the refined features.

CHAPTER 6
EXPERIMENTS

Experimental results are detailed, affirming the efficacy of the segmentation guidance process incorporated in the suggested GAN.

6.1 EXPERIMENTAL SETTINGS

To guide segmentation in the proposed method, the referring image segmentation model [17], pretrained on RefCOCO [39], is adopted. This model employs multimodal graph reasoning, achieving high accuracy in generating segmentation masks that extract text-related regions. In the implementation, prior to GAN training, segmentation masks are obtained from all image and text description sets in each dataset using the referring image segmentation model. Throughout GAN training, the parameters of the referring image segmentation model remain unchanged, as the obtained masks are used without running the model each time. The segmentation mask in this network is transformed into feature maps using the FTM

Table 6.1: Detailed statistics for each dataset.

Dataset	Sample	Train:test	Text/image
Oxford-102	8,189	6,149:2,040	10
CUB	11,788	8,855:2,933	10
CUB-based	11,788	8,855:2,933	10
Category	102	200	200

The evaluated image manipulation performance on the Oxford-102 [40] dataset, the Caltech-UCSD-Birds (CUB) [41] dataset, and a new more complicated dataset. Each image in each dataset represents the details of a flower or bird, respectively. Table 5.1 shows the detailed statistics for each dataset. The segmentation guidance process is expected to be powerful in situations involving multiple objects in an image; thus, it

cannot be validated sufficiently on the CUB dataset, consisting of a single bird image. Therefore, to demonstrate the effectiveness of the image manipulation process, Here constructed a CUB-based unique dataset.Applying a semantic segmentation model [42], the regions of birds and backgrounds were automatically separated from all images in the CUB dataset. However, unlike the segmentation model [17] used in the proposed method, the model [42] takes only an image as input and generates the segmentation mask of objects in the image. Then manually selected 20 images to be used as backgrounds by cropping out areas of images in the CUB dataset that did not contain birds. Then,the selected one of the backgrounds created above and randomly aligned the two extracted birds in vertical or horizontal alignment.

The comparison involved evaluating the proposed method against state-of-the-art text-guided image manipulation techniques, including TAGAN [13], ManiGAN [14], Li'20 [15], and Haruyama'21 [16]. These methods, tailored for manipulating single objects in images, have shown promising results on the CUB dataset. Notably, there are no known extensions to these methods capable of manipulating specific objects among multiple objects in an image based on provided text descriptions. Through this comparative analysis, the demonstrated robustness of image manipulation by the proposed method extends to complex images.

During the training process, a methodology consistent with the literature [14] was adhered to, involving the independent training of the two units to stabilize GAN output. Specifically, optimization of parameters in the main unit was conducted first through adversarial training, followed by training the GDCM unit using fixed parameters from the main unit. In ensuring a fair comparison with prior studies that share the same multi-stage architecture [14], [16] as the proposed method, the experimental setup of those methods was adopted. Both the main and GDCM units underwent training for 600 and 100 epochs, respectively, employing the Adam optimizer [43] with a learning rate of 0.0002. It is noteworthy that the proposed and comparison methods were trained on the original CUB dataset, and subsequent image manipulation was performed on

images in the CUB-based unique dataset

Demonstrating the efficacy of segmentation guidance, the application of metrics such as Inception Score (IS) [44], Fréchet Inception Distance (FID) [45], and Kernel Inception Distance (KID) [46] confirms that the quality of manipulated images achieved by the proposed method is comparable to that of the comparison methods. The values of these metrics were computed for the test images on each dataset based on insights from prior studies on text-guided image manipulation. This computation of IS, FID, and KID across datasets and in alignment with previous research underscores that the introduction of segmentation guidance does not compromise the quality of images manipulated using the proposed method.

A subjective experiment on the CUB-based unique dataset was conducted to assess image manipulation precision in complex scenarios. In this experiment, 30 sets of images and text descriptions were randomly selected from the CUB-based unique dataset. Subsequently, words related to the attributes of the chosen text description were randomly altered (e.g., from “yellow” to “white”) to create a new text description representing the user’s desired image manipulation. To generate samples for evaluation by subjects, the proposed and comparison methods were employed to manipulate an image using the selected image and the new text description as inputs.

In the experimental setup, participants were presented with manipulated images generated using the proposed and comparison methods. Their task was to evaluate each image in terms of accuracy and realism based on criteria outlined in [15]. The metrics used for evaluation were as follows: (1) Accuracy, which assesses whether the text-related region is manipulated according to the given text description and whether text-unrelated regions are preserved; (2) Realism, indicating whether the manipulated image appears realistic. Informed consent was obtained from 24 participants, who were then instructed to assign scores ranging from 1 to 5 (1: inaccurate or unreal to 5: accurate or real) to each image according to these two metrics.

6.2 QUANTITATIVE RESULT

Table 6.1 presents the IS, FID, and KID results for the generated images. The IS, FID, and KID values of the proposed method consistently rank as the first or second- best, indicating that the quality of images generated using the proposed method is comparable to or better than those generated by the state-of-the-art comparison methods. These results affirm that the segmentation guidance process employed does not adversely impact the quality of the generated images.

In Table 6.2, accuracy and realism results from the subjective experiment are listed. The presented accuracy and realism values represent the mean scores for each text-guided image manipulation method, calculated based on participant assessments. The accuracy results highlight the effectiveness of the proposed method in complex scenarios, with the proposed method achieving the highest accuracy among all comparison methods. Specifically, the accuracy metric assesses the precision of manipulated images generated from an image with two birds and a text description indicating the manipulation of only a single bird. These results underscore the advantages of the proposed image manipulation method in handling complex situations.

6.3 QUALITATIVE RESULT

Fig. 6.1 compares the visual quality of manipulated images obtained using the proposed method and comparison methods. Specifically, samples (A1), (A2), (B1), and (B2) from datasets featuring a single object showcase that the proposed method achieves enhanced adherence to the text description in the text-related region, with comparable image quality to the comparison methods. It's important to note that TAGAN [13], lacking a multi-stage architecture, exhibits inferior image quality in its manipulated images compared to the other methods. In sample (A1), only the proposed method successfully generates the blue attribute while preserving the white background. For sample (A2), the comparison methods either lose the texture of the stripes or retain the purple attribute, while the proposed method successfully retains

texture and modifies colors. In sample (B1), only the proposed method generates a manipulated image with the red attribute while preserving the background and bird details, including texture. For sample (B2), the precision of image manipulation is compared when the text description contains two attribute words (i.e., blue and white). The proposed method successfully reflects all elements of the text description, such as a white breast and blue tail, in the manipulated image by leveraging the segmentation guidance process for text-related region selection, enabling the GAN to focus on generating new attributes based on the CLIP loss.

For samples (C1) and (C2) involving complex situations from the CUB-based unique dataset, the proposed method effectively concentrates on text-related objects, maintaining substantial image manipulation precision. Trained on the CUB dataset without phrases like “left bird” in the text description, comparison methods are primarily influenced by the word “bird” and manipulate both birds in the image. In contrast, the proposed method introduces segmentation guidance based on the referring image segmentation model to suppress manipulation of text-unrelated regions, selecting regions identified by the text description. Notably, manipulated images produced by the proposed method preserve the original colors of text-unrelated birds, such as a right bird in sample (C1) and a bottom bird in sample (C2). While some comparison methods achieve partial success in suppressing background manipulation, the proposed method appears to most effectively suppress background manipulation in the manipulated images. These results underscore the advantages of the proposed method over comparison state-of-the-art methods [13], [14], [15], [16], particularly in terms of image manipulation precision.

CHAPTER 7
DISCUSSION

Table 7.1: Quantitative results of proposed and comparison methods.

	Oxford-102 [40]			CUB [41]			CUB-based unique dataset		
	IS(↑)	FID(↓)	KID(↓)	IS(↑)	FID(↓)	KID(↓)	IS(↑)	FID(↓)	KID(↓)
TAGAN [13]	2.86	66.48	0.384	3.64	57.20	0.205	3.29	58.75	0.334
ManiGAN [14]	2.90	52.53	0.223	4.58	11.30	0.036	3.60	19.45	0.069
Li'20 [15]	3.81	36.78	0.171	4.64	**9.10**	0.018	3.31	34.96	0.181
Haruyama'21 [16]	3.83	27.52	0.125	4.54	9.47	0.020	3.66	20.22	0.074
Ours	**3.95**	**19.67**	**0.052**	**6.91**	9.24	**0.016**	**3.85**	**16.97**	**0.067**

7.1 EFFECTIVENESS OF SEGMENTATION GUIDANCE

Fig. 7.3 depicts the segmentation masks that are transformed into feature maps SGACM or SGDCM and used as the segmentation guidance in the proposed method's GAN. The segmentation mask successfully extracts only text-related regions and can indicate to the network which regions should be manipulated. With this segmentation guidance, the genera- tors in the proposed method can focus on generating images based on the text description and reconstructing input images without selecting regions. Specifically, as depicted in Fig. 6.3, the temporarily generated images produced by G3 reflect the text descriptions in the regions extracted by the segmentation masks; thus, the GACM works effectively. The image produced by the last generator G3 in the main unit has a rich representation of the visual attributes indicated by the text description, whereas the text-unrelated regions are not properly reconstructed. By utilizing segmentation guidance based on feature maps SGDCM , the GDCM reconstructs the contents of the input image in text-unrelated regions. The difference between the image generated by G3 and the final manipulated image is due to the contribution of the GDCM.

Figure 7.1: Qualitative results of proposed method and four comparison methods [13], [14], [15], [16].

7.2 REPRESENTATION CAPABILITIES OF GENERATORS

In Fig. 7.3, segmentation masks are illustrated, which are transformed into feature maps SGACM or SGDCM and serve as segmentation guidance in the GAN of the proposed method. These segmentation masks adeptly extract only text-related regions, signaling to the network the areas to be manipulated. Empowered by this segmentation guidance, the generators in the proposed method concentrate on generating images based on the text description and reconstructing input images without the need for region selection.

Specifically, as depicted in Fig. 7.3, the interim images produced by G3 reflect the

text descriptions in the regions extracted by the segmentation masks, indicating the effectiveness of GACM. The final image generated by the last generator G3 in the main unit exhibits a rich representation of the visual attributes indicated by the text description, while text-unrelated regions are not reconstructed accurately. Leveraging segmentation guidance based on feature maps SGDCM, the GDCM unit reconstructs the content of the input image in text-unrelated regions. The disparity between the image generated by G3 and the final manipulated image is attributed to the contribution of the GDCM.

In the section on "Representation capabilities of generators," the generated images corresponding to Gi are presented in the three columns on the right side of Fig. 6.3, showcasing images at resolutions of 64 × 64, 128 × 128, and 256 × 256 using the multi-stage architecture. These images, as demonstrated in the figure, exhibit increasingly sophisticated visual representations as the resolution increases. Guided by segmentation based on feature maps SGACM, these images accurately capture the visual attributes indicated by the text description in each text-related region. The generators showcase high representational capability through the introduction of the CLIP loss, enabling them to reflect fine-grained words in the generated images. Furthermore, aided by segmentation guidance based on feature maps SGDCM, the generator GGDCM in the GDCM unit adeptly reconstructs the content of the input image without negating the attributes generated by G3.

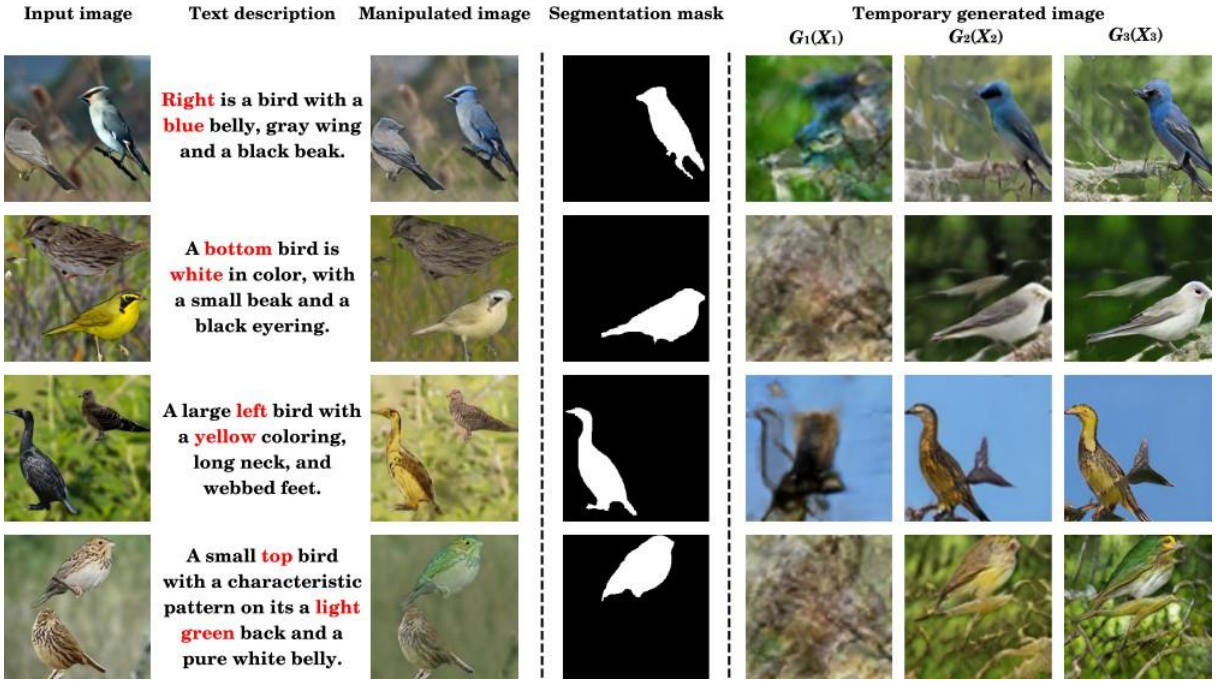

Figure 7.2: Detailed visual analyses of the proposed method.

Table 7.2: Accuracy and realism of the results

Model	Accuracy(↑)	Realism(↑)
TAGAN [13]	2.74	3.20
ManiGAN [14]	2.84	3.21
Li'20 [15]	2.83	3.68
Haruyama'21 [16]	2.83	3.25
Ours	**4.12**	**3.47**

7.3 EXAMPLES OF FAILED IMAGE MANIPULATION

Although the proposed method enhances image manipulation precision through the introduction of segmentation guidance, there are instances where images could not be manipulated effectively. The segmentation guidance assumes a crucial role in refining image manipulation precision, and the results outputted by the pre-trained referring image segmentation model exert a significant influence on the proposed method's performance. Illustrated in Fig. 7.3 is an example where the proposed method falls short in achieving effective image manipulation on the CUB-based unique dataset. In this instance, the segmentation mask fails to extract a right bird, leading to a manipulated image region that does not align with the text description. Furthermore, the incorrectly extracted region by the segmentation mask incompletely covers the object, resulting in an unnatural interruption in the region of a left bird. Consequently, future efforts will focus on exploring methods to enhance the matching between images and texts, facilitating the extraction of small objects and further refining the performance of the segmentation guidance process. Although a segmentation mask can be directly applied to an image to distinguish the regions to be reconstructed and manipulated, it attempt to reduce disconnectedness by applying a segmentation mask to image features. However, few artifacts certainly remain in the manipulated image generated using the proposed method. In future studies, it will consider introducing a module to mitigate artifacts at the end of the network as well as using a segmentation mask of the state of the score map prior to the binarization.

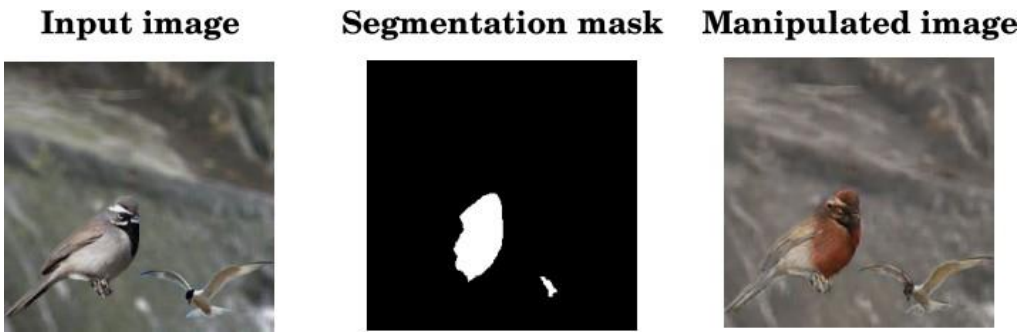

Figure 7.3: Examples of failed image manipulation on the CUB-based unique dataset.

CHAPTER 8

TEXT-TO-IMAGE GENERATION METHODS

This section provides an overview of relevant studies on text-to-image generative models. Due to the diversity of the generative models and the vast amount of associated literature, this study narrows its focus to the two cutting edge types of deep learning generative models: **GANs and diffusion models.**

8.1 TEXT-TO-IMAGE GENERATION USING GANS

- **StackGAN**: A two-stage GAN that first generates low-resolution images (Stage-I) and refines them into high-resolution images (Stage-II).
- **AttnGAN**: Uses an attention mechanism to focus on key textual features, generating fine-grained image details.
- **MirrorGAN**: A model that ensures semantic alignment by reconstructing text from generated images, maintaining consistency.
- **DM-GAN**: Employs a dynamic memory module to enhance unclear image regions based on the text description.
- **ManiGAN**: Enables image manipulation by modifying specific visual attributes based on text inputs while preserving irrelevant content.

Figure 8.1: Random image samples on the CUB dataset, generated by

DM-GAN, Attn-GAN, StackGAN, and GAN-INT-CLS.

8.2 TEXT-TO-IMAGE GENERATION USING DIFFUSION MODELS

These models iteratively add noise to images and learn to reverse this process, gradually generating a detailed image from text prompts. Diffusion models, like **Imagen** and **Stable Diffusion**, have proven to be highly effective in generating photorealistic images with fine details and better alignment to the provided text compared to GANs. **Diffusion Models**, represents an advanced approach to text-to-image generation. Unlike GANs, which use adversarial networks to generate images, diffusion models rely on a step-by-step denoising process that gradually creates a detailed image from noise. Diffusion models are based on probabilistic processes that add noise to the data (in this case, images) in incremental steps, and then the model learns to reverse this noise process to generate clear, structured images. These models have recently shown superior performance in tasks like image synthesis, surpassing GANs in generating high-quality and realistic images.

Key Features:

1. **Noise Injection and Removal**: Diffusion models work in two phases: forward and reverse. In the **forward diffusion** phase, noise is gradually added to the image data through a Markov chain until the image is mostly noise. Then, in the **reverse diffusion** phase, the model learns to progressively remove the noise, reconstructing a realistic image from what started as random noise.
2. **Two-Step Learning Process**:
    - The forward process is simple: random noise is progressively added to the input image.
    - The reverse process is more complex and requires learning how to effectively remove noise, step by step, to reconstruct an image that matches the text input.

Applications and Models:

Several recent models use diffusion processes for text-to-image generation:

- **VQ-Diffusion**: This model uses a combination of Vector Quantized Variational Autoencoders (VQ-VAE) and diffusion to model latent space during the generation process. It produces images based on text prompts by denoising from a latent code.
- **Blended Diffusion**: A technique that allows local image modifications based on text prompts, where the model is guided to modify specific regions of an image.
- **GLIDE**: A text-conditional diffusion model that can generate realistic images from text descriptions. It uses CLIP (Contrastive Language-Image Pretraining) guidance to align generated images with the text, making it highly effective in terms of text-image coherence.
- **Imagen**: Uses a diffusion process to create high-resolution, detailed images from text prompts. It employs a text encoder and diffusion models to transform textual inputs into images, focusing on ensuring high fidelity and alignment between the generated images and the text.

Advantages:

- **Better Image Quality**: Diffusion models produce sharper, more detailed images compared to GANs, especially when generating high-resolution images.
- **Text-Alignment**: They often perform better in ensuring that the generated image accurately reflects the text description. Models like **GLIDE** and **Imagen** use mechanisms such as CLIP to ensure the image matches the text input effectively.

Challenges:

- **Computational Complexity**: Diffusion models are computationally intensive due to the multi-step process of adding and removing noise, which requires significant processing power and time.
- **Scalability**: Handling large datasets and producing high-resolution images can be more resource-demanding compared to GANs.

DALL-E 2 is a prime example of a text-to-image generation model based on diffusion, marking a significant advancement in the field. Unlike GANs, DALL-E 2 uses a two-stage approach that leverages diffusion for improved quality and alignment with text.

8.2.1 How DALL-E 2 Works:

1. **Text Encoding**: DALL-E 2 starts by encoding the text description into an image embedding using a model like CLIP (Contrastive Language-Image Pretraining). This embedding captures the semantic information of the text, which will guide the image generation.
2. **Diffusion Process**: After obtaining the text embedding, DALL-E 2 uses a **diffusion decoder** to generate an image from noise. The diffusion process works by reversing the noise-adding steps, gradually refining the random noise into a coherent image that aligns with the provided text. This process allows DALL-E 2 to create highly detailed and creative visuals that closely match the complexity and nuances of the text description.

Figure 8.2: Samples generated by DALL-E 2 given the prompt: ‘‘a bowl of soup that is a portal to another dimension as digital art’’.

CHAPTER 9
EVALUATION METRICS

The majority of current metrics evaluate a model's quality by considering two main factors: the quality of the images it produces and the alignment between text and images. Fréchet Inception Distance (FID) [96] and Inception Score (IS) [97] are commonly used metrics for appraising the image quality of a model. These metrics were initially developed for traditional GAN tasks focused on assessing image quality. To evaluate text-image alignment, the R-precision [47] metric is widely employed.

Figure 9.1: Random image samples on MS-COCO, generated by DALL-E, GLIDE, and DALL-E 2.

For more in-depth details, we refer to [98]. Moreover, the Clip Score [99] is used in evaluating common sense and mentioned objects, while Human Evaluation offers a comprehensive insight into multiple aspects of image generation. In the following a detailed description of each metric.

CHAPTER 10

CHALLENGES AND LIMITATIONS

Although there has been significant progress made in the area of creating visual representations of textual descriptions, there are still some challenges and limitations that will be discussed below.

- **Computational Complexity**: Diffusion models require significant computational resources due to their multi-step processes of adding and removing noise, making them resource-intensive and slower to train and run.
- **Language Support**: Most models are trained on English text due to the availability of datasets, and extending them to other languages, particularly complex ones like Arabic, remains a challenge.
- **Open-Source Availability**: While some models like **Stable Diffusion** are open-source, others like **DALL-E 2** are not fully publicly available, limiting access for research and development.
- **Ethical Concerns**: There is potential for misuse of these models, such as generating fake or misleading content. Additionally, biases present in the training data can lead to prejudiced or stereotypical image outputs.

CHAPTER 11

CONCLUSION

The field of text-to-image synthesis has made significant progress in recent years. The development of GANs and diffusion models has paved the way for more advanced and realistic image generation from textual descriptions. These models have demonstrated an outstanding ability to generate high-quality images across a wide range of domains and datasets. This study offers a comprehensive review of the existing literature on text-to-image generative models, summarizing the historical development, popular datasets, key methods, commonly used evaluation metrics, and challenges faced in this field. Despite these challenges, the potential of text-to-image generation in expanding creative horizons and enhancing AI systems is undeniable. The ability to generate realistic and diverse images from textual inputs opens up new possibilities in various fields, including art, design, advertising, and others. Therefore, researchers and practitioners should continue to explore and refine text-to image generative models.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2414–2423.

[2] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 649–666.

[3] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, “Generative visual manipulation on the natural image manifold,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 597–613.

[4] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014, pp. 2672–2680.

[6] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2536–2544.

[7] V. Lempitsky, A. Vedaldi, and D. Ulyanov, “Deep image prior,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 9446–9454.

[8] Z. Cheng, Q. Yang, and B. Sheng, “Deep colorization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 415–423

[9] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” 2015, arXiv:1508.06576.

[10] A. Madaan, A. Setlur, T. Parekh, B. Poczos, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, and S. Prabhumoye, “Politeness transfer: A tag and generate approach,” 2020, arXiv:2004.14257.