

# Research on Deep Learning for Artificially Controllable Image Synthesis

ZHANG, Zhiqiang

---

(開始ページ / Start Page)

1

(終了ページ / End Page)

114

(発行年 / Year)

2023-09-15

(学位授与番号 / Degree Number)

32675甲第586号

(学位授与年月日 / Date of Granted)

2023-09-15

(学位名 / Degree Name)

博士(工学)

(学位授与機関 / Degree Grantor)

法政大学 (Hosei University)

(URL)

<https://doi.org/10.15002/00030035>

# Research on Deep Learning for Artificially Controllable Image Synthesis

Zhiqiang Zhang

Doctoral Dissertation Reviewed by  
Hosei University

Research on Deep Learning for Artificially  
Controllable Image Synthesis

Zhiqiang Zhang

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>List of Publications</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Introduction . . . . .	1
1.2 Research Objective and Specific Scheme . . . . .	3
1.3 Overview of the Proposed Methods . . . . .	4
1.4 Main Contributions . . . . .	5
1.5 Dissertation Outline . . . . .	6
<b>2 High Quality Oriented Image Synthesis Methods</b>	<b>8</b>
2.1 Introduction . . . . .	8
2.2 Related Works . . . . .	9
2.3 Preliminaries . . . . .	10
2.3.1 Generative Adversarial Networks . . . . .	10
2.3.2 Image-Text Matching . . . . .	11
2.3.3 Introduction to Experimental Datasets . . . . .	12
2.3.4 Introduction to Evaluation Methods . . . . .	13
2.4 Method 1 — DrawGAN: Text to Image Synthesis with Drawing Generative Adversarial Networks . . . . .	14
2.4.1 Network Structure . . . . .	15
2.4.2 Loss Function . . . . .	18
2.4.3 Implementation Details . . . . .	20
2.4.4 Experiments . . . . .	21

2.5	Method 2 — Text-to-Image Synthesis: Starting Composite from the Foreground Content . . . . .	27
2.5.1	Network Structure . . . . .	28
2.5.2	Loss Function . . . . .	32
2.5.3	Implementation Details . . . . .	33
2.5.4	Experiments . . . . .	33
2.6	Method 3 — Text to Image Synthesis with Erudite Generative Adversarial Networks . . . . .	47
2.6.1	Network Structure . . . . .	48
2.6.2	Loss Function . . . . .	51
2.6.3	Implementation Details . . . . .	51
2.6.4	Experiments . . . . .	52
2.7	Internal Comparison of Proposed Methods . . . . .	55
2.7.1	Qualitative Comparison . . . . .	55
2.7.2	Quantitative Comparison . . . . .	56
2.8	Chapter Conclusions . . . . .	57
<b>3</b>	<b>High Controllability Oriented Image Synthesis Methods</b>	<b>58</b>
3.1	Introduction . . . . .	58
3.2	Related Works . . . . .	59
3.3	Method 1 — Customizable GAN: A Method for Image Synthesis of Human Controllable . . . . .	61
3.3.1	Network Sturcture . . . . .	62
3.3.2	Loss Function . . . . .	66
3.3.3	Implementation Details . . . . .	67
3.3.4	Experiments . . . . .	67
3.4	Method 2 — TCGIS: Text and Contour Guided artificially controllable Image Synthesis . . . . .	77

3.4.1	Network Structure . . . . .	78
3.4.2	Loss Function . . . . .	81
3.4.3	Implementation Details . . . . .	82
3.4.4	Experiments . . . . .	82
3.5	Internal Comparison of Proposed Methods . . . . .	85
3.6	Chapter Conclusion . . . . .	86
<b>4</b>	<b>High Practicality Oriented Image Synthesis Methods</b>	<b>87</b>
4.1	Introduction . . . . .	87
4.2	Related Works . . . . .	88
4.3	Text-guided Image Manipulation based on Sentence-aware and Word-aware Network . . . . .	89
4.3.1	Network Structure . . . . .	89
4.3.2	Loss Function . . . . .	93
4.3.3	Implementation Details . . . . .	93
4.3.4	Experiments . . . . .	93
4.4	Image Synthesis Methods with High Practicality . . . . .	98
4.4.1	Text-guided Image Synthesis and Manipulation . . . . .	98
4.4.2	Text-guided Controllable Image Synthesis and Manipulation . .	100
4.5	Chapter Conclusion . . . . .	101
<b>5</b>	<b>Conclusion</b>	<b>102</b>
	<b>List of Abbreviations</b>	<b>113</b>

# Abstract

Image synthesis has been one of the most important topics in computer vision research. In recent years, with the rise of artificial intelligence, research in the image synthesis field has made many breakthroughs. Especially the introduction of deep learning has made the field advance by leaps and bounds, the most notable of which is generative adversarial networks (GAN). However, since the input received by GAN is randomly generated Gaussian distribution noise, this makes the image synthesis process artificially uncontrollable, resulting in poor practicability of the whole method. In order to solve this problem, in recent years, text-to-image synthesis (T2I) has been proposed and gained extensive attention. T2I generates corresponding images through simple, intuitive, and easy-to-enter text information. Due to the text information conforming to people’s input habits, this method can realize the artificially controllable image synthesis effect to a certain extent. Nevertheless, T2I still faces the following challenges: 1) The **quality** of image synthesis needs to be further improved. Quality is reflected in the realism of the synthesized content. The current T2I methods still produce poorly realistic image results, so the overall quality needs to be improved. 2) The **controllability** of the image synthesis process needs to be further improved. Controllability is reflected in the control degree over the synthetic content. By using text information, the current T2I methods can only control the basic content of the synthesized object but cannot control the shape, size, and position information of the synthesized object, so the overall controllability is insufficient. 3) The **overall practicability** of the image synthesis method needs to be further improved. Practicality is reflected in the application degree of the synthetic method. The current T2I methods can synthesize the corresponding image based on the input text, but it cannot continue to input new text to modify the content of the generated image, which makes the overall practicability of the current method insufficient.

Facing the above challenges, this research is committed to realizing the artificial

controllable image synthesis method in the whole process, which is divided into three parts: 1) Developing better T2I methods to achieve higher-quality image results; 2) Developing controllable image synthesis methods to improve the controllability of the synthesis process; 3) Based on the first and second parts, introduce the image manipulation method to achieve controllable image synthesis and manipulation with high quality, thereby further improving the practicability of the synthesis method.

In the first part, we propose three methods to achieve higher-quality image synthesis results. The basic idea of the first method is to synthesize simple contour information at first, and then synthesize foreground content, and then synthesize the final image result; The basic idea of the second method is first to synthesize the foreground content based on the text information, and then synthesize the final image result based on the synthesized foreground and the input text information. The basic idea of the third method is to introduce additional image discrimination types into the GAN’s discriminator to improve its discriminative ability, and then better discriminant is fed back to the generator to improve the quality of the synthesis result. Extensive experimental results have proved that the three methods proposed above all achieve higher-quality image synthesis results.

In the second part, we propose a more controllable approach to image synthesis. Specifically, text description and simple contour information are used to synthesize corresponding image results, where text description can control the synthesis content, contour information can control the basic shape and position of the synthesized object, and both text and contour information can be manually input. Therefore, using text and contour information to synthesize the corresponding image has better controllability. In this idea, we proposed two network structures. The first is to simply combine text and contour information, and then achieve corresponding image synthesis through the residual and upsampling operations. This method preliminarily achieves the effect of controllable image synthesis, but the overall quality of the synthesis is mediocre.

Therefore, the second network structure is proposed. The core of the second structure is to introduce an attention mechanism to fine-tune the synthesis result to improve the quality of image synthesis. Experimental results demonstrate that our proposed second network structure achieves better controllable and higher-quality image synthesis results.

In the third part, the core is to introduce the image manipulation method on top of the first and second parts to form the high practicality image synthesis methods. Therefore, we first propose a text-guided image manipulation (TGIM) method. The basic idea of this method is to design a sentence-aware and word-aware network structure to achieve better image manipulation effects. After that, by fusing the proposed text-guided image manipulation method and the image synthesis methods proposed in the first and second parts, we finally achieve the text-guided image synthesis and manipulation and text-guided controllable image synthesis and manipulation methods. The former allows the input text manually to synthesize the corresponding image, and then continue to input new text manually to modify the content of the synthesized image. The latter allows input text and simple contour information artificially to synthesize the corresponding image result, and then can artificially continue to input new text to modify the content of the previously synthesized image. From the experimental results, these two methods have achieved good practicability. In contrast, the second approach has better human controllability and practicability because it can control the basic content of image synthesis and the shape and position information of the synthetic object at the beginning.

# Acknowledgements

This dissertation presents the research conducted during the years of 2020-2023 at the Graduate School of Science and Engineering, Hosei University. During the time of carrying out this research, I have received great support from the beloved people I would like to thank.

Foremost, I would like to express my sincere gratitude to Professor Jinjia Zhou for her enthusiasm supervision, and encouragement from day one. I am very grateful and proud of the knowledge and skills I have learned from her. There is no doubt that her scientific feedback and recommendations have brought my research to a higher level.

I would like to thank Professor Koichi Ogawa and Professor Hitoshi Iyatomi for the feedback they provided as part of my dissertation committee. Your suggestions brought my dissertation to a higher level than it would have been.

I would like to thank Professor Kazuo Yana. When I was a master student, he came to our school to introduce Hosei University, which gave me the opportunity to learn about Hosei University and ultimately led me to pursue a doctoral degree at Hosei University. I would also thank the Graduate School Section of Hosei University, Koganei campus for providing quick answers and kind assistance.

I would also send my special thanks to my laboratory members and alumni: Chen Fu, Xin Cheng, Jiayao Xu, Jian Yang, Alaa Zein, Keren He, Sato Mirai, Morita Ryugo, Man M. Ho.

Last but not least, I thank my wonderful mom, dad, and brother for their unconditional and endless love. You are always the reason I keep going on every day.

# List of Publications

## Journal papers

- [1] **Zhiqiang Zhang**, Jinjia Zhou, Wenxin Yu, and Ning Jiang, “Text-to-image synthesis: Starting composite from the foreground content,” in *Information Sciences*, vol. 607, pp. 1265-1285, 2022, DOI: 10.1016/j.ins.2022.06.044. **(IF: 8.233)**
- [2] **Zhiqiang Zhang**, Chen Fu, Wei Weng, and Jinjia Zhou, “Text-Guided Customizable Image Synthesis and Manipulation,” in *Applied Sciences*, vol. 12, no. 20, pp. 10645, 2022, DOI: 10.3390/app122010645. **(IF: 2.838)**
- [3] **Zhiqiang Zhang**, Wenxin Yu, Jinjia Zhou, Xuewen Zhang, Ning Jiang, Gang He, and Zhuo Yang, “Customizable GAN: A Method for Image Synthesis of Human Controllable,” in *IEEE Access*, vol. 8, pp. 108004-108017, 2020, DOI: 10.1109/ACCESS.2020.3001070. **(IF: 3.476)**
- [4] Wenxin Yu, Xuewen Zhang, Yunye Zhang, **Zhiqiang Zhang**, and Jinjia Zhou, “Blind Image Quality Assessment for a Single Image From Text-to-Image Synthesis,” in *IEEE Access*, vol. 9, pp. 94656-94667, 2021, DOI: 10.1109/ACCESS.2021.3094048. **(IF: 3.476)**
- [5] Chen Fu, Heming Sun, **Zhiqiang Zhang**, and Jinjia Zhou, “A Highly Pipelined and Highly Parallel VLSI Architecture of CABAC Encoder for UHD TV Applications,” in *Sensors*, vol. 23, no. 9, pp. 4293, 2023, DOI: 10.3390/s23094293. **(IF: 3.847)**

## Conference papers

- [1] **Zhiqiang Zhang**, Chen Fu, Man M. Ho, Jinjia Zhou, Ning Jiang, and Wenxin Yu, “Text-Guided Image Manipulation Based on Sentence-Aware and Word-Aware

- Network,” IEEE International Conference on Multimedia and Expo (ICME), 2022, pp. 1-6, DOI: 10.1109/ICME52920.2022.9859585.
- [2] **Zhiqiang Zhang**, Jinjia Zhou, Wenxin Yu, and Ning Jiang, “Drawgan: Text to Image Synthesis with Drawing Generative Adversarial Networks,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 4195-4199, DOI: 10.1109/ICASSP39728.2021.9414166.
- [3] **Zhiqiang Zhang**, Wenxin Yu, Ning Jiang, and Jinjia Zhou, “Text To Image Synthesis With Erudite Generative Adversarial Networks,” IEEE International Conference on Image Processing (ICIP), 2021, pp. 2438-2442, DOI: 10.1109/ICIP42928.2021.9506487.
- [4] **Zhiqiang Zhang**, Chen Fu, Jinjia Zhou, Wenxin Yu, and Ning Jiang, “Text to Image Synthesis based on Multi - Perspective Fusions,” International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1-8, DOI: 10.1109/IJCNN52387.2021.9533925.
- [5] **Zhiqiang Zhang**, Wenxin Yu, Jinjia Zhou, Xuewen Zhang, Jialiang Tang, Siyuan Li, Ning Jiang, Gang He, Gang He, and Zhuo Yang, “Customizable GAN: customizable image synthesis based on adversarial learning,” International Conference on Neural Information Processing (ICONIP), 2020, pp. 336-344, DOI: 10.1007/978-3-030-63820-7\_38.
- [6] Ryugo Morita, **Zhiqiang Zhang**, Man M. Ho, and Jinjia Zhou, “Interactive Image Manipulation with Complex Text Instructions,” IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1053-1062, DOI: 10.1109/WACV56688.2023.00111.
- [7] Man M.Ho, Heming Sun, **Zhiqiang Zhang**, Jinjia Zhou, “On Pre-chewing Compression Degradation for Learned Video Compression,” IEEE Visual Communica-

- tions and Image Processing (VCIP), 2022, pp. 1-5, DOI: 10.1109/VCIP56404.2022.1-0008873.
- [8] Shinko Hayashi, **Zhiqiang Zhang**, and Jinja Zhou, “HA Recurrent Point Clouds Selection Method for 3D Dense Captioning,” International Conference on Neural Information Processing (ICONIP), 2022, pp. 263-274, DOI: 10.1007/978-3-031-30111-7\_23.
- [9] Shinko Hayashi, **Zhiqiang Zhang**, and Jinjia Zhou, “Improving descriptive deficiencies with a Random Selection Loop for 3D Dense Captioning based on Point Clouds,” IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022.
- [10] Jiayao Xu, Chen Fu, **Zhiqiang Zhang**, Jinjia Zhou, “Real-time FPGA Design for OMP Targeting 8K Image Reconstruction,” International Conference on Multimedia Modeling (MMM), 2022, pp. 518-529, DOI: 10.1007/978-3-030-98358-1\_41.
- [11] Chen Fu, Heming Sun, Jiayao Xu, **Zhiqiang Zhang**, Jinjia Zhou, “Optimizing CABAC architecture with prediction based context model prefetching,” IEEE International Workshop on Multimedia Signal Processing (MMSP), 2022, pp. 1-6, DOI: 10.1109/MMSP55362.2022.9949499.
- [12] Ning Jiang, Jialiang Tang, **Zhiqiang Zhang**, Wenxin Yu, Xin Deng, and Jinjia Zhou, “Search-and-Train: Two-Stage Model Compression and Acceleration,” International Conference on Neural Information Processing (ICONIP), 2020, pp. 642-649, DOI: 10.1007/978-3-030-63823-8\_73.
- [13] Siyuan Li, Lu Lu, **Zhiqiang Zhang**, Xin Cheng, Kepeng Xu, Wenxin Yu, Gang He, Jinjia Zhou, Zhuo Yang, “Interactive Separation Network for Image Inpainting,” IEEE International Conference on Image Processing (ICIP), 2020, pp. 1008-1012, DOI: 10.1109/ICIP40778.2020.9191263.

Among the above-mentioned published papers, those published as the first author (Zhiqiang Zhang) are related to this doctoral dissertation research.

# List of Figures

1.1	Research scheme. . . . .	3
1.2	Method overview. . . . .	4
2.1	The exhibit of preliminary results of DrawGAN. . . . .	15
2.2	The generation structure of DrawGAN. . . . .	17
2.3	Processed foreground image results. . . . .	21
2.4	The comparison bird results of DrawGAN. . . . .	22
2.5	The comparison flower results of DrawGAN. . . . .	22
2.6	The comparison complex results of DrawGAN. . . . .	23
2.7	The exhibit of preliminary results of proposed method 2. . . . .	27
2.8	The generation structure in the first stage. . . . .	28
2.9	The generation structure in the second and third stage. . . . .	29
2.10	First cause for the three stages' results. . . . .	34
2.11	Second cause for the three stages' results. . . . .	35
2.12	Fine tuning results for birds in the first cause. . . . .	36
2.13	Fine tuning results for birds in the second cause. . . . .	36
2.14	Fine tuning results for flowers in the first cause. . . . .	36
2.15	Fine tuning results for flowers in the second cause. . . . .	37
2.16	Fine tuning results for complex image in the first cause. . . . .	37
2.17	Fine tuning results for complex image in the second cause. . . . .	37
2.18	Comparison results in the CUB dataset. . . . .	38
2.19	Comparison results in the Oxford-102 dataset. . . . .	38
2.20	Comparison results in the MS COCO dataset. . . . .	39
2.21	Three stage comparison results in the CUB dataset. . . . .	41
2.22	Three stage comparison results in the Oxford-102 dataset. . . . .	42
2.23	Three stage comparison results in the MS COCO dataset. . . . .	43

2.24	The exhibit of preliminary results of EruditeGAN. . . . .	48
2.25	The generator and discriminator structure of EruditeGAN. . . . .	49
2.26	Processed segmentation image results. . . . .	52
2.27	The comparison results of EruditeGAN. . . . .	53
2.28	The qualitative comparison results among our proposed T2I methods .	55
3.1	The exhibit of preliminary results of CustomizableGAN. . . . .	62
3.2	The generator structure of CustomizableGAN. . . . .	63
3.3	The discriminator structure of CustomizableGAN. . . . .	64
3.4	The comparison between our method and the existing T2I method . . .	68
3.5	The first group comparison between our method and GAWWN. . . . .	68
3.6	The second group comparison between our method and GAWWN. . . .	69
3.7	The comparison bird results of our method without VGG and with VGG16, with VGG19. . . . .	70
3.8	The comparison flower results of our method without VGG and with VGG16, with VGG19. . . . .	71
3.9	Artificial controllable image synthesis results on the CUB dataset. . . .	75
3.10	Artificial controllable image synthesis results on the Oxford-102 dataset.	76
3.11	Controllable image synthesis results on the MS COCO dataset. . . . .	77
3.12	The exhibit of preliminary results of TCGIS. . . . .	78
3.13	The generation structure of TCGIS. . . . .	79
3.14	The comparison between our method and the existing T2I methods. . .	83
3.15	The comparison between our method and CustomizableGAN. . . . .	83
4.1	The exhibit of preliminary results of our proposed TGIM method. . . .	89
4.2	The network structure of our proposed TGIM method. . . . .	90
4.3	The SWN structure of our proposed TGIM method. . . . .	90
4.4	The ACM&DCM structure of our proposed TGIM method. . . . .	91
4.5	The comparison bird results of our proposed TGIM method. . . . .	94

4.6	The comparison flower results of our proposed TGIM method. . . . .	94
4.7	The comparison complex results of our proposed TGIM method. . . . .	95
4.8	The results of Text-guided Image Synthesis and Manipulation. . . . .	99
4.9	The results of Text-guided Controllable Image Synthesis and Manipulation. . . . .	100

# List of Tables

2.1	The basic information for the CUB, Oxford-102, and MS COCO datasets.	13
2.2	The IS and FID comparison results of our DrawGAN and the existing methods on the CUB dataset. . . . .	24
2.3	The IS and FID comparison results of our DrawGAN and the existing methods on the Oxford-102 flower dataset. . . . .	24
2.4	The IS and FID comparison results of our DrawGAN and the existing methods on the MS COCO dataset. . . . .	24
2.5	The R-precision comparison results of AttnGAN, DMGAN, and our DrawGAN. . . . .	25
2.6	The ablation experiment results on the CUB dataset. . . . .	26
2.7	The ablation experiment results on the Oxford-102 dataset. . . . .	26
2.8	The ablation experiment results on the MS COCO dataset. . . . .	26
2.9	The IS and FID comparison results of our method and the existing methods on the CUB dataset. . . . .	43
2.10	The IS and FID comparison results of our method and the existing methods on the Oxford-102 dataset. . . . .	44
2.11	The IS and FID comparison results of our method and the existing methods on the MS COCO dataset. . . . .	45
2.12	The R-precision comparison results of AttnGAN, DMGAN, DrawGAN, and our method. . . . .	45
2.13	The ablation experiment analysis of the three-generation stages corresponding to the <i>fore_1</i> . . . . .	45
2.14	The ablation experiment analysis of the three-generation stages corresponding to the <i>fore_1&amp;2</i> . . . . .	46
2.15	The comparison experiment results between <i>fore_1</i> and <i>fore_1&amp;2</i> on the CUB, Oxford-102, and MS COCO dataset. . . . .	47

2.16	The comparison results of IS and FID on the CUB, Oxford-102 flower, and MS COCO datasets between our method and existing state-of-the-art methods. . . . .	53
2.17	The R-precision comparison results of AttnGAN, DMGAN, and our method. . . . .	53
2.18	The comparison results of IS and FID on the CUB, Oxford-102 flower, and MS COCO datasets between our method and existing state-of-the-art methods. . . . .	54
2.19	The quantitative comparison results of IS and FID on the CUB, Oxford-102 flower, and MS COCO datasets among our proposed methods. . . .	56
3.1	The results of quantitative comparison between our three methods and GAWWN. . . . .	72
3.2	The quantitative comparison results between our method with GAWWN in CUB dataset. . . . .	73
3.3	The internal quantitative comparison results of our methods in CUB dataset. . . . .	73
3.4	The internal quantitative comparison results of our methods in Oxford-102 flower dataset. . . . .	74
3.5	The MS-SSIM, SSIM, and FSIM comparison results of our method and existing T2I methods. . . . .	84
3.6	The IS and FID comparison results of our method, existing T2I methods, and CustomizableGAN . . . . .	85
4.1	The quantitative comparison results of our method and other existing methods on the CUB dataset. . . . .	96
4.2	The quantitative comparison results of our method and other existing methods on the Oxford-102 dataset. . . . .	96

4.3	The quantitative comparison results of our method and other existing methods on the MS COCO dataset. . . . .	97
4.4	The ablation comparison results of our proposed SWN. . . . .	98

# Chapter 1

## Introduction

### 1.1 Problem Introduction

In computer vision, image synthesis has always been a widely concerned research field. Due to the rapid development of deep learning, many breakthroughs have been made in the field of image synthesis, especially the recent introduction of Generative Adversarial Networks (GAN) [1], which has obtained many encouraging results in this field. Nevertheless, the traditional GAN only uses the noise vectors obtained by Gaussian or uniform distribution for image synthesis, making the image synthesis category of the training model entirely dependent on training datasets. For example, when using the bird image dataset for training, the corresponding training model can synthesize the bird images, and when using the flower image dataset, the model can generate the flower images. Therefore, using only noise vectors as input leads to the trained model not having good flexibility and controllability.

In order to solve this problem, conditional Generative Adversarial Networks (CGAN) [2] is proposed. CGAN introduces conditional variables into the input to achieve reasonable control of composite image types. For example, using the image category of birds or flowers in training, the trained model can generate the corresponding bird or flower images. CGAN achieves the flexibility and controllability required in image synthesis to a certain extent. However, CGAN can only determine the specific type of the synthesis image through the category label but cannot determine the specific content of the composite image. For example, the category label ‘bird’ is input, the model can synthesize a bird image, but the specific color, size, and other information of the bird can not be determined. To further improve

the overall flexibility and control of image synthesis, text-to-image synthesis (T2I) research has been proposed. The T2I can synthesize the corresponding image results by inputting the text description. The text includes more basic information so that it can determine the specific content of the composite image. For example, input a text description “this grey bird has an impressive wingspan, a grey bill, and a white stripe that surrounds the feathers near the bill”, and the model can synthesize image results related to the semantic information of the input text. Hence text-to-image synthesis research has better flexibility and controllability. Besides, it has tremendous potential applications, such as computer-aided design, art generation, image editing, video games, and so on. Based on the above reasons, the T2I research field has received extensive attention at present.

At present, many promising results have been achieved in T2I research. Reed et al. [3] first proposed an end-to-end GAN structure and realized the image synthesis from the text description. However, the overall clarity and authenticity of synthesized images are poor in this work. To further improve the quality of the synthesized image, many improved methods have been proposed later. Zhang et al. [4][5] proposed a stack generation method and achieved high-quality synthesis results. Xu et al. [6] introduced the attention mechanism to obtain high-resolution results. Then, the methods of hierarchical nesting [7], mirror text comparison [8], and prior knowledge guidance [9] were proposed and achieved higher-quality image results. Although the current T2I methods have achieved encouraging results, there is still room for improvement in the quality of synthetic images. Hence the study of T2I is still challenging. In addition, the text description can only control the basic information of the synthesized image, but it is powerless to synthesize the shape, position, and other information of the synthesized object, which makes the existing T2I methods still insufficient at the level of human controllability. Furthermore, in terms of practicability, the existing T2I methods can only synthesize corresponding image results based on one input text description, and

cannot continue to input text information to modify the content of the generated image, which makes the practicability of the existing T2I methods mediocre.

## 1.2 Research Objective and Specific Scheme

In view of the three problems existing in the existing T2I method (the quality of image synthesis needs to be further improved; the controllability of the synthesis method is still insufficient; and the practicability of the synthesis process is mediocre), this research is dedicated to realizing a highly artificially controllable image synthesis method, which can achieve higher synthesis quality and have better controllability and practicability, so as to promote the development of artificially controllable image synthesis towards practical applications.

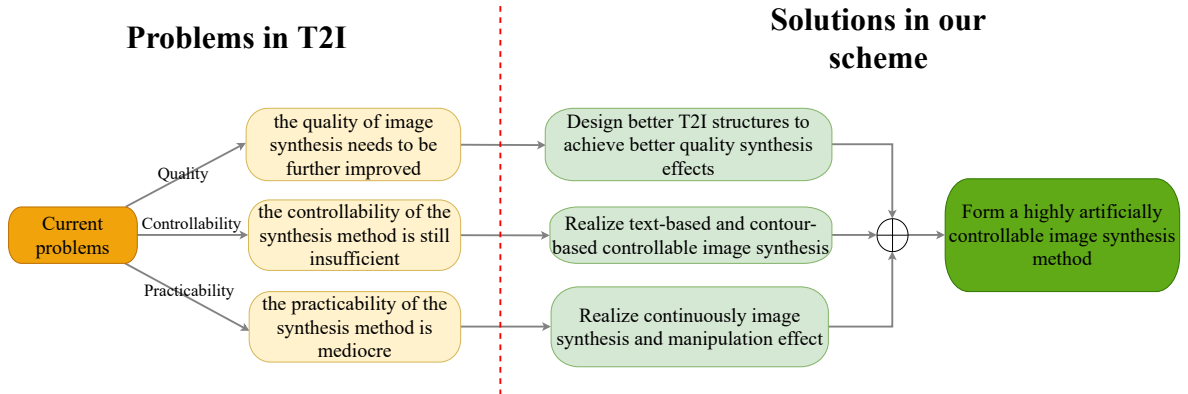


Figure 1.1: For the existing problems of T2I, our specific research scheme is shown in the figure above.

To achieve our research objective, the specific scheme is shown in Fig. 1.1. For the problem that the quality needs to be improved, we aim to design better T2I structures to improve the quality of image synthesis. For the problem that the controllability is still insufficient, we aim to design a controllable image synthesis method based on text and contour information to achieve a more controllable synthesis effect. For the problem of the mediocre practicality of current synthesis methods, we aim to achieve a continuous image synthesis and manipulation effect to further improve the practica-

bility of the method. Overall, the whole scheme is divided into three parts, namely improving the quality of synthesis, improving the controllability of synthesis, and improving the practicability of the method. The research of the above three parts can not only better solve the problems existing in T2I, but also form a highly artificially controllable image synthesis method by combining the research of these three parts.

### 1.3 Overview of the Proposed Methods

The core of our research scheme includes three parts: improving quality, improving controllability, and improving practicability. The overview of our proposed methods in each part is shown in Fig. 1.2.

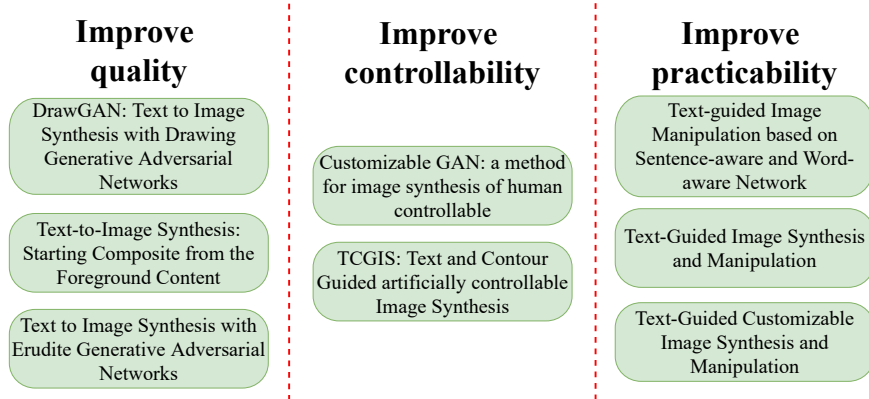


Figure 1.2: The method overview corresponding to each part of our proposed scheme is shown above.

In the part of improving quality, we have proposed three methods, namely DrawGAN: Text to Image Synthesis with Drawing Generative Adversarial Networks; Text-to-Image Synthesis: Starting Composite from the Foreground Content; Text to Image Synthesis with Erudite Generative Adversarial Networks. The basic idea of the first method is to simulate painting process. Specifically, synthesize the contour information based on the text at first, then synthesize the foreground information, and then synthesize the final image result. The basic idea of the second method is to synthesize the foreground content based on the text information, and then synthesize the final image

result. The basic idea of the third method is to improve its discriminative ability by introducing additional discriminative types in the discriminator, so as to promote the improvement of the generator’s generation ability and achieve better quality synthesis.

In the part of improving controllability, we have proposed two methods, namely Customizable GAN: a method for image synthesis of human controllable; TCGIS: Text and Contour Guided artificially controllable Image Synthesis. The first method synthesizes corresponding image results by simply fusing text and contour information; The second method introduces some more complex information processing modules (such as attention mechanism, affine combination module, etc.) based on the first method so as to achieve a better customizable synthesis effect.

In the part of improving practicability, we have proposed a TGIM method at first, namely Text-guided Image Manipulation based on Sentence-aware and Word-aware Network. The proposed method has achieved a better image manipulation effect by introducing sentence-aware and word-aware network. After that, the proposed TGIM method integrates the methods proposed in the first and second parts to achieve two highly human-controllable image synthesis methods. The first method is to first synthesize the corresponding image based on the text, and then continue to input text to continuously modify the content of the synthesized image. The second method allows humans to input text and contour to achieve customized image synthesis initially, and then allows humans to continue to input new text to modify the synthesized image content so that the synthesis results are more conform with human subjective wishes.

## 1.4 Main Contributions

This research focuses on solving the problems of existing T2I methods and is committed to realizing a highly human-controllable image synthesis method with satisfactory practicability. Fig. 1.2 shows that our proposed method is mainly composed of three parts, and the main contribution of each part is as follows:

**Improve quality.** To further improve the synthesis quality of T2I, we propose three methods. Extensive experimental verifications show that each proposed method achieves higher-quality image results, which further promotes the development of the T2I field while solving the existing problem.

**Improve controllability.** In order to solve the problem of insufficient controllability of existing synthesis methods, we propose two controllable image synthesis methods. The core is to allow human input of text and simple contour information to achieve controllable image synthesis. Experimental results demonstrate the effectiveness of our proposed method. Furthermore, our proposed method facilitates the development of image synthesis toward human controllability.

**Improve practicability.** In order to improve the practicality of the synthesis method, we first propose a TGIM method, which can modify the image’s content based on the text information. Then, combining this method with the previously proposed methods can achieve the effect of manipulating the content of the generated image. Finally, two highly human-controllable and practical image synthesis methods are realized, which greatly promotes the development of the practical image synthesis field.

## 1.5 Dissertation Outline

This dissertation is composed of four main chapters. Each chapter is a complete and independent research that could be read separately by the readers. This chapter introduces the problems existing in the existing T2I methods, indicates the main objective and basic scheme of the research, and gives a preliminary overview of the methods in the scheme.

In Chapter 2, three T2I methods we propose are introduced in detail. Specifically, the network structure of each method is introduced in detail, and corresponding extensive experimental results are shown to demonstrate the effectiveness of the proposed

method in terms of improving synthesis quality.

Chapter 3 presents our proposed controllable image synthesis method. In this part, two controllable image synthesis methods are proposed and the corresponding network structures and experimental results are shown in detail.

Chapter 4 first presents our proposed TGIM method and corresponding experimental results, and then presents the corresponding experimental results after combining the proposed TGIM method with the previously proposed image synthesis methods.

Chapter 5 presents the conclusion of this dissertation.

# Chapter 2

## High Quality Oriented Image Synthesis Methods

### 2.1 Introduction

Compared with the traditional method of image synthesis using generative adversarial networks (GAN) [1], using category labels or attributes in conditional generative adversarial networks (CGAN) [2] makes the generation more flexible and controllable. However, CGAN can only determine the category of synthetic objects and cannot determine the specific details. Recently, text-to-image synthesis has attracted more and more attention due to its high flexibility and has a wide range of application prospects, such as scene restoration, computer-aided design, *etc.* On the one hand, it defines the specific details of the synthetic image. On the other hand, the text description conforms to people's input habits, making the whole task highly flexible.

Reed et al. [3] first proposed end-to-end generative adversarial networks and successfully realized the task of text-to-image synthesis. This structure has played a proper role in the synthesis of simple and complex images. However, in terms of the synthesis results' quality, the overall resolution and authenticity are low. To further improve the resolution and authenticity of the results, Zhang et al. [4] [5] and Xu et al. [6] put forward the idea of stack generation and attention mechanism fine-tuning, respectively. Their results have been greatly improved in resolution and authenticity. Then, [7] [9] [8] [10] [11] based on the idea of stack synthesis or attention mechanism, more effective text or image processing modules are introduced to achieve better results. Although these methods have achieved excellent results, there is still a gap between the quality

of generated images and real images in the task of text-to-image synthesis.

## 2.2 Related Works

**Image Synthesis.** Effective construction of image generation modeling is the fundamental problem in computer vision. For image synthesis research, the core task is to establish a useful image generation model to synthesize more realistic image results. There has been remarkable progress in the image synthesis field with the emergence of deep learning techniques. Variational Autoencoders (VAE) [12] utilized a probability graph model to achieve a better generation by maximizing the lower bound of data likelihood. Generative Adversarial Networks (GAN) achieved remarkable image synthesis results through adversarial learning. For example, Salimans et al. [13] proposed the ImprovedGAN that can enhance the stability of results by passing the discriminator learning characteristics to the generator. DCGAN proposed by Radford et al. [14] combined GAN and CNN (Convolutional Neural Networks) together, and it has a promising performance in obtaining high-quality images. Gnanha et al. [15] proposed a parameterized robust loss function, which solves the problems of mode collapse and unstable training in GAN, and achieves higher quality image synthesis results. Although these GAN structures can obtain stunning image synthesis results, their original inputs are only noise vectors obtained by Gaussian or uniform distribution, so the overall flexibility and human controllability of these models are poor. In order to solve this problem and make the image synthesis model more useful, conditional image synthesis research has been explored. The initial condition generation model is based on simple image attributes (such as category labels [2] or sketch information [16][17]) to achieve effective control synthesis. After that, some works try to generate images based on images (pixel to pixel), including image style conversion [18][19][20], image super-resolution [21][22][23], image editing [24][25], and so on. However, because the input is images, these image synthesis methods' overall flexibility is poor, and the degree of

artificial controllability is low. In order to achieve a more flexible and effective image synthesis, image synthesis based on text description is proposed.

**Text-to-Image Synthesis.** Since the input of text-to-image synthesis research is the text description that more conforms to people’s input habits, this research field has better flexibility. Reed et al. [3] first proposed an end-to-end text-to-image synthesis structure, which can determine the composite image’s content information through the input text description. After that, in order to further improve the quality of text-to-image synthesis, multi-stage synthesis [4][5] and attention mechanism [6] methods are proposed and achieved pretty results. Based on the multi-stage synthesis and attention mechanism methods, many improved methods [7][8][9][10][26][27][28][11][29] have been proposed and further improve the quality of synthesis results. Although these methods have achieved stunning results, the overall synthesis quality can still be further improved. Unlike these improvement methods that focus on the external image or text coding module, our method’s idea is to generate the refined foreground results based on the text description and then synthesize the final image results based on the generated foreground content. The refined foreground results can better promote the synthesis of the final results, thus generating higher-quality synthetic images.

## 2.3 Preliminaries

### 2.3.1 Generative Adversarial Networks

Generative Adversarial Networks consist of a generator  $G$  and a discriminator  $D$ . The performance of  $G$  and  $D$  can be improved simultaneously through adversarial learning. Among them, the goal of  $G$  is to synthesize the data distribution similar to the original data so that it can deceive  $D$  into believing, while the goal of  $D$  is not to be deceived by

$G$ . The specific process is a min-max game. The corresponding equation is as follows:

$$\begin{aligned} \min_G \max_D V(D, G) = & \sum_{x \sim p_{data}} [\log D(x)] \\ & + \sum_{z \sim p_z} [\log(1 - D(G(z)))], \end{aligned} \quad (2.1)$$

where  $x$  and  $z$  are the original data and noise vectors, respectively.  $P_{data}$  and  $P_z$  are the distribution of the original data and Gaussian, respectively.  $\log$  represents logarithmic function.  $\min_G$  indicates that the image generated by the generator is expected to be as real as possible, that is, the loss function of the generator is expected to be minimized.  $\max_D$  represents the hope that the discriminator can maximize the distinction between generated image and the real image.

Based on GAN, CGAN introduces the conditional variable  $c$  to the generator and discriminator to determine the specific category of the synthesis image. The specific equation is as follows:

$$\begin{aligned} \min_G \max_D V(D, G) = & \sum_{x \sim p_{data}} [\log D(x, c)] \\ & + \sum_{z \sim p_z} [\log(1 - D(G(z, c)))], \end{aligned} \quad (2.2)$$

### 2.3.2 Image-Text Matching

The main task of image-text matching is to map the image features and text features to a common semantic space to measure the similarity of image and text. To better calculate semantic consistency between text and image, we employ Deep Attentional Multimodal Similarity Model (DAMSM) [6], which is a pretty way to judge semantic consistency between text and image. DAMSM is a word-level fine-grained image-text matching method. It employs the Bidirectional Long Short-Term Memory (BiLSTM) [30] to extract sentence features and word features of text description and utilizes the Inception-v3 model [31] to obtain global features and sub-region features of the image. Then, the combination of the two sets of features (text features and global

image features, word features and image sub-region features) are mapped to a common semantic space to calculate the semantic consistency.

The DAMSM’s training loss includes four aspects, and the related equations are as follows:

$$L_1 = -\log P(s \mid I_F), \quad (2.3)$$

$$L_2 = -\log P(I_F \mid s), \quad (2.4)$$

$$L_3 = -\sum_{i=1}^{N_w} \sum_{j=1}^{N_r} \log P(w_i \mid f_j), \quad (2.5)$$

$$L_4 = -\sum_{j=1}^{N_r} \sum_{i=1}^{N_w} \log P(f_j \mid w_i), \quad (2.6)$$

where  $s$  and  $w_i$  denote sentence features and  $i^{th}$  word features, respectively.  $I_F$  and  $f_j$  represent image features and  $j^{th}$  image sub-region features, respectively.  $N_w$  and  $N_r$  denote the number of word and sub-region, respectively. The first equation represents the matching degree with the sentence features under the condition of the global image features, while the second equation is the opposite. The third equation shows the matching degree with word features under the condition of image sub-region features, while the fourth equation is the opposite. The related four losses are expected to be minimized due to the use of negative log posterior probability.

Finally, the loss of DAMSM is as follows:

$$L_{DAMSM} = L_1 + L_2 + L_3 + L_4 \quad (2.7)$$

### 2.3.3 Introduction to Experimental Datasets

For our proposed T2I methods, we verify the performance on the CUB [32], Oxford-102 [33], and MS COCO [34] datasets. Table 2.3.3 shows the basic information for these datasets. The CUB dataset includes 11,788 images with 200 classes, where 8,855 images with 150 classes are utilized for training, and the rest of 2,933 images with

Table 2.1: The basic information for the CUB, Oxford-102, and MS COCO datasets.

Dataset	CUB [32]		Oxford-102 [33]		MS COCO [34]	
	Train	Test	Train	Test	Train	Test
Samples	8,855	2,933	7,034	1,155	82,783	40,504

50 classes are employed for testing. The Oxford-102 dataset contains 8,189 images with 102 categories, 7,034 images with 82 categories of which are for training, and the remaining 1,155 images with 20 categories for testing. The MS COCO dataset consists of a training set of 82,783 images and a test set of 40,504 images. Each image in the CUB and Oxford-102 contains ten text descriptions, while each image in the MS COCO contains five text descriptions.

### 2.3.4 Introduction to Evaluation Methods

To evaluate our proposed T2I methods, we employ three quantitative evaluation methods, including inception score (IS) [13], Fréchet Inception Distance (FID) [35], and R-precision [6].

The Inception Score (IS), Fréchet Inception Distance (FID), and R-precision are employed to evaluate our method quantitatively. The IS [13] uses a pre-trained Inception model [31] to evaluate the authenticity and diversity of the results. The higher the IS score, the better quality and diversity of the synthesis results. The specific IS evaluation equation is as follows:

$$IS = \exp(\sum_x KL(p(y|x) || p(y))), \quad (2.8)$$

where  $x$  represents the synthetic image, and  $y$  is the label predicted by the Inception model,  $KL$  denotes Kullback-Leibler divergence.  $\exp$  means exponential function.

FID [35] first utilizes the Inception model to extract corresponding features from the synthetic image set and original image set. It then calculates the Fréchet distance between the two sets of features through the Gaussian model. The lower score implies

closer to the real image. The specific FID evaluation equation is as follows:

$$d^2((m_{syn}, C_{syn}), (m_{ori}, C_{ori})) = \|m_{syn} - m_{ori}\|_2^2 + Tr(C_{syn} + C_{ori} - 2(C_{syn} C_{ori})^{\frac{1}{2}}), \quad (2.9)$$

The  $m_{syn}$  and  $C_{syn}$  respectively represent the means and covariance metrics obtained from synthetic image distribution  $p_{syn}$ .  $m_{ori}$  and  $C_{ori}$  respectively represent the means and covariance metrics obtained from the original image distribution  $p_{ori}$ .  $Tr$  denotes the matrix trace, which is the sum of the diagonal elements of the matrix.

R-precision is proposed by Xu et al. [6] that can evaluate whether the generated image is consistent with the given text description. For a given image query, R-precision can be measured by retrieving relevant text. Specifically, the DAMSM encoder is used to encode the generated image and candidate text descriptions to extract the corresponding feature vectors. Then the cosine similarity between global image features and candidate text features is calculated. The candidate text descriptions contain R ground truth and 100-R randomly selected mismatched text descriptions. If there are  $r$  results that are relevant in the top R ranked retrieval descriptions, then the value of R-precision is  $\frac{r}{R} \times 100\%$ . In this paper, we compute the R-precision with  $R = 1$ , and the generated images are divided into 10 folds to retrieve, and then the mean and standard deviation of the result scores are regarded as the final R-precision results. A higher score means more consistency with the semantic information of the text.

## 2.4 Method 1 — DrawGAN: Text to Image Synthesis with Drawing Generative Adversarial Networks

In order to achieve better image synthesis quality, we analogy the text-to-image synthesis task to the painting process. In the painting process, the first step is to draw object’s basic contour, fill in the specific details based on the contour, and finally com-

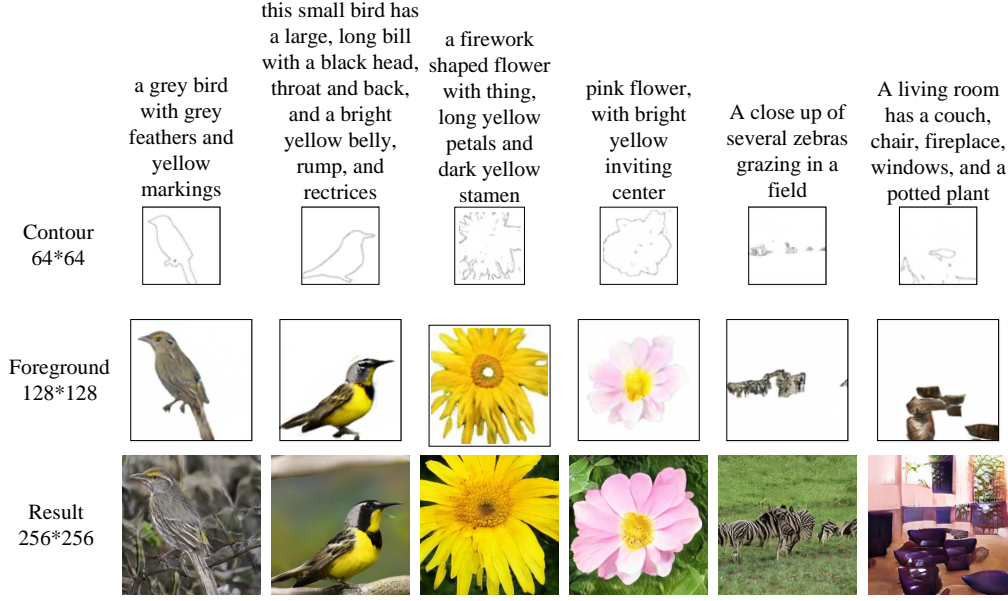


Figure 2.1: The display of three-stage synthesis results in the CUB, Oxford-102, and MS COCO datasets. Through this piece-wise synthesis from simple to complex, our model finally synthesizes high-quality images.

plete the whole drawing. Referring to such a rendering process, we first synthesize the corresponding contour information based on the text description to determine the basic shape of the object, then synthesize the corresponding foreground result with details based on the contour information, and finally synthesize the result with complete information based on the foreground image. Compared with other synthesis methods, our proposed method defines each stage’s synthesis task, which can make the network structure of each stage pay more attention to its own task to obtain high-quality synthesis results. By adding the information stage by stage, our method finally achieves excellent results, which surpasses the performance of the existing state-of-the-art methods. Fig. 2.1 shows the progressive results obtained by our method.

### 2.4.1 Network Structure

The network structure of DrawGAN on text-to-image synthesis is shown in Fig. 2.2. The input text description generates the corresponding sentence features and word features through a pre-trained text encoder [6]. For the whole sentence features, firstly,

the enhanced features are obtained by the conditioning augmentation technology [4]. And then the enhanced features combined with the noise vector of uniform distribution or Gaussian distribution to form the final input. The features of contour information (Image feature\_0) can obtain through the fully-connected layer and continuous up-sampling [36]. Through a convolution layer, the 2D contour image (Image\_0) of the first step is obtained.

For conditional augmentation technology [4], it can randomly sample latent variables from an independent Gaussian distribution ( $N(\mu(s), \sum(s))$ ), where  $\mu(s)$  and  $\sum(s)$  are the mean and diagonal covariance matrix of the sentence features ( $s$ ). Therefore, it can expand the number of training. On the other hand, to avoid the over-fitting problem, it adds the following regularization term to the objective of the generator during training.

$$D_{KL}(N(\mu(s), \sum(s)) \parallel N(0, 1)), \quad (2.10)$$

where  $s$  is the sentence features,  $KL$  is the Kullback-Leibler divergence.

For the word features, with the support of the consistency calculation method, it will select the words associated with the previous stage's image features acquired.

For the specific consistency calculation method used in our model, we make the following definitions: word features  $W$ , image local features  $F$ .

$$W = \{w_1, w_2, \dots, w_i, \dots, w_N\} \quad (2.11)$$

$$F = \{f_1, f_2, \dots, f_j, \dots, f_{N_r}\} \quad (2.12)$$

where  $w_i$  and  $f_j$  represent  $i^{th}$  word feature and  $j^{th}$  image sub-region feature, respectively.  $N$  is the number of the words, and  $N_r$  is the number of image sub-region. Following AttnGAN [6], we use a pre-trained image encoder [31] to extract image local features  $F \in \mathbb{R}^{768 \times 289}$ . Each column of  $F$  is the feature vector of an image sub-region. Therefore, for one image, there are a total of 289 image sub-regions.

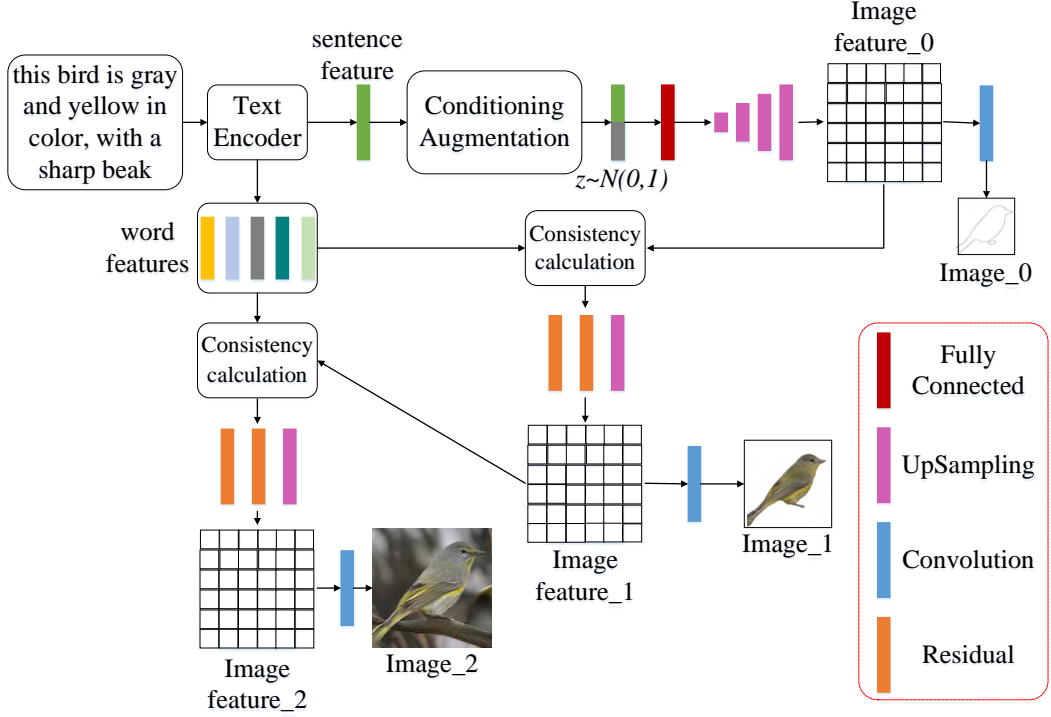


Figure 2.2: The basic architecture of our DrawGAN. From simple to complex, the model gradually synthesizes the contour image, foreground image, and final result image.

After acquiring word and image local features, it needs to establish the relationship between word and image local features, which means to retrieve the corresponding image feature content for each word feature. The weight calculation equation for the similarity probability between each word feature and image feature is as follows:

$$c_{i,j} = \frac{\exp(DT(w_i) \cdot f_j)}{\sum_{n=1}^N \exp(DT(w_n) \cdot f_j)} \quad (2.13)$$

where  $c_{i,j}$  represents the similarity probability between the  $i^{th}$  encoded word feature and the  $j^{th}$  image sub-region feature.  $DT$  denotes the dimension transformation operation, which can transfer the dimension of  $w_i$  to map the image sub-region feature's dimension so that they can perform matrix multiplication operation (denoted as ' $\cdot$ ').  $\exp$  means exponential function.

According to the calculated similarity probability, the output of consistency calculation is expressed as:

$$out_j = \sum_{i=1}^N c_{i,j} * f_j \quad (2.14)$$

After completing the consistency calculation, the output consistency calculation features are combined with the current image features to generate the next stage image features through the residual block [37] and up-sampling. Finally, the corresponding synthetic results can obtain through a  $3 \times 3$  convolution layer. The foreground image (Image\_1) of the second step and the final image (Image\_2) of the last step are all obtained in this way.

Our image synthesis process is divided into three stages, and each stage generates the contour image, foreground image, and final image result, respectively. For the image results generated by each stage, we use a corresponding discriminator to discriminate them. Following AttnGAN [6], the structure of the discriminator is designed to extract image features through multiple consecutive downsampling operations. On the one hand, the extracted image features are used to distinguish the authenticity of the image. On the other hand, image features will be combined with text features to judge the semantic consistency between them.

## 2.4.2 Loss Function

The loss function of the generator consists of two parts: adversarial loss and consistency judgment loss between image and text. The equation of adversarial loss is as follows:

$$L_{G_i} = -\frac{1}{2} [\sum_{x_i \sim P_{G_i}} \log D_i(x_i) + \sum_{x_i \sim P_{G_i}} \log D_i(x_i, s)] \quad (2.15)$$

$x_i \sim P_{G_i}$  denotes that  $x_i$  belongs to the generated image, where the value of  $i$  is 0, 1, 2, corresponding to three generation stages. Therefore,  $x_0$ ,  $x_1$ , and  $x_2$  represent the generated contour image, foreground image, and final image result, respectively.  $D_i$

represents the discriminator corresponding to the  $i^{th}$  stage.  $\log$  represents log means logarithmic function. In Eq. 2.15, the first term is unconditional loss, which is used to discriminate whether the image is true or false; the second term is conditional loss, which is used to judge whether the image matches the text.  $s$  represents the sentence features.

For the loss of semantic consistency between image and text, we use Deep Attentional Multimodal Similarity Model (DAMSM) [6] loss, which is the best way to judge semantic consistency between image and text. The equation of DAMSM loss is shown Eq. 2.7.

In summary, the final loss equation of the generator is:

$$L_G = \sum_i L_{G_i} + \lambda L_{DAMSM} \quad (2.16)$$

The discriminator's loss function only includes the adversarial loss, and the specific equation is as follows:

$$L_D = \sum_i -\frac{1}{2} [\sum_{r_i \sim P_{data}} \log D_i(r_i) + \sum_{x_i \sim P_{G_i}} \log(1 - D_i(x_i)) + \sum_{r_i \sim P_{data}} \log D_i(r_i, s) + \sum_{x_i \sim P_{G_i}} \log(1 - D_i(x_i, s))] \quad (2.17)$$

$r_i \sim P_{data}$  denotes that  $r_i$  belongs to the real image. Therefore,  $r_0$ ,  $r_1$ , and  $r_2$  represent the real contour image, foreground image, and image with background in the dataset, respectively. In Eq. 2.17, the first two items are unconditional losses for judging the authenticity of the image, and the latter two items are conditional losses for judging whether the image and text match.

The loss functions of the generator and the discriminator show that, unlike the existing T2I methods [6][10], in our proposed method, the generator sequentially synthesizes the contour image, the foreground image, and the final result with the background in three stages, while the generator of [6][10] synthesizes the image result of the background

in all three stages. In contrast, we have the ability to finally synthesize high-quality image results by refining the generation process in the generator, allowing it to be generated sequentially from simple information to complex information. Furthermore, corresponding to the three stages of the generator, our discriminator can receive three different image types (contour, foreground, and image with background) respectively and uses them for loss calculation at different stages. In this way, our proposed method can generate and discriminate specific types of images in different stages.

### 2.4.3 Implementation Details

During the specific training, Adam optimizer [38] is used with batch size 10 on the CUB, Oxford-102 flower, and MS COCO datasets, and the initial learning rate is 0.0002. The whole training process is iterated 600 epochs on the CUB and Oxford-102 flower datasets, and 120 epochs on the MS COCO dataset. The value of  $\lambda$  in Eq. 2.16 is 5 in the CUB and Oxford-102, and 50 in the MS COCO. For the text encoder and image encoder, following [6], we use a pre-trained text encoder [30] model and a pre-trained image encoder [31] model to extract corresponding text features and image features.

The related model structure details are as follows. In the up-sampling block, the scale factor is 2. After the up-sampling, it is processed by a convolution layer and a batch normalization (BN) [39]. In the residual block, it contains two convolution layers and BN operations. In the down-sampling of image features, a spectral normalization (SpectralNorm) [40] is utilized first, and then the leaky-ReLU [41] activation function is employed. For the image results of the previous stage, four SpectralNorm and leaky-ReLU operations are used continuously in the down-sampling coding.

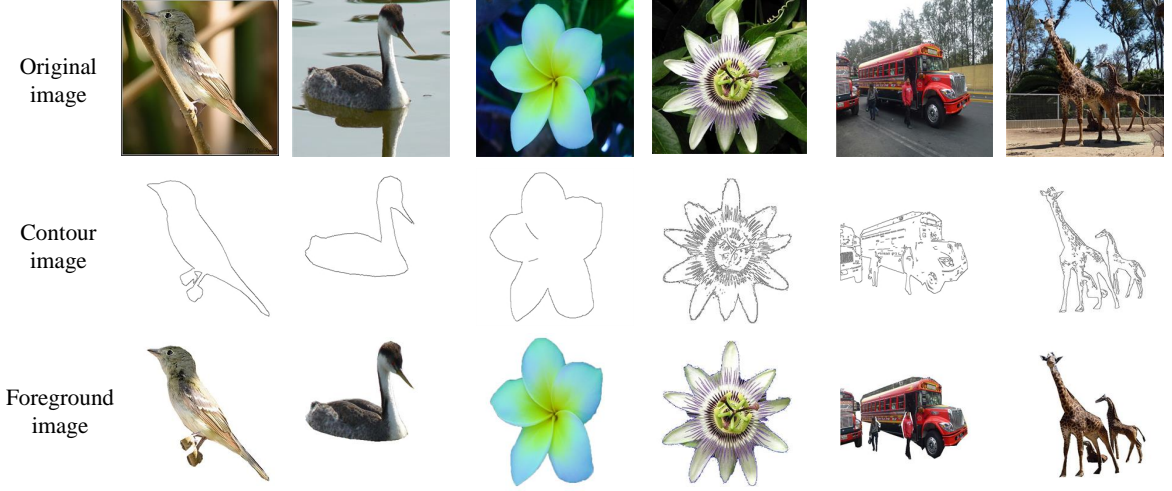


Figure 2.3: Some foreground images processed in the CUB, Oxford-102, and MS COCO datasets are displayed above.

#### 2.4.4 Experiments

There is no contour and foreground image on the CUB, Oxford-102, and MS COCO datasets, so the corresponding contour and foreground results are obtained by pre-processing at first. For the pre-processing of the CUB dataset, because it provides the binary image of the original image on the official website, we directly convert the binary image and reserve the edge to get the corresponding contour results. Besides, the foreground image can be obtained by turning the corresponding background into white in the original image by comparing the binary image with the original image. The Oxford-102 dataset provides the segmentation image with a pure blue background. Hence it can obtain the foreground image of Oxford-102 by turning the background of the segmentation image to white directly. For the pre-processing of the MS COCO dataset, the dataset provides the coordinates of the mask segmentation result, so it can directly read these coordinates to extract the foreground image. After obtaining the foreground content of the Oxford-102 and MS COCO datasets, we use a Canny operator to process the foreground image to obtain the contour image. Some pre-processing results are shown in Figure 2.3.



Figure 2.4: The comparison results of AttnGAN [6], DMGAN [10], and our method on the CUB dataset.

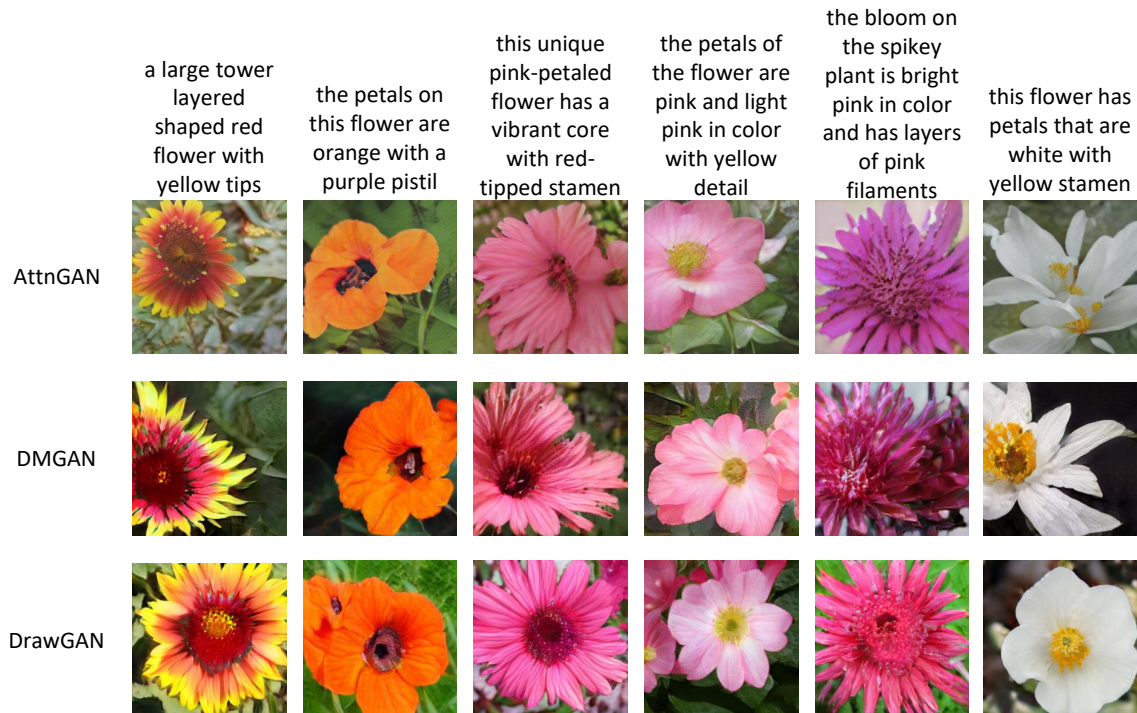


Figure 2.5: The comparison results of AttnGAN [6], DMGAN [10], and our method on the Oxford-102 flower dataset.

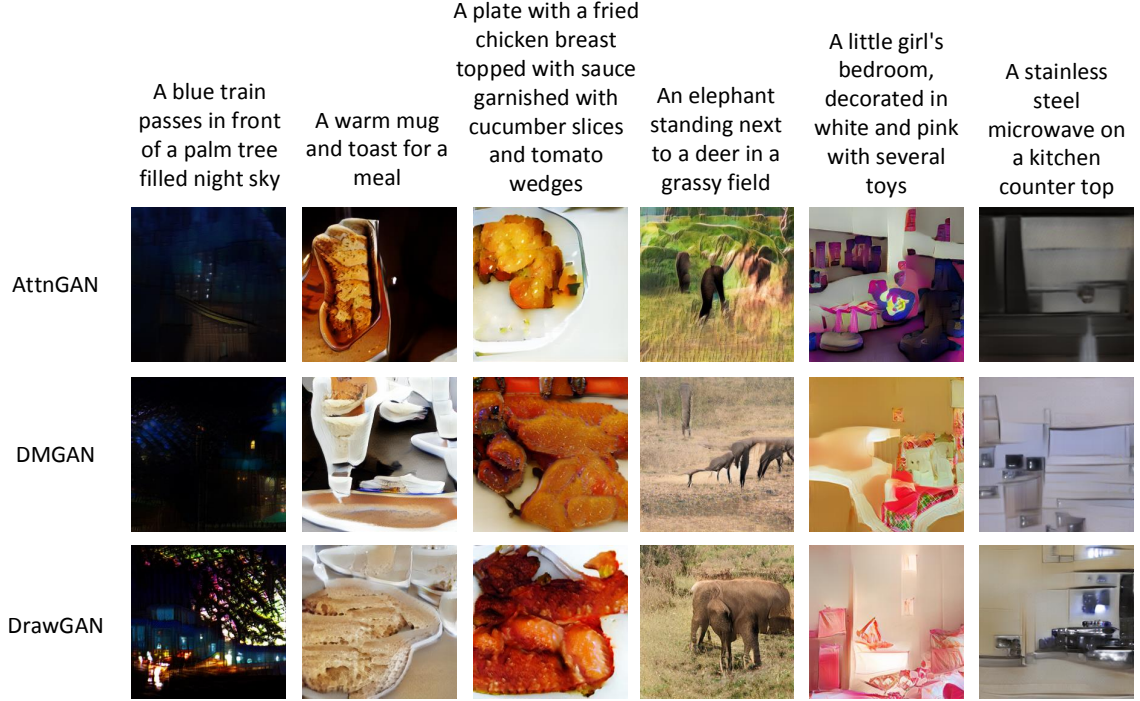


Figure 2.6: The comparison results of AttnGAN [6], DMGAN [10], and our DrawGAN on the MS COCO dataset.

### Qualitative results

The results obtained by our method are compared with those of AttnGAN and DMGAN. The specific comparison results are shown in Fig. 2.4, 2.5, and 2.6. In the bird comparison results, our method shows better performance in detail synthesis (such as eyes, pecks, and tails), overall smoothness, and authenticity. Our results are closest to the real image effect. In the flower comparison results, our results have brightness and detail processing so as to have better authenticity. In the comparison results of MS COCO, our method has better performance in overall quality. It has better subjective authenticity than the results of AttnGAN and DMGAN. Besides, the results of DrawGAN synthesis are more consistent with the content of the text description.

Table 2.2: The IS and FID comparison results of our DrawGAN and the existing methods on the CUB dataset.

Model	IS $\uparrow$	FID $\downarrow$
GAN-CLS-INT [3]	2.88 $\pm$ 0.04	68.79
GAWWN [42]	3.62 $\pm$ 0.07	53.51
StackGAN [4]	3.70 $\pm$ 0.04	35.11
StackGAN++ [5]	4.04 $\pm$ 0.05	18.02
HDGAN [7]	4.15 $\pm$ 0.05	22.70
AttnGAN [6]	4.36 $\pm$ 0.03	23.98
MirrorGAN [8]	4.56 $\pm$ 0.05	29.81
ControlGAN [26]	4.58 $\pm$ 0.09	-
LeicaGAN [9]	4.62 $\pm$ 0.06	-
DMGAN [10]	4.75 $\pm$ 0.07	16.09
DrawGAN	<b>4.76<math>\pm</math>0.04</b>	<b>9.87</b>

Table 2.3: The IS and FID comparison results of our DrawGAN and the existing methods on the Oxford-102 flower dataset.

Model	IS $\uparrow$	FID $\downarrow$
GAN-CLS-INT [3]	2.66 $\pm$ 0.03	79.55
StackGAN [4]	3.20 $\pm$ 0.01	55.28
StackGAN++ [5]	3.26 $\pm$ 0.01	48.68
HDGAN [7]	3.45 $\pm$ 0.07	-
AttnGAN [6]	3.75 $\pm$ 0.02	37.94
LeicaGAN [9]	3.92 $\pm$ 0.03	-
DMGAN [10]	4.03 $\pm$ 0.05	21.36
DrawGAN	<b>4.07<math>\pm</math>0.04</b>	<b>20.24</b>

Table 2.4: The IS and FID comparison results of our DrawGAN and the existing methods on the MS COCO dataset.

Model	IS $\uparrow$	FID $\downarrow$
GAN-CLS-INT [3]	7.88 $\pm$ 0.07	60.62
StackGAN [4]	8.45 $\pm$ 0.03	74.05
StackGAN++ [5]	8.30 $\pm$ 0.10	81.59
HDGAN [7]	11.86 $\pm$ 0.18	-
ISL [11]	12.40 $\pm$ 0.08	-
AttnGAN [6]	25.89 $\pm$ 0.47	35.49
MirrorGAN [8]	26.47 $\pm$ 0.41	-
DMGAN [10]	30.49 $\pm$ 0.57	32.64
DrawGAN	<b>31.11<math>\pm</math>0.67</b>	<b>31.51</b>

Table 2.5: The R-precision comparison results of AttnGAN [6], DMGAN [10], and our DrawGAN.

Dataset	AttnGAN	DMGAN	Our
CUB	67.82±4.43	72.31±0.91	<b>77.99±0.72</b>
Oxford-102	67.64±0.90	77.25±1.13	<b>77.70±1.00</b>
MS COCO	85.47±3.69	88.56±0.28	<b>89.20±0.40</b>

## Quantitative results

The quantitative comparison results between our method and other methods on the CUB, Oxford-102 flower, and MS COCO are shown in Tables 2.2, 2.3, 2.4 and 2.5. The results reflect our method’s excellent performance in terms of synthetic quality and match degree with text, which surpasses the existing state-of-the-art methods. This shows the effectiveness of our proposed method. By first synthesizing the contour image, then the foreground image, and then the final result image, our method refines the synthesis tasks at each stage and finally achieves better image synthesis performance.

For the results on the CUB and Oxford-102 datasets (in Tables 2.2 and 2.3), we can find that our method has little improvement in IS (about 0.21% and 0.99%) and obvious improvement in FID (about 38.6% and 5.2%) compared with the current best performance method. IS can measure the overall quality and diversity of synthetic results, and FID can measure the quality of synthetic images. Therefore, the results on the CUB and Oxford-102 datasets show that on these two datasets, our proposed method has a better promotion effect on the quality of image synthesis and has a certain promotion effect in improving the diversity of synthetic images. For the results on the MS COCO dataset (in Table 2.4), compared with the current best method, our results improve IS and FID by 2.03% and 3.46%, respectively, which shows the effectiveness of our method on the MS COCO dataset.

Table 2.6: The ablation experiment results on the CUB dataset.

	First Stage	Second Stage	Third Stage
IS	$1.08 \pm .00$	$3.80 \pm .05$	$4.76 \pm .04$
FID	247.21	55.48	9.87

Table 2.7: The ablation experiment results on the Oxford-102 dataset.

	First Stage	Second Stage	Third Stage
IS	$1.40 \pm .01$	$2.82 \pm .03$	$4.07 \pm .04$
FID	293.55	110.21	20.24

### Ablation Study

The specific ablation results are shown in Tables 2.6, 2.7, and 2.8. The first stage, second stage, and third stage in the tables represent the corresponding generated contour image (Image\_0), foreground image (Image\_1), and final image (Image\_2). Tables 2.6 and 2.7 show the ablation experimental results under the CUB and Oxford-102 flower datasets. It reflects that the first stage’s contour results are very poor according to IS and FID, while the second stage’s foreground results have been significantly improved. On this basis of the second stage, the third stage finally synthesizes higher-quality results. Table 2.8 shows the ablation experimental results under the MS COCO dataset. It shows that the results of both IS and FID in the first and second stages are very poor, and better results are obtained only in the third stage. The main reason is that the MS COCO dataset’s images are complex images, and its complexity is reflected in the foreground and background content. Due to the lack of background information, the foreground image in the second stage can not achieve the essential improvement. In contrast, the core of the bird and flower images is mainly reflected in the foreground objects, so the second stage results in the CUB and Oxford-102 datasets can be improved

Table 2.8: The ablation experiment results on the MS COCO dataset.

	First Stage	Second Stage	Third Stage
IS	$2.03 \pm .03$	$4.00 \pm .05$	$31.11 \pm .67$
FID	308.02	171.28	31.51

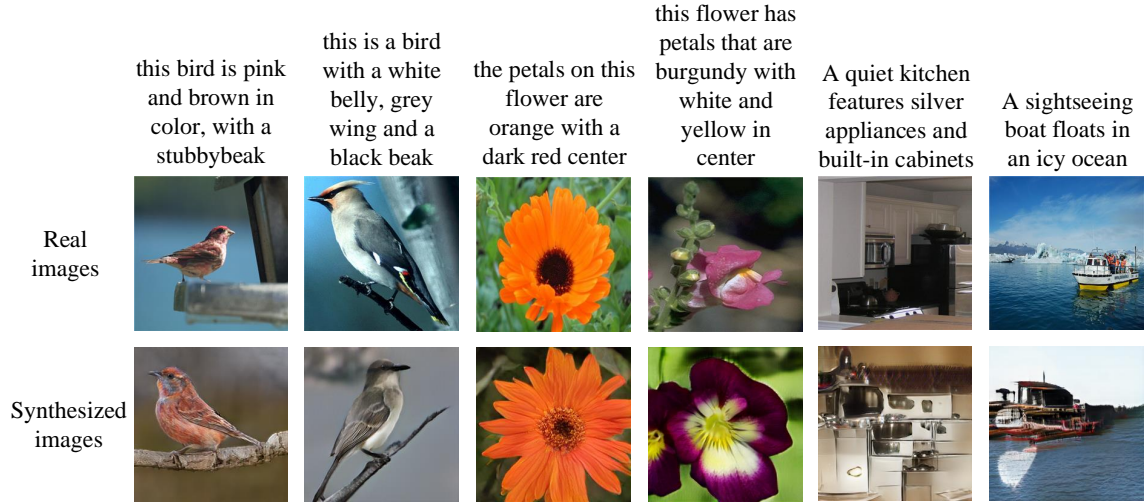


Figure 2.7: The above figure shows the comparison between our synthetic image and the real image. It can be seen that the synthesized results of our proposed method are basically equivalent to the real image effect.

essentially.

## 2.5 Method 2 — Text-to-Image Synthesis: Starting Composite from the Foreground Content

Our proposed method 1 (denoted as DrawGAN) improves the quality of image synthesis by refining the three-stage synthesis task. In the first-stage synthesis task, DrawGAN is to synthesize the contour image. However, the input text information does not contain contour information, which may hinder the improvement of image quality for subsequent synthesis. Therefore, to further improve the quality of image synthesis, we propose a multi-stage synthesis method starting the composite from the foreground content. Different from DrawGAN, this method first synthesizes the foreground result based on the text information, and then synthesizes the final image result. The content of the foreground result is highly correlated with the content of the input text information, which can better promote the quality of the final synthetic image result. Like DrawGAN, this method also includes three synthesis stages. Specifically, it syn-

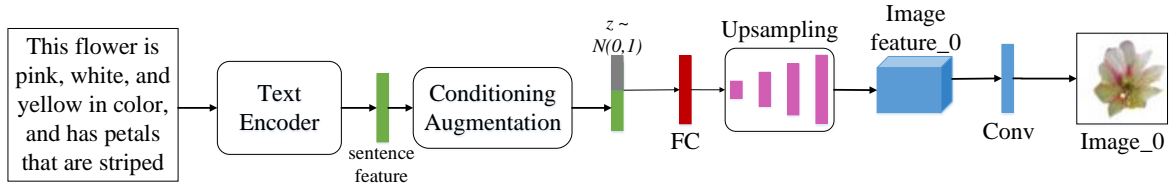


Figure 2.8: The first stage generation structure of our method is shown above. It mainly carries out the continuous up-sampling operation on text features to realize image synthesis. Image\_0 indicates the synthesis result of the first stage.

thesizes foreground results in the first stage. For the synthesis of the second stage, there are two cases: continue to synthesize the foreground result or initially synthesize the image result with background information. In the third stage, the final result with background information is synthesized based on the result synthesized in the second stage. It is worth mentioning that in the following content, the situation of continuing to synthesize the foreground result in the second stage is denoted as *fore\_1*, while the preliminary synthesis of the image result with background information is denoted as *fore\_1&2*.

In our proposed method, In the foreground synthesis stage, the whole architecture can pay more attention to the synthesis of foreground objects so as to generate the refined foreground result. The refined foreground result can play a good role in promoting the subsequent image synthesis, and the higher-quality image result can finally be achieved. Figure 2.7 shows the comparison between the synthesized results of our method and the real images. The comparison shows that our results have basically equivalent to the real image synthesis effect.

### 2.5.1 Network Structure

In our specific method, the synthesis process is divided into three stages. The first stage’s generation structure is shown in Figure 2.8. For the input text description, a text encoder [6] is used to encode the sentence feature, and then the sentence features are enhanced by conditional augmentation technology [4]. The corresponding content

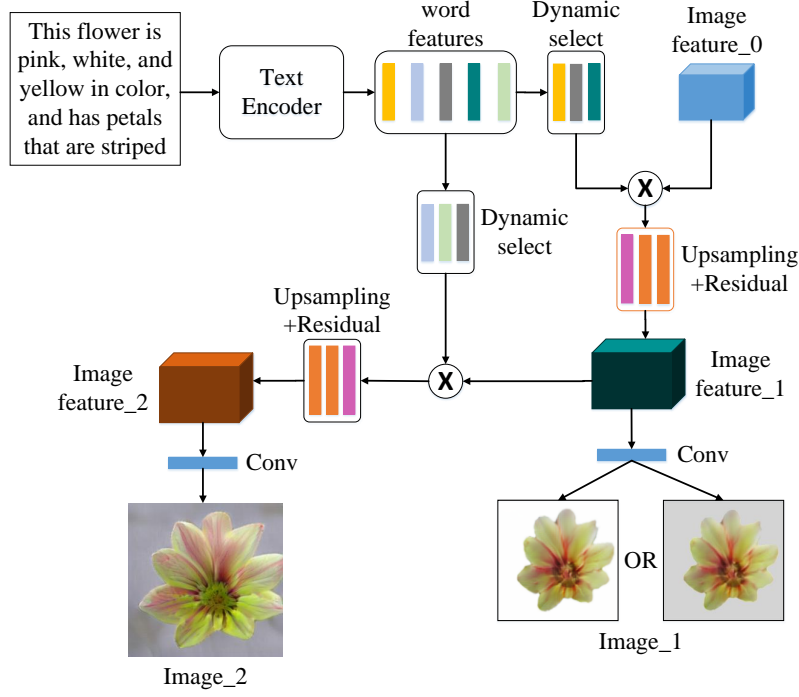


Figure 2.9: The figure above shows our proposed method’s second and third stages generation structure. Image\_1 and Image\_2 indicate the synthesis results of the second and third stages, respectively. The results of the second stage include two cases: continue to synthesize foreground results or initially synthesize the image result with background information.

of the conditional augmentation is shown near Eq. 2.10.

After the conditional augmentation, the sentence features are first transformed by a fully connected (FC) layer and then obtain the corresponding image features by continuously up-sampling. Finally, the image features are transformed into the corresponding image by a  $3 \times 3$  convolution layer.

Figure 2.9 shows the synthesis structure of the second and third stages. Unlike the first stage that encodes the input text description as the sentence features, the text encoder [6] in the second and third stages encodes the text description as corresponding word features. In the second stage, the dynamic selection method [10] is employed to select the word features that match the image features synthesized in the first stage from the encoding word features and then combine the selected word features with the image features of the first stage to generate the image features of the second stage

through up-sampling and residual [37] operations. Figure 2.9 shows that in the second stage, the foreground result or image result with background is synthesized. Similarly, the third stage also uses the dynamic selection method to select the word features that match the second stage image features and then combine them to synthesize the third stage image features through up-sampling and residual operations. The image features in the second and third stages also through a  $3 \times 3$  convolution layer to generate the corresponding image results.

For the dynamic selection method in this work, the corresponding specific details are presented. In the beginning, the dynamic selection method first selects the related words to refine the previous stage's image. It utilizes the sigmoid function to process word features and image features to calculate the importance of each word and then obtains the prior knowledge by combining the image features and the word features. The relevant equations are as follows:

$$h_i = \sigma(DT(w_i) + DT(\bar{R})), \quad (2.18)$$

$$k_i = DT(w_i) * h_i + DT(\bar{R}) * (1 - h_i), \quad (2.19)$$

where  $w_i$  is the  $i^{th}$  word feature,  $\bar{R} = \frac{1}{N_r} \sum_{i=1}^{N_r} f_i$ ,  $f_i$  is the  $i^{th}$  image region feature,  $\bar{R}$  represents the average of image sub-region features.  $N_r$  denotes the number of image regions.  $\sigma$  is a sigmoid function.  $h_i$  is a value that indicates the importance of the  $i^{th}$  word.  $DT$  represents dimension transformation operation, which is to enable  $w_i$  and  $\bar{R}$  to perform normal matrix operations.  $k_i$  is considered prior knowledge obtained by fusing image features and word features according to the calculated word importance.

The acquired prior knowledge initially constructs the relationship between word and image features. Then, in order to further strengthen this connection, the similarity of each prior knowledge and image region feature is calculated. The specific calculation

equation is as follows:

$$c_{i,j} = \frac{\exp((k_i)^T \cdot f_j)}{\sum_{n=1}^N \exp((k_n)^T \cdot f_j)}, \quad (2.20)$$

where  $c_{i,j}$  represents the similarity probability between the  $i^{th}$  prior knowledge feature ( $k_i$ ) and the  $j^{th}$  image sub-region feature ( $f_j$ ).  $\exp$  means exponential function.  $T$  represents the matrix transpose operation.  $N$  denotes the number of words in the sentence.

After the similarity calculation, the prior knowledge will be updated as follows:

$$k_j' = \sum_{i=1}^N c_{i,j} * k_j, \quad (2.21)$$

After updating the prior knowledge, the current image features and the updated prior knowledge are combined to form the new features. After that, the next stage's image features can be obtained by up-sampling and residual block processing. The specific combination equations of image features and prior knowledge are as follows:

$$q_i = \sigma(W(k_i', f_i) + b), \quad (2.22)$$

$$f_i' = k_i' * q_i + f_i * (1 - q_i), \quad (2.23)$$

where  $k_i'$  and  $f_i$  denote  $i^{th}$  updated prior knowledge feature and  $i^{th}$  image sub-region feature, respectively.  $q_i$  is a value that uses to fuse the features of updated prior knowledge and image sub-region.  $f_i'$  is  $i^{th}$  updated image sub-region feature after fusion.  $W$  and  $b$  represent the weight and bias in the training process, respectively.

### 2.5.2 Loss Function

The whole model's loss function includes the loss of the GAN and the encoding loss (DAMSM loss) of the text encoder [6]. For the loss function of the generator, it includes the adversarial loss and the DAMSM loss. The details are as follows:

$$L_G = \sum_i L_{G_i} + \lambda L_{DAMSM}, \quad (2.24)$$

$$L_{G_i} = -\frac{1}{2} \sum_{x_i \sim P_{G_i}} \log D_i(x_i) - \frac{1}{2} \sum_{x_i \sim P_{G_i}} \log D_i(x_i, s), \quad (2.25)$$

The specific content of DAMSM loss is shown in Section 2.3.2.  $i$  indicates the  $i^{th}$  generation stage.  $x_i \sim P_{G_i}$  denotes that  $x_i$  belongs to the generated image. Specifically,  $x_0$  and  $x_2$  represent the generated foreground image and final image result with background information, respectively. For  $x_1$ , it can be the foreground image that continues to be generated or the initially generated image with background information.  $D_i$  represents the discriminator corresponding to the  $i^{th}$  stage.  $\log$  represents log means logarithmic function. In Eq. 2.25, the first item is to distinguish whether the synthesized image is realistic, and the second item is to distinguish whether the synthesized image matches the input text description.

For the discriminator's loss function, it only includes the adversarial loss. The specific equation is as follows:

$$L_D = \sum_i -\frac{1}{2} [\sum_{r_i \sim P_{data}} \log D_i(r_i) + \sum_{x_i \sim P_{G_i}} \log(1 - D_i(x_i))] - \frac{1}{2} [\sum_{r_i \sim P_{data}} \log D_i(r_i, s) + \sum_{x_i \sim P_{G_i}} \log(1 - D_i(x_i, s))], \quad (2.26)$$

$r_i \sim P_{data}$  denotes that  $r_i$  belongs to the real image. Therefore,  $r_0$  and  $r_2$  represent the real foreground image and image with the background in the dataset, respectively. For

$r_1$ , if  $x_1$  is the generated foreground result, it represents the real foreground image. if  $x_1$  is the generated image result with the background, it represents the real image with the background. In Eq. 2.26, the first term is to judge whether the image is real, and the second term is to judge whether the image and the text match.

### 2.5.3 Implementation Details

The up-sampling operation after the conditional augmentation in the generator consists of four up-sampling operations. The up-sampling operation after dynamic selection includes two residual blocks and one up-sampling block. In the up-sampling and residual block, except for the last convolution, each convolution operation is followed by a Batch Normalization (BN) [39]. In the down-sampling operation, spectral normalization [40] and leaky-ReLU [41] are used after each convolution. The value of leaky is 0.2.

The model uses Adam optimizer [38] to train 600 epochs on the CUB and Oxford-102 datasets, and 120 epochs on the MS COCO dataset. The batch size in CUB, Oxford-102, and MS COCO is 10, and the learning rate is 0.0002. The value of  $\lambda$  in Eq. 2.24 is 5 for the CUB and Oxford-102 datasets, 50 for the MS COCO dataset. Besides, for the text encoder and image encoder, we still use a pre-trained text encoder [30] model and a pre-trained image encoder [31] model to extract corresponding text features and image features.

### 2.5.4 Experiments

For the method of obtaining foreground results, we have explained in Section 2.4.4, and some processed foreground results are also shown in Figure 2.3.

#### Qualitative Results

**Stage Results.** The staged results are shown in Figures 2.10 and 2.11. Figure 2.10 corresponds to the situation that the foreground content is synthesized only in the first

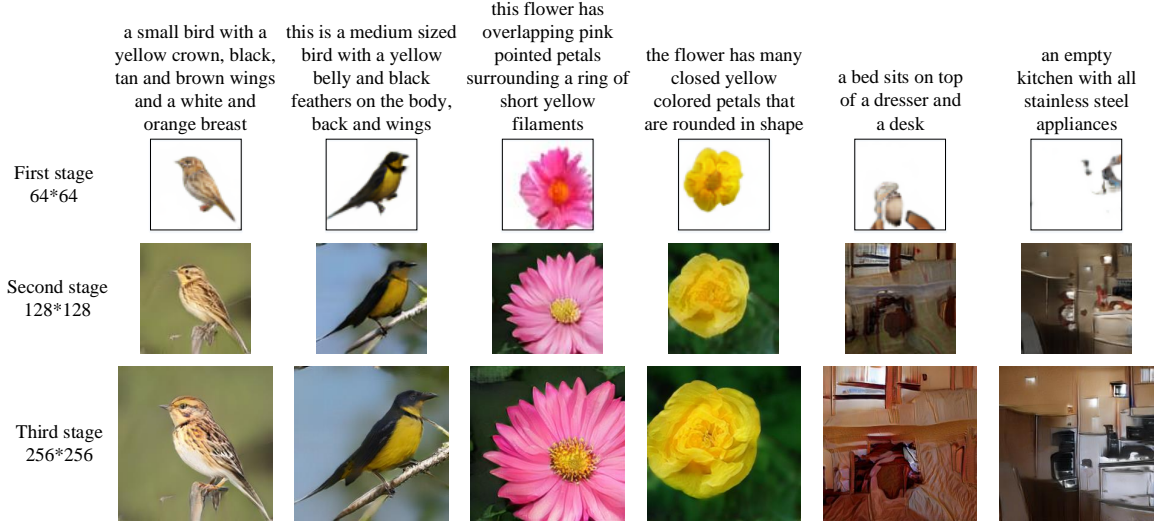


Figure 2.10: The results of the three stages corresponding to *fore\_1* are shown above. *fore\_1* means that the foreground result is synthesized only in the first stage.

stage, while Figure 2.11 is the situation that the foreground image is synthesized in the first two stages. For the bird and flower image results, whether the foreground image is synthesized only in the first stage or in the first two stages, the foreground image's synthesis effect is relatively real. It is obvious from the intuitive point of view that the generated result is a bird or a flower. But for complex images, the synthesized foreground results can not reflect the content of the text description subjectively. In particular, in the first or second stage, the foreground result basically can not reflect the corresponding text content. This phenomenon implies that in the complex image synthesis, the foreground result synthesized in the second stage can not promote the final image synthesis better because it still does not have good authenticity.

At the same time, the process of fine-tuning based on the dynamic selection method is shown in Figures 2.12- 2.17. The dynamic selection method makes local fine-tuning on the results of the second and third stages, which reflects that it can select the words related to the image region features and fine-tune them to achieve higher-quality results. In Figures 2.12- 2.17, the first row is the input text and the corresponding generated result. The second row is the fine-tuning result of the second stage. The third row is the fine-tuning result of the third stage. Comparing the fine-tuning results

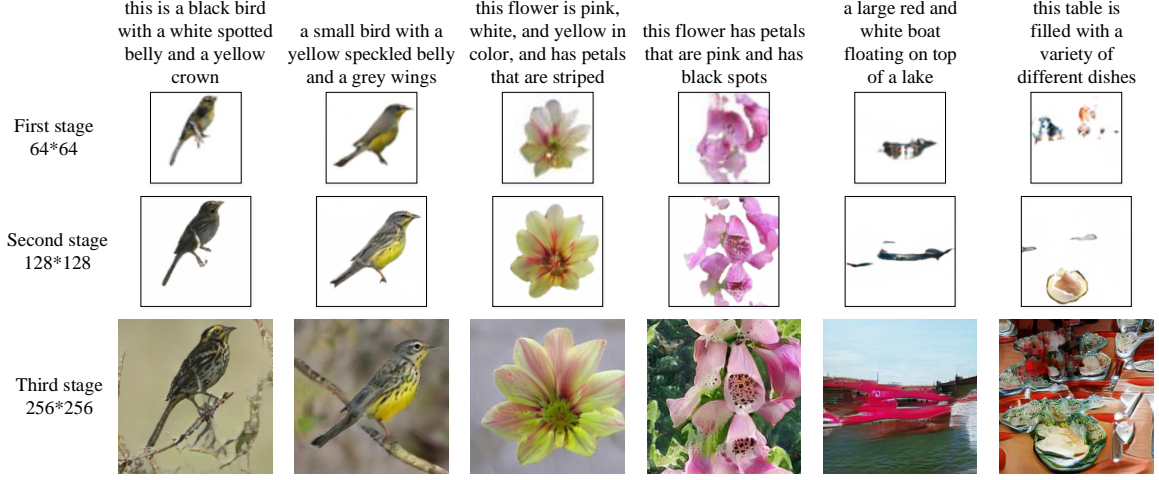


Figure 2.11: The results of the three stages corresponding to *fore\_1&2* are shown above. *fore\_1&2* indicates that both the first two stages synthesize the foreground image.

in the second row and the third row, we can find that in different synthesis stages, the dynamic selection method can select different word information and image regions for fine-tuning. This dynamicity enables the method to adjust the selection of word information and image regions to achieve better fine-tuning performance. In addition, Figures 2.12 and 2.13, 2.14 and 2.15, 2.16 and 2.17 respectively show the comparison fine-tune results of birds, flowers, and complex images corresponding to *fore\_1* and *fore\_1&2* under the same text description. In contrast, although the synthesis process of these two cases is different, they can both synthesize satisfactory image results. The quantitative comparison between them is shown in Table 2.15. The results in Table 2.15 show that *fore\_1* is more suitable for generating bird and flower images, while *fore\_1&2* is more suitable for generating complex images.

### Comparison Results.

In the quantitative results, we first present the comparison results between our method’s synthetic images and the generated images of current excellent performance methods. The specific comparison results on the three datasets are shown in Figures 2.18- 2.20. For the bird results, the overall authenticity and clarity of StackGAN results are poor, while the clarity of AttnGAN, DMGAN, and DrawGAN is improved

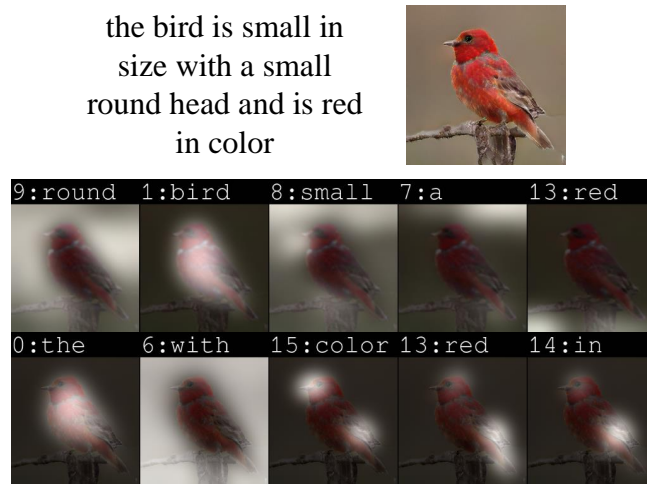


Figure 2.12: The fine-tuning results of birds in situation *fore\_1* are shown above.

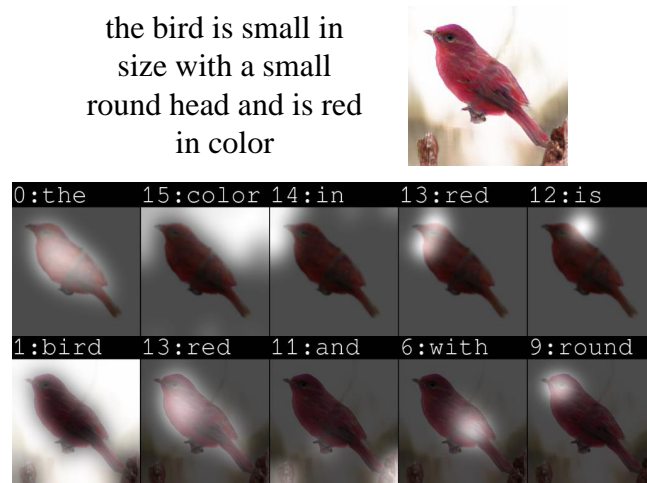


Figure 2.13: The fine-tuning results of birds in situation *fore\_1&2* are shown above.

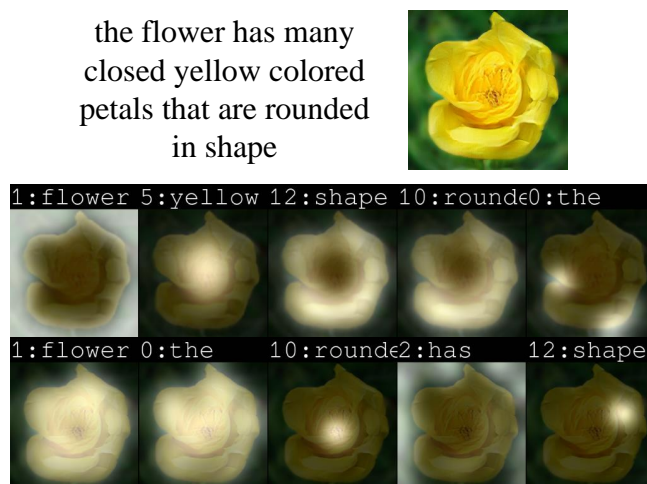


Figure 2.14: The fine-tuning results of flowers in case *fore\_1* are shown above.

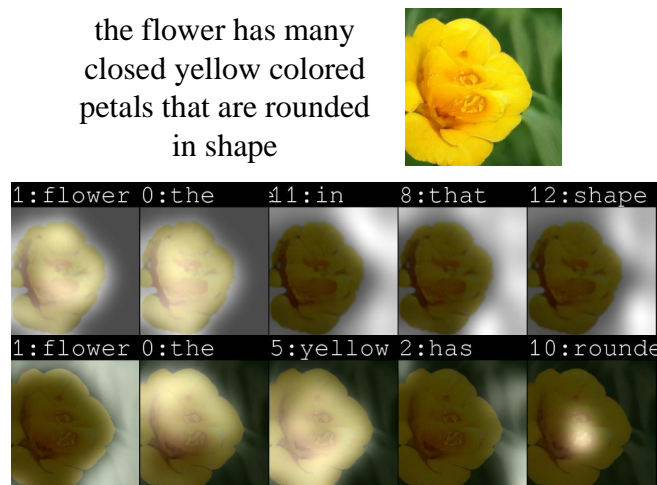


Figure 2.15: The fine-tuning results of flowers in case *fore\_1&2* are shown above.



Figure 2.16: The MS COCO'S fine-tuning results in case *fore\_1* are shown above.

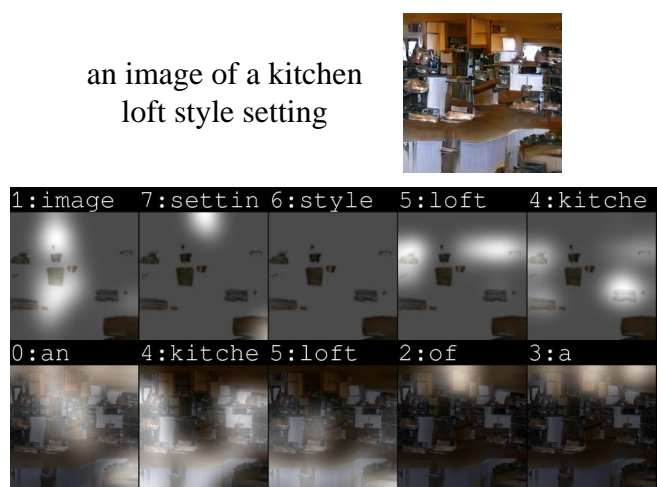


Figure 2.17: The MS COCO'S fine-tuning results in case *fore\_1&2* are shown above.



Figure 2.18: The comparison results between StackGAN [4], AttnGAN [6], DMGAN [10], DrawGAN, and our method on the CUB dataset are shown above. Our results are subjectively closest to the real image effect.

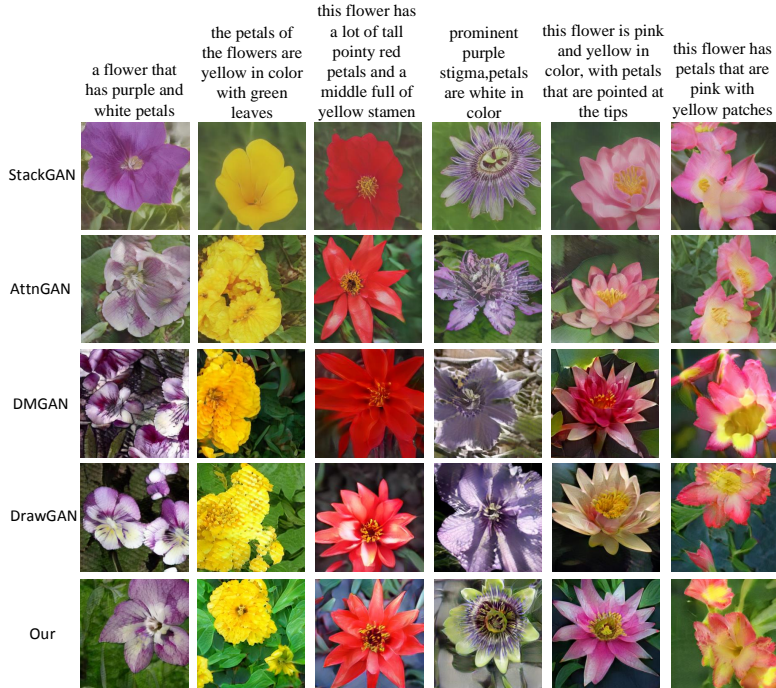


Figure 2.19: The comparison results between StackGAN [4], AttnGAN [6], DMGAN [10], DrawGAN and our method on the Oxford-102 dataset are shown above. Our flower images are the best in terms of overall shape, clarity, and authenticity.



Figure 2.20: The comparison results between StackGAN [4], AttnGAN [6], DMGAN [10], DrawGAN and our method on the MS COCO dataset are shown above. Overall, our results are more authentic than other methods.

effectively, but the overall authenticity is still weak. In contrast, the results obtained by our method have excellent performance in authenticity and clarity. Besides, our method performs better in detail processing, such as the bird’s eyes, peck, feathers, and overall smoothness.

For the flower comparison results, the flower images synthesized by all the methods have pretty authenticity. But in terms of clarity, StackGAN is still unsatisfactory. The overall clarity of AttnGAN, DMGAN, and DrawGAN is better than StackGAN. However, they are not as good as our method in detail processing, such as petals and stamens. Our method’s results are clearly visible in detail, which makes the results have the best authenticity.

For the comparison results of complex images, the results of our method also have better clarity and subjective authenticity, and the results are

more authentic than those of other methods.

From the comparison results of these three datasets, our method’s results have stunning subjective authenticity on the whole. Simultaneously, our method shows excellent performance in detail synthesis so that the generated results are basically equivalent to the real image effect.

**Stage Comparison Results.** Figures 2.21- 2.23 show the three-stage comparison results between our proposed method and AttnGAN [6], DMGAN [10], DrawGAN. For better subjective comparison, we expand the images synthesized in the first stage and the second stage into 256\*256 size. In these results, we can find that the three-stage results of AttnGAN and DMGAN both contain foreground information and background information. The three-stage result of DrawGAN is to first synthesize the simple contour information, then synthesize the foreground content, and finally synthesize the image result with background information. In contrast, our proposed method includes two ways, one is to synthesize foreground information in the first stage, the second and third stages synthesize image results with background information, and the other is to synthesize foreground information in the first and second stages, the image results with background information are synthesized in the third stage.

Compared with AttnGAN and DMGAN, which synthesize foreground and background results directly, DrawGAN makes the tasks of each synthesis stage more clear through the simple to complex synthesis method, thereby reducing the difficulty of the entire synthesis task and finally realizing better realistic image synthesis. However, DrawGAN’s synthesis method from contour to foreground and then to the final result lacks the fine-tuning process of foreground or background content, so there is still room for improvement in the quality of synthetic images. Our proposed method synthesizes the foreground content in the first stage, synthesizes the image results with background information or continues to synthesize the foreground content in the second stage, and synthesizes the image results with background information in the third stage. The

**this small black bird has black eyes, a white belly and red wingbars**

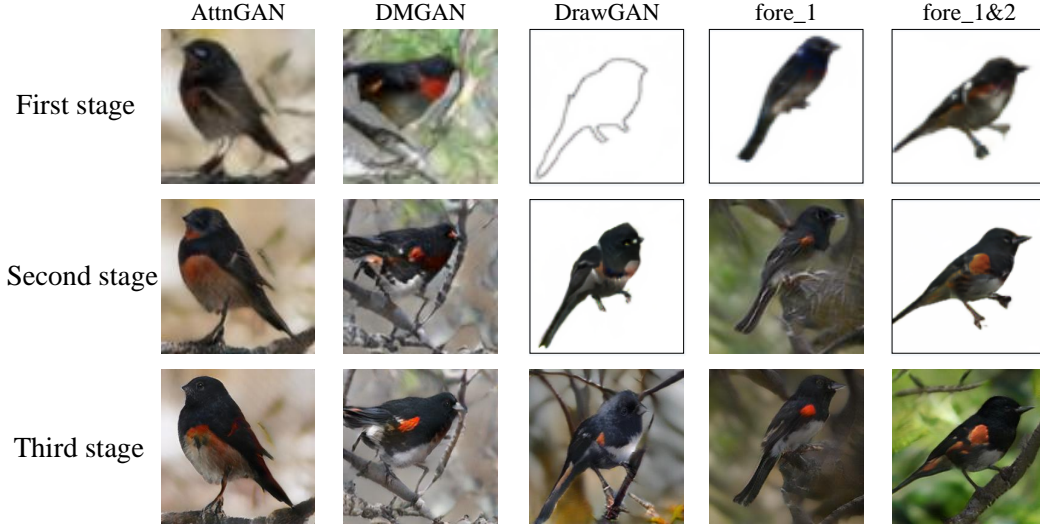


Figure 2.21: The bird results of the three stages corresponding to AttnGAN [6], DMGAN [10], DrawGAN, and our proposed *fore\_1* and *fore\_1&2* are shown above. *fore\_1* means that the foreground result is synthesized only in the first stage, while *fore\_1&2* indicates that both the first two stages synthesize the foreground image.

way of continuing to synthesize foreground content in the second stage can fine-tune the foreground synthesized in the first stage, thereby improving the synthesis quality of the foreground. When the image results with background information are synthesized in the second and third stages, the third stage can fine-tune the background information synthesized in the second stage to improve the overall synthesis quality. Therefore, compared with DrawGAN, our proposed method can better fine-tune the background or foreground information during the stage synthesis process to synthesize higher-quality image results. Besides, our method also splits the task of the entire synthesis process in each stage, thus reducing the difficulty of the synthesis task to a certain extent.

## Quantitative results

**Comparison Results.** The comparison results of the IS and FID on the CUB, Oxford-102, and MS COCO datasets are shown in Tables 2.9- 2.11. The results on the CUB

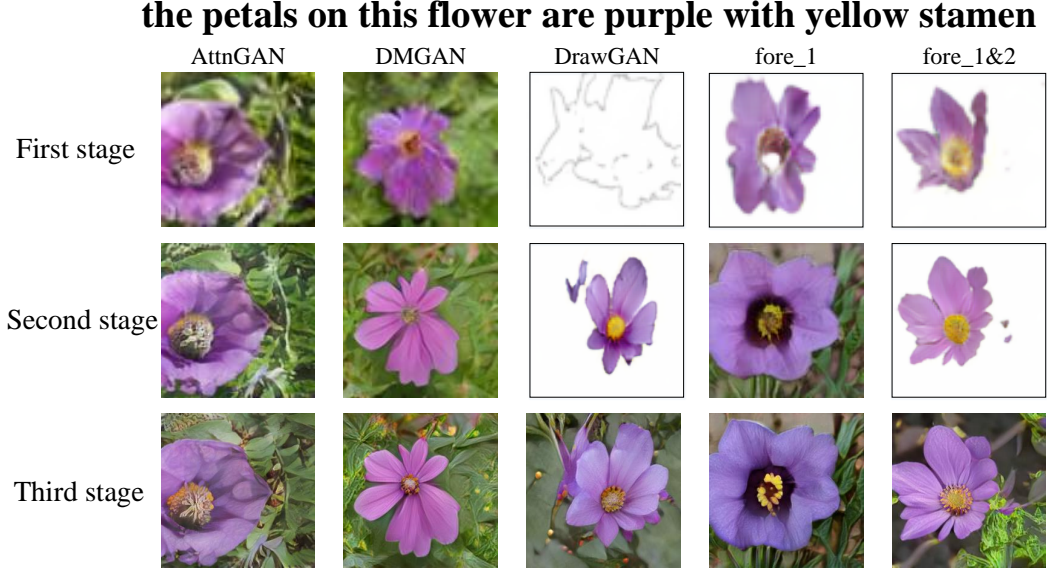


Figure 2.22: The flower results of the three stages corresponding to AttnGAN [6], DMGAN [10], DrawGAN, and our proposed *fore\_1* and *fore\_1&2* are shown above.

and Oxford-102 datasets show that our method outperforms the existing state-of-the-art methods, reflecting that our method has an excellent effect on the diversity and authenticity of the generated bird and flower images. In the MS COCO dataset, our method also shows stunning performance. It achieves excellent performance in terms of IS and FID. Especially in FID, our method performs best, demonstrating our method’s effectiveness and superiority in the MS COCO dataset.

For the comparison results of R-precision, we select AttnGAN [6], DMGAN [10], and DrawGAN with the pretty performance for comparison. The specific comparison results are shown in Table 2.12. The comparison results show that the results of bird, flower, and complex images synthesized by our method have excellent consistency with the input text description information, which further reflects our method’s effectiveness and superiority in synthetic images.

**Ablation Study.** Since our method is divided into three stages, the first stage is to synthesize the foreground result, the second stage is to generate the foreground result or the image result with background, and the third stage is to synthesize the final image result. Because there are two situations in the second stage, we conduct

**A living room with furniture in it, including a black couch  
and sofa table, shelf and TV**

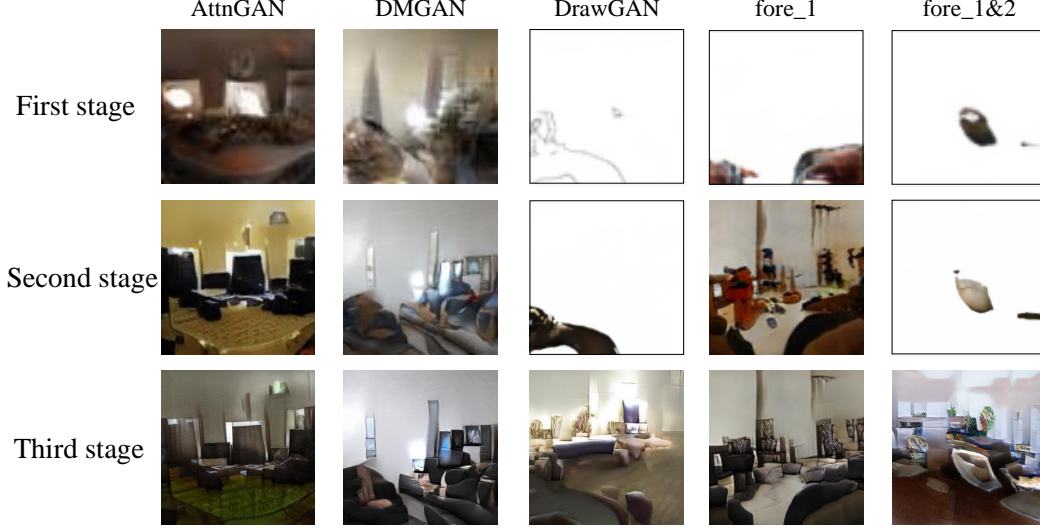


Figure 2.23: The complex results of the three stages corresponding to AttnGAN [6], DMGAN [10], DrawGAN, and our proposed *fore\_1* and *fore\_1&2* are shown above.

Table 2.9: The IS and FID comparison results of our method and the existing methods on the CUB dataset.

Model	IS $\uparrow$	FID $\downarrow$
GAN-CLS-INT [3]	$2.88 \pm 0.04$	68.79
GAWWN [42]	$3.62 \pm 0.07$	53.51
StackGAN [4]	$3.70 \pm 0.04$	35.11
StackGAN++ [5]	$4.04 \pm 0.05$	18.02
C4Synth [43]	$4.07 \pm 0.13$	-
HDGAN [7]	$4.15 \pm 0.05$	-
LDCGAN [44]	$4.18 \pm 0.06$	-
AttnGAN [6]	$4.36 \pm 0.03$	23.98
PPAN [27]	$4.38 \pm 0.05$	-
MirrorGAN [8]	$4.56 \pm 0.05$	29.81
ControlGAN [26]	$4.58 \pm 0.09$	-
LeicaGAN [9]	$4.62 \pm 0.06$	-
SEGAN [28]	$4.68 \pm 0.04$	18.16
SAM-GAN [45]	$4.61 \pm 0.03$	20.49
ICSD-GAN [46]	$4.66 \pm 0.04$	9.35
DMGAN [10]	$4.75 \pm 0.07$	16.09
TVBi-GAN [47]	-	11.83
MA-GAN [48]	$4.76 \pm 0.05$	21.66
DrawGAN	$4.76 \pm 0.04$	9.87
<b>Our</b>	<b><math>4.79 \pm 0.05</math></b>	<b>9.29</b>

Table 2.10: The IS and FID comparison results of our method and the existing methods on the Oxford-102 dataset.

Model	IS $\uparrow$	FID $\downarrow$
GAN-CLS-INT [3]	2.66 $\pm$ 0.03	79.55
StackGAN [4]	3.20 $\pm$ 0.01	55.28
StackGAN++ [5]	3.26 $\pm$ 0.01	48.68
HDGAN [7]	3.45 $\pm$ 0.07	-
LDCGAN [44]	3.45 $\pm$ 0.08	-
C4Synth [43]	3.52 $\pm$ 0.15	-
PPAN [27]	3.52 $\pm$ 0.02	-
AttnGAN [6]	3.75 $\pm$ 0.02	37.94
LeicaGAN [9]	3.92 $\pm$ 0.02	-
DMGAN [10]	4.03 $\pm$ 0.05	21.36
ICSD-GAN [46]	3.87 $\pm$ 0.05	32.64
MA-GAN [48]	4.09 $\pm$ 0.08	41.85
DrawGAN	4.07 $\pm$ 0.04	20.24
Our	<b>4.19<math>\pm</math>0.05</b>	<b>18.96</b>

ablation experiments. There are two aspects to the ablation experiment. One is the quantitative results of the three stages corresponding to the two situations on the CUB, Oxford-102, and MS COCO datasets. The other is the comparison results of two situations on these three datasets.

The three-stage results of “*fore\_1*” and “*fore\_1&2*” are shown in Tables 2.13 and 2.14. For the case that only the foreground result is synthesized in the first stage, IS and FID are significantly improved in the second stage and then achieve better results in the third stage. For the R-precision, there is no big difference between the corresponding three-stage results in CUB and Oxford-102, reflecting that the results synthesized in each stage of the two datasets are consistent with the semantic information of the text. However, the R-precision of the first stage is relatively poor in MS COCO results. The MS COCO dataset’s image content is complex, with diverse foreground objects and complex background information. Therefore, the R-precision of the foreground results synthesized in the first stage is poor in MS COCO due to the lack of background content.

For the case that the foreground result is synthesized in the first and second stage,

Table 2.11: The IS and FID comparison results of our method and the existing methods on the MS COCO dataset.

Model	IS $\uparrow$	FID $\downarrow$
GAN-CLS-INT [3]	7.88 $\pm$ 0.07	60.62
StackGAN [4]	8.45 $\pm$ 0.03	74.05
StackGAN++ [5]	8.30 $\pm$ 0.10	81.59
ChatPainter [49]	9.74 $\pm$ 0.02	-
PPGN [29]	9.58 $\pm$ 0.21	-
HDGAN [7]	11.86 $\pm$ 0.18	-
ISL [11]	12.40 $\pm$ 0.08	-
AttnGAN [6]	25.89 $\pm$ 0.47	35.49
MirrorGAN [8]	26.47 $\pm$ 0.41	-
ControlGAN [26]	24.06 $\pm$ 0.60	-
SEGAN [28]	27.86 $\pm$ 0.31	32.28
SAM-GAN [45]	27.31 $\pm$ 0.23	33.41
DMGAN [10]	30.49 $\pm$ 0.57	32.64
TVBi-GAN [47]	-	31.97
DrawGAN	<b>31.11<math>\pm</math>0.67</b>	31.51
Our	30.70 $\pm$ 0.45	<b>29.55</b>

Table 2.12: The R-precision comparison results of AttnGAN, DMGAN, DrawGAN, and our method.

Model	CUB	Oxford-102	MS COCO
AttnGAN [6]	67.82 $\pm$ 4.43	67.64 $\pm$ 0.90	85.47 $\pm$ 3.69
DMGAN [10]	72.31 $\pm$ 0.91	77.25 $\pm$ 1.13	88.56 $\pm$ 0.28
DrawGAN	<b>77.99<math>\pm</math>0.72</b>	77.70 $\pm$ 1.00	89.20 $\pm$ 0.40
Our	77.86 $\pm$ 0.3	<b>79.32<math>\pm</math>0.67</b>	<b>90.36<math>\pm</math>0.52</b>

Table 2.13: The ablation experiment analysis of the three-generation stages corresponding to the *fore*. 1, 2, and 3 in the table indicate the first stage, the second stage, and the third stage, respectively.

Dataset		IS	FID	R-precision
CUB	1	2.99 $\pm$ 0.03	130.5	76.84 $\pm$ 0.69
	2	4.34 $\pm$ 0.05	48.74	76.86 $\pm$ 0.67
	3	4.62 $\pm$ 0.0.7	14	76.88 $\pm$ 0.70
Oxford-102	1	2.28 $\pm$ 0.02	183.66	76.98 $\pm$ 0.88
	2	3.55 $\pm$ 0.05	59.31	76.99 $\pm$ 0.90
	3	3.90 $\pm$ 0.05	22.12	77.85 $\pm$ 0.57
MS COCO	1	2.28 $\pm$ 0.03	260.39	20.4 $\pm$ 0.16
	2	9.92 $\pm$ 0.15	86.04	89.26 $\pm$ 0.42
	3	30.70 $\pm$ 0.45	29.55	90.36 $\pm$ 0.52

Table 2.14: The ablation experiment analysis of the three-generation stages corresponding to the *fore\_1&2*.

Dataset		IS	FID	R-precision
CUB	1	3.23 $\pm$ 0.03	121.1	76.94 $\pm$ 1.11
	2	3.64 $\pm$ 0.04	59.31	76.99 $\pm$ 1.99
	3	4.79 $\pm$ 0.05	9.29	77.86 $\pm$ 0.53
Oxford-102	1	2.32 $\pm$ 0.02	184.16	77.61 $\pm$ 0.89
	2	2.75 $\pm$ 0.03	111.49	77.64 $\pm$ 0.89
	3	4.19 $\pm$ 0.05	18.22	79.32 $\pm$ 0.67
MS COCO	1	2.87 $\pm$ 0.04	234.94	20.0 $\pm$ 0.24
	2	3.84 $\pm$ 0.04	181.85	20.2 $\pm$ 0.24
	3	29.94 $\pm$ 0.62	31.62	89.63 $\pm$ 0.52

the results in Table 2.14 reflect that excellent performance only can be achieved in the third stage, while the results in the first and second stages are relatively poor. The reason is that the results of the first two stages are the foreground image. However, compared with the results of the first and third stages in Tables 2.13 and 2.14, “*fore\_1&2*” shows better performance than “*fore\_1*” in terms of IS and FID. It demonstrates that continuing to synthesize the foreground image in the second stage can make the foreground object more refined, which can further promote the final result’s synthesis quality in the third stage. Besides, better training in the third stage can also promote the training of the first two stages so that the results of the first stage of “*fore\_1&2*” are better than those of “*fore\_1*”.

The comparison results between the two situations are shown in Table 2.15. For the CUB and Oxford-102 results, the method of synthesizing foreground content in the first two stages achieves better results in terms of IS, FID, and R-precision. This demonstrates that using two stages to synthesize refined foreground objects can better promote the final synthesis effect and obtain higher-quality image results. For the results of MS COCO, “*fore\_1*” performs better, which indicates that the results with background information synthesized in the second stage can make the final generated complex image results better authentic. The reason is that in the situation “*fore\_1*”,

the background information generated in the second stage can be adjusted in the third stage so that it can improve the overall authenticity further. There is no such phenomenon in CUB and Oxford-102 because the authenticity of the images in these two datasets is more reflected in the foreground content, and the background information has little effect on it.

Table 2.15: The comparison experiment results between *fore\_1* and *fore\_1&2* on the CUB, Oxford-102, and MS COCO dataset.

Dataset	fore_1			fore_1&2		
	IS	FID	R-precision	IS	FID	R-precision
CUB	4.62±0.07	14.00	76.88±0.70	<b>4.79±0.05</b>	<b>9.29</b>	<b>77.86±0.53</b>
Oxford-102	3.90±0.05	25.49	77.85±0.57	<b>4.19±0.05</b>	<b>18.22</b>	<b>79.32±0.67</b>
MS COCO	<b>30.70±0.45</b>	<b>29.55</b>	<b>90.36±0.52</b>	29.94±0.62	31.62	89.63±0.52

## 2.6 Method 3 — Text to Image Synthesis with Erudite Generative Adversarial Networks

The core of our proposed method 1 (denoted as DrawGAN) and method 2 (denoted as INS\_fore) is to refine the generation process of the generator to synthesize higher-quality image results. In generative adversarial networks (GAN), in addition to the generator, the discriminator is also crucial, which needs to provide high-quality discriminative feedback to promote the generator to achieve high-quality image synthesis. Based on this, in this method, we are committed to improving the discriminative ability of the discriminator so that it can improve the generation ability of the generator and finally synthesize higher-quality results.

Specifically, we propose an erudite generative adversarial network (denoted as EruditeGAN). In EruditeGAN, the foreground images and segmentation images related to the original images are introduced into the discriminator. The introduction of these images can be regarded as additional discriminant types so as to improve the discrim-

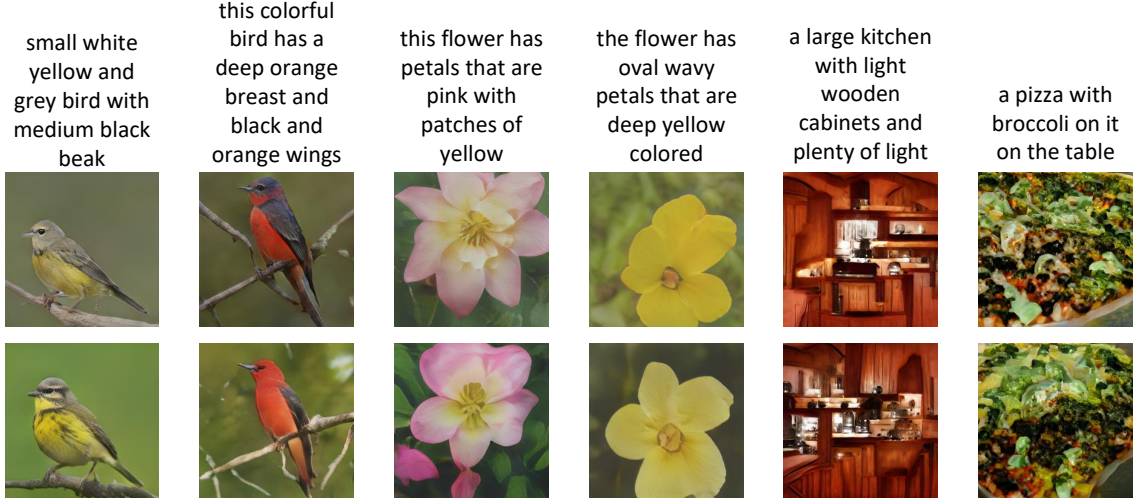


Figure 2.24: Some image results synthesized by our proposed EruditeGAN based on text description are shown above.

inant ability of the discriminator. Due to the confrontation characteristics of GAN, the improvement of the discriminator ability can circuitously improve the generator’s generation ability so that the generator can finally synthesize higher-quality results. Fig. 2.24 shows some results obtained by our proposed EruditeGAN.

### 2.6.1 Network Structure

The structure of the generator is shown on the left of Figure 2.25. The text description is encoded as word features and sentence features by a pre-trained text encoder [30]. Among them, sentence features are expanded by conditional augmentation (CA) [4] technology (the details are shown near Equation 2.10), and then combined with the noise vector, the image features are generated through a fully-connected layer and continuous upsampling operations. The generated image features can be converted into image results through a  $3 \times 3$  convolution layer. Simultaneously, the synthesized image features and word features will use the dynamic selection method (consistent with the dynamic selection method in Section 2.5, the details are shown in Equation 2.18-2.23) to improve semantic consistency. After the dynamic select processing, the features will be processed by upsampling and residual block [12] operations to synthesize the next

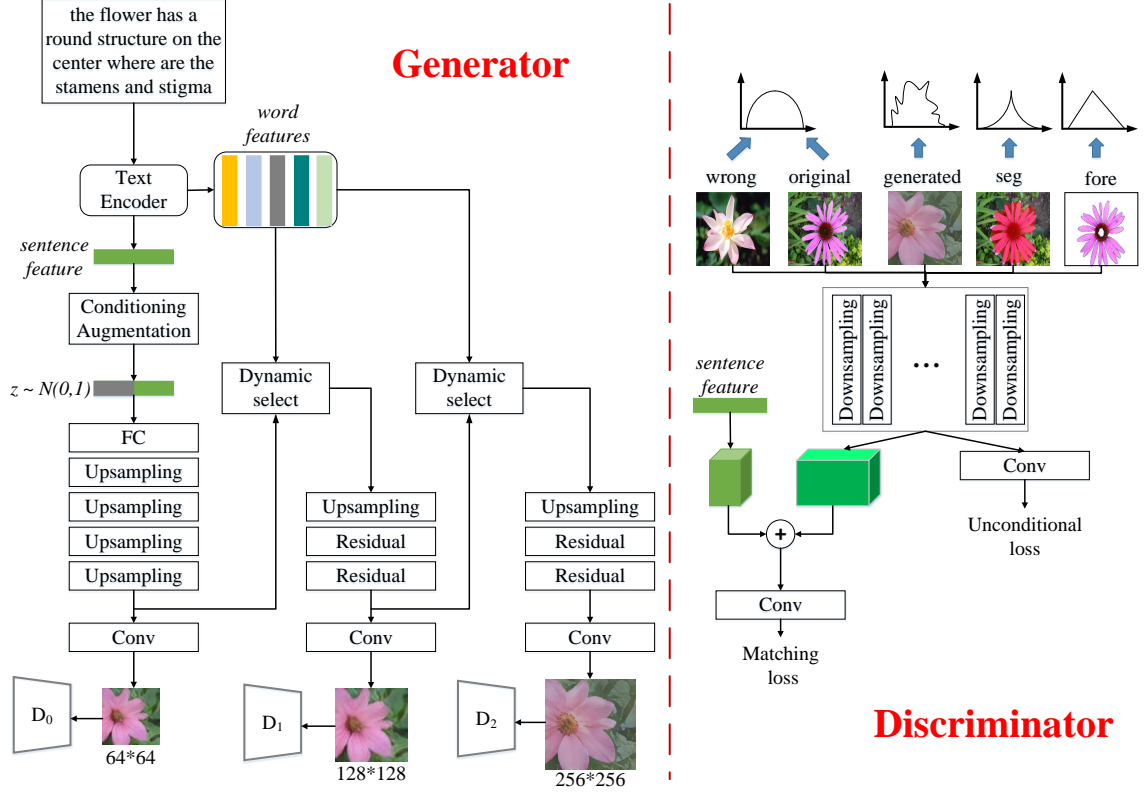


Figure 2.25: The network structure in the generator and discriminator is shown above.

stage's image features. This process can be continued to synthesize higher-resolution images.

The discriminator structure is shown on the right Figure 2.25. For the received images, the discriminator discriminates from two aspects: whether the image is true and whether the image and text match. The downsampling features of the image can be used directly to distinguish real or generated images. For consistency matching, the downsampling features are combined with the extended dimension text vector to identify the consistency.

## Training algorithm

For our proposed method, the specific training process is shown in Algorithm 1. There are four kinds of input images in the algorithm, i.e., real/wrong/foreground/segmentation image. The generated image synthesized by the generator is the fifth type. The

---

**Algorithm 1** EruditeGAN training algrotihm

---

```
1: Input: original real image  $I_r$ , wrong image  $I_w$ , segmentation image  $I_{seg}$ ,  
foreground image  $I_{fore}$ , text  $t$   
2: for  $n = 1$  to  $epochs$  do  
3:    $z = N(0, 1)$   
4:    $s = \phi(t)$   
5:    $I_{gen} = G(z, s)$   
6:    $d_{r\_unc}, d_r = D(I_r, s)$   
7:    $d_{w\_unc}, d_w = D(I_w, s)$   
8:    $d_{gen\_unc}, d_{gen} = D(I_{gen}, s)$   
9:    $d_{seg\_unc}, d_{seg} = D(I_{seg}, s)$   
10:   $d_{fore\_unc}, d_{fore} = D(I_{fore}, s)$   
11:   $L_{D\_t\_unc} = (\log(d_{r\_unc}) + \log(d_{w\_unc}))/2$   
12:   $L_{D\_f\_unc} = (\log(1 - d_{gen\_unc}) + \log(1 - d_{fore\_unc})) + \log(1 - d_{seg\_unc})/3$   
13:   $L_{D\_unc} = L_{D\_t\_unc} + L_{D\_f\_unc}$   
14:   $L_{D\_t\_mat} = \log(d_r)$   
15:   $L_{D\_f\_mat} = (\log(1 - d_{gen}) + \log(1 - d_w) + \log(1 - d_{fore})) + \log(1 - d_{seg})/4$   
16:   $L_{D\_mat} = L_{D\_t\_mat} + L_{D\_f\_mat}$   
17:   $L_D = L_{D\_unc} + L_{D\_mat}$   
18:   $D = D - ss * \partial L_D / \partial D$   
19:   $L_G = \log(d_{gen})$   
20:   $G = G - ss * \partial L_G / \partial G$   
21: end
```

---

discriminator discriminates these five kinds of images and obtains their unconditional loss and matching loss, respectively.  $L_{D\_t}$  represents the loss of the true sample of the discriminator, and  $L_{D\_f}$  represents the loss of the wrong sample. *unc* indicates the unconditional case. *mat* stands for the matching case. In the case of unconditional loss, both real and wrong images come from datasets, so they are true labels. Under the matching loss, only the real image comes from the dataset, so it is the true label. The rest are false labels. *ss* in the algorithm represents step size.

### 2.6.2 Loss Function

According to the content of algorithm 1, the specific loss functions are as follows:

$$\begin{aligned}
L_D = \sum_{I_r, I_w \sim p_{data}, I_{gen} \sim p_G, I_{fore} \sim p_{fore}, I_{seg} \sim p_{seg}} \{ & [\log D_0(I_r, s) + \log D_0(I_w, s)]/2 \\
& [\log(1 - D_0(I_{gen}, s)) + \log(1 - D_0(I_{fore}, s)) + \log(1 - D_0(I_{seg}, s))]/3\} \\
& + \{\log D_1(I_r, s) + [\log(1 - D_1(I_{gen}, s)) + \log(1 - D_1(I_w, s)) + \\
& \log(1 - D_1(I_{fore}, s)) + \log(1 - D_1(I_{seg}, s))]/4\}
\end{aligned} \tag{2.27}$$

$$L_G = \sum_{(I_{gen}) \sim p_G} \log D_0(I_{gen}, s) + \log D_1(I_{gen}, s) \tag{2.28}$$

where  $D_0$  represents the discriminator's first output (unconditional discrimination, that is, to distinguish the authenticity of the image), and  $D_1$  represents the second output (conditional discrimination, that is, to determine whether image and text match).  $I_r$  and  $I_w$  represent real images and wrong images that conform to the image distribution of the original dataset.  $I_{gen}$ ,  $I_{fore}$ , and  $I_{seg}$  represent the generated image, foreground image, and segmentation image conform to the generated image distribution, foreground image distribution, and segmented image distribution.  $\log$  represents log means logarithmic function.

### 2.6.3 Implementation Details

During the up-sampling, in addition to the last convolution, batch normalization (BN) [39] is performed after each convolution. For text embedding, it performs leaky-ReLU [41] activation after encoding. The leaky value is 0.2. In our proposed method, we use ADAM optimization [38] to train 600 epochs for the CUB and Oxford-102 datasets, and 120 for the MS COCO dataset, with an initial learning rate of 0.0002 and a batch size of 10. For the text encoder and image encoder, we still use a pre-trained text encoder [30] model and a pre-trained image encoder [31] model to extract corresponding text



Figure 2.26: Some segmentation images processed in the CUB, Oxford-102, and MS COCO datasets are displayed above.

features and image features. For the acquisition of mask segmentation images, we directly use a pre-trained mask RCNN [50] model to process the original image so as to obtain the corresponding mask segmentation image. Some obtained mask segmentation results are shown in Figure 2.26. For the input original image, mask RCNN can detect the key target and mark the target with the mask.

## 2.6.4 Experiments

### Qualitative results

In the qualitative results, the existing state-of-the-art methods are compared with our EurditeGAN. The comparison results are shown in Figure 2.27. Among them, the results of AttnGAN and DMGAN are general in terms of overall authenticity. By contrast, our results show better authenticity and are closer to the real image. Besides, our results also have better performance in the fine-grained synthesis, such as eyes, pecking, tail of bird, smoothness, the brightness of flower, and the detail of pizza and kitchen, which reflect that our proposed method is excellent in detail processing.

### Quantitative results

The IS and FID comparison results on the CUB, Oxford-102 flower, and MS COCO datasets are shown in Table 2.16. The comparison results show that the performance



Figure 2.27: The comparison results between our EruditeGAN and AttnGAN, DMGAN are shown above.

Table 2.16: The comparison results of IS and FID on the CUB, Oxford-102 flower, and MS COCO datasets between our method and existing state-of-the-art methods.

Model	CUB		Oxford-102		MS COCO	
	IS	FID	IS	FID	IS	FID
GAN-CLS [3]	2.88±0.04	68.79	2.66±0.03	79.55	7.88±0.07	60.62
GAWWN [42]	3.62±0.07	51.89	-	-	-	-
StackGAN [4]	3.70±0.04	35.11	3.20±0.01	55.28	8.45±0.03	74.05
StackGAN++ [5]	4.04±0.05	15.30	3.26±0.01	48.58	8.30±0.10	81.59
HDGAN [7]	4.15±0.05	22.70	-	29.55	-	-
AttnGAN [6]	4.36±0.03	23.98	3.75±0.02	35.49	25.89±0.47	35.49
MirrorGAN [8]	4.56±0.05	29.81	-	-	26.47±0.41	-
SEGAN [28]	4.68±0.04	18.17	-	-	27.86±0.31	32.38
DMGAN [10]	<b>4.75±0.07</b>	16.09	4.03±0.05	21.36	30.49±0.57	32.64
TVBi-GAN [47]	-	11.83	-	-	-	31.97
EruditeGAN	4.69±0.07	<b>9.58</b>	<b>4.07±0.05</b>	<b>17.69</b>	<b>31.94±0.47</b>	<b>28.79</b>

Table 2.17: The R-precision comparison results of AttnGAN, DMGAN, and our method.

Model	CUB	Oxford-102	MS COCO
AttnGAN [6]	67.82±4.43	67.64±0.89	85.47±3.69
DMGAN [10]	72.31±0.91	77.25±0.77	88.56±0.28
EruditeGAN	<b>77.62±0.90</b>	<b>80.25±0.69</b>	<b>91.26±0.40</b>

Table 2.18: The comparison results of IS and FID on the CUB, Oxford-102 flower, and MS COCO datasets between our method and existing state-of-the-art methods. ‘√’ represents to use this type of image and ‘−’ means not to use it.

fore	seg	CUB			Oxford-102			MS COCO		
		IS	FID	R	IS	FID	R	IS	FID	R
√	-	4.63	12.01	73.22	4.01	22.75	77.95	29.68	30.83	89.64
-	√	4.67	10.22	75.53	4.03	19.45	79.12	30.05	31.79	90.48
√	√	<b>4.69</b>	<b>9.58</b>	<b>77.62</b>	<b>4.07</b>	<b>17.69</b>	<b>80.25</b>	<b>31.94</b>	<b>28.79</b>	<b>91.26</b>

of our method is better than the existing state-of-the-art methods. IS can measure the quality and diversity of the generated results, and FID can measure the distance between the synthetic image and the real image. Therefore, our method performs excellently on IS and FID can demonstrate that the synthetic results of our proposed method have better quality and are closer to the real image effect. The R-precision comparison results among AttnGAN, DMGAN, and our EruditeGAN are shown in Table 2.17. The results reflect that our method’s synthesized images are most consistent with the input text’s semantic information.

## Ablation Study

In order to further analyze the effectiveness of different image types introduced into the discriminator, we conduct ablation experiments. The specific ablation results are shown in Table 2.18. The first two lines’ results reflect that it can play a useful role in promoting whether the foreground image or the segmentation image is added to the discriminator. The results of the last line show that the best results can be achieved when both types of images are added to the discriminator.

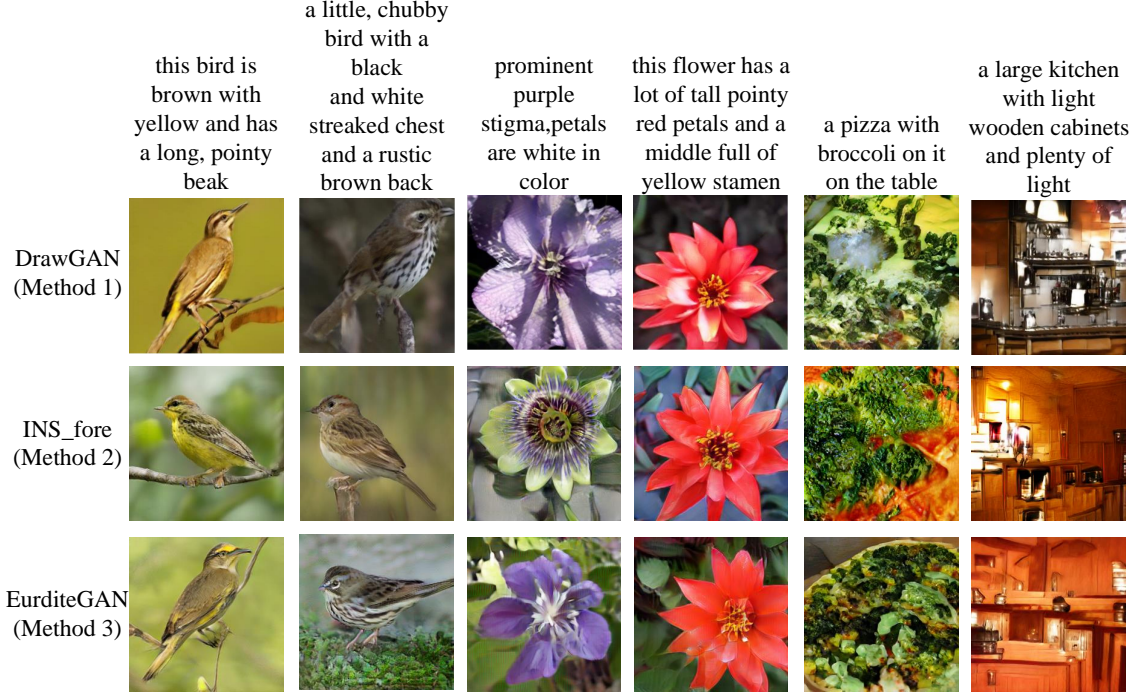


Figure 2.28: The qualitative comparison results among our proposed T2I methods, including DrawGAN, INS\_fore, and EruditeGAN.

## 2.7 Internal Comparison of Proposed Methods

### 2.7.1 Qualitative Comparison

In order to improve the quality of T2I synthetic images, we have proposed three methods: DrawGAN: Text to Image Synthesis with Drawing Generative Adversarial Networks; Text-to-Image Synthesis: Starting Composite from the Foreground Content; Text to Image Synthesis with Erudite Generative Adversarial Networks., respectively expressed as DrawGAN, INS\_fore, EruditeGAN.

The qualitative comparison results of the three proposed methods are shown in Fig. 2.28. From the generated results, the synthetic results of each method have good quality, which reflects the effectiveness of our proposed methods in improving the quality of image synthesis. In contrast, the synthesis effect of INS\_fore is better than that of DrawGAN, which shows that the method of first synthesizing forward content

Table 2.19: The quantitative comparison results of IS and FID on the CUB, Oxford-102 flower, and MS COCO datasets among our proposed methods.

Method	CUB			Oxford-102			MS COCO		
	IS	FID	R	IS	FID	R	IS	FID	R
DrawGAN	4.76	9.87	<b>77.99</b>	4.07	20.24	77.70	31.11	31.51	89.20
INS_fore	<b>4.79</b>	<b>9.29</b>	77.86	<b>4.19</b>	18.96	79.32	30.70	29.55	90.36
EruditeGAN	4.69	9.58	77.62	4.07	<b>17.69</b>	<b>80.25</b>	<b>31.94</b>	<b>28.79</b>	<b>91.26</b>

based on text information and then synthesizing the final result is better than the method of gradually synthesizing contour, foreground, and final content. The reason is that the text information describes the specific content that needs to be synthesized, but the contour information does not match it. Therefore, in DrawGAN, the method of synthesizing contour information based on the text at first increases a little difficulty of the whole synthesis task, resulting in limited quality improvement. In comparison, the foreground information is matched with the text information, so the INS\_fore method makes the whole synthesis task simpler and can achieve better quality improvement.

For INS\_fore and EruditeGAN, there is no obvious difference in image synthesis quality, which indicates that both approaches achieve satisfactory quality improvement.

## 2.7.2 Quantitative Comparison

The quantitative comparison results of the three proposed methods are shown in Table. 2.19. Overall, the results of INS\_fore and EruditeGAN methods are slightly better than DrawGAN. Compared with INS\_fore and EruditeGAN, the performance of the EruditeGAN method on the MS COCO dataset is better. The main reason is that for the composition of complex images, both foreground content and background content are important. INS\_fore only focuses on the synthesis of foreground content at the beginning, resulting in limited quality improvement of its final result. EruditeGAN does not have such a problem and can improve the quality of synthesis through multi-class discriminant feedback, so it achieves the best performance in complex image synthesis.

## 2.8 Chapter Conclusions

In this section, we propose three methods (DrawGAN, INS\_fore, and EruditeGAN) to improve the synthesis quality of T2I. For each proposed method, we conduct a detailed introduction, including network structure, loss function, implementation details, and experiments. At the same time, we also compared the three methods we proposed internally and obtained the following conclusions:

- The three methods we proposed all have a good role in improving the quality of T2I image synthesis;
- The overall performance of INS\_fore and EruditeGAN is slightly better than DrawGAN;
- In complex image synthesis, EruditeGAN achieves the best synthesis performance.

In the next chapter, we will introduce the high controllability oriented image synthesis methods.

# Chapter 3

## High Controllability Oriented Image Synthesis Methods

### 3.1 Introduction

In order to make the image generation structure more valuable, it is necessary to provide high-level control information. The current research mainly starts from two directions: one is to control the shape of synthesis, and the other is to control the content of synthesis. The main form of shape control is to enter a profile, such as a simple outline of a shoe or bag. Then the input contour is used to synthesize the image. The biggest problem of this method is that only shape information can be controlled, but not the specific details. For example, if input the contour of a package, this method cannot determine the color information of the package in the synthesis result. In the related model [51], [52], the specific details (such as color) are determined by the image in the training set. If the training set has a yellow packet, it is possible to synthesize the yellow packet based on the contour of the packet. However, if there is no blue package in the training set, the model cannot synthesize the blue package. This reflects that the degree of control for this approach is limited.

The method to control the content of synthesis starts with the use of text information control. At first, conditional GAN (CGAN) [2], [53] used the category attributes of images (such as flower and bird) to control the categories of image synthesis. This method can only control the category of the composite content, but not for more specific details. For example, if the category label is a bird, the model can synthesize a bird image, but the color, size, and other information of the bird cannot be controlled.

Furthermore, Reed *et al.* [3] proposed image synthesis based on text description information (like “this bird is black with white and has a long, pointy beak”), which makes the whole synthesis process more flexible and conforms to human input habits. This approach demonstrates great flexibility and more control over the content. Since it is more conform to people’s input habits, it has better application prospects because the current research of artificial intelligence is more inclined to serve people. Nevertheless, the text description controls both the object and detailed information, but for the shape, size, and position of the object, it seems to be ineffective. For the research of image synthesis based on the text description, many works have been done, and encouraging results have been achieved. However, none of these works can control the shape, size, and position of the synthesized object.

To alleviate this problem and achieve better control of the synthesis details, Reed *et al.* proposed the Generative Adversarial What-Where Network (GAWWN) [42], using the bounding box and the key points to determine the location and shape of the target, and then generated specific content based on the text description. GAWWN is more flexible and controllable. On the one hand, the input information (bounding box, key point, and text description) can be determined artificially. On the other hand, the overall control degree is higher than that of only using the text description. Although GAWWN has achieved some success, it has two obvious problems. Firstly, the authenticity of the result is comparatively poor. Secondly, the control implemented by using the bounding box or key points is relatively rough, which does not achieve the real fine-grained control effect.

## 3.2 Related Works

Compared with the traditional research of image processing, image generation is more challenging. Mansimov *et al.* [54] proposed the alignDRAW model, which is an extension of the Deep Recurrent Attention Writer (DRAW) [55] model, by learning to

estimate alignment between generating results and text. Autoregressive models [56], [57] obtained arresting results by modeling the conditional distribution of pixel space using the neural network. [58], [59] realized image synthesis by using the deterministic network as function approximation. Variational Autoencoders (VAE) [12], [60] defined the generation problem as a probability graph model and achieved the final generation by maximizing the lower bound of data likelihood. Besides these, the best overall performance ability is Generative Adversarial Networks (GAN) [1][14][61][13][62][63]. It has shown encouraging image generation results. Because of the instability of training, many improvement works have been proposed to stabilize the training process and improve the quality of synthesis.

In order to make the generative image model useful, conditional image synthesis has been explored. The initial condition generation is based on simple image attributes or class labels [2], [53], which has achieved some better results, but it is not suitable for human basic input habits because it may require some professional knowledge. Besides, using property or category labels can not control the details. After that, there are some works of image generation conditioned on the image (pixel to pixel), including image super-resolution [64], [21], image editing [24]–[65], image style transfer [51], [66], [18]. Since the image is as the input, the overall content cannot be changed greatly, which limits the artificial control factors to a certain extent. In these works, there is a way of simple input and strong control, that is to utilize simple contour to synthesize image. This method is more practical than using labels because it fixes the basic shape of the synthetic image. Nevertheless, it can only control the shape and but not detailed information. At present, the image generation, which accords with the habit of human input, is using text description to synthesize images. Reed *et al.* [3] first implemented text-to-image synthesis using the end-to-end GAN architecture based on adversarial learning, which generated realistic images. Subsequently, Zhang *et al.* [4], [5] proposed StackGAN to generate more realistic results through multi-stage

adjustment. Xu *et al.* [6] used the attention mechanism to make local fine-tuning to obtain better results. Based on the attention mechanism, Qiao *et al.* [8] and Zhu *et al.* [10] respectively utilized text reproduction and dynamic memory to improve the quality of results further. Zhang *et al.* [7] proposed a hierarchical nesting structure, and could generate larger and more vivid images. Qiao *et al.* [9] employed prior knowledge to improve the quality of synthetic images further. Its prior knowledge is obtained from the result with the mask. Although the results of text-to-image synthesis are more real and more abundant, there is the same problem — for the same text description, the model can generate a variety of results that conform to the text description but have different shapes, sizes, and orientations, which means that the input text can only control the generated content, but not the specific shape. This problem makes the current image synthesis model based on text description less practical.

For better flexible and effective control, based on the text description, Reed *et al.* [42] proposed the GAWWN structure and realized the controllable image generation process for the first time by combining the object location and other annotations. The size and position of the object are determined by inputting the bounding box and key points information. No matter the bounding box, key points, or text description, GAWWN can be input artificially, which makes GAWWN have pretty practicability. However, their results are not satisfactory as well as the bounding box, and key points are rough information, which cannot accurately determine the specific shape of the object.

### **3.3 Method 1 — Customizable GAN: A Method for Image Synthesis of Human Controllable**

To achieve better fine-grained control and generate more authentic results, we propose a customized GAN. The image is generated by combining the contour and text descrip-

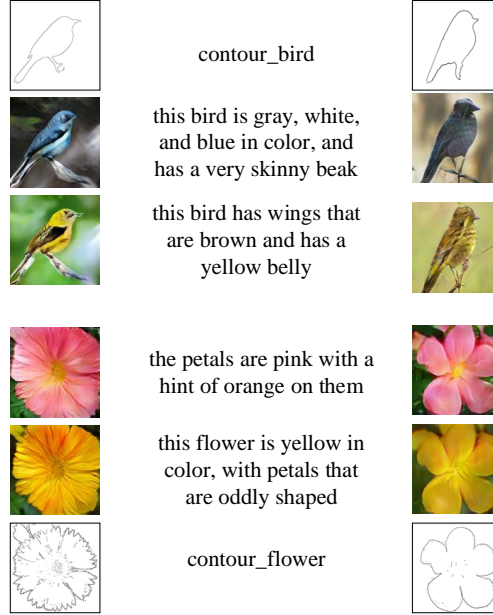


Figure 3.1: The results of the corresponding birds and flowers under different texts and contours. They are consistent with the corresponding text description while retaining the contour shape. The left contours are obtained by pre-processing the original dataset. The contours on the right are drawn by hand.

tion, as shown in Fig. 3.1. The contour is used to determine the specific shape, size, and position information of the object. Then the text description is used for generating the specific content. Finally, high-quality images based on the hand-drawing contour and artificial text description are obtained by our method. It realizes fine-grained control while also completing the generation of the realistic image.

### 3.3.1 Network Structure

#### Structure diagram

The architecture of our method is shown in Fig. 3.2 and 3.3. It is built upon conditional GAN framework conditioning on both contour and text description. Fig. 3.2 shows the network structure of the generator. In the generator, the contour and text description in the input is encoded in different ways and combined together, and then the corresponding result is synthesized by de-convolution [36].

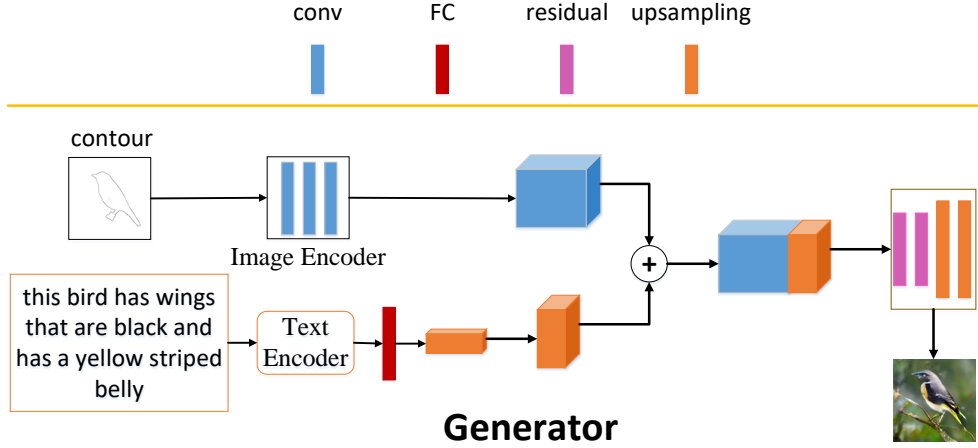


Figure 3.2: The generator structure of the model. The generator synthesizes the corresponding image based on the text description and contour. The synthesized image not only retains the contour shape but also conforms to the text description information.

Specifically, in the generator, we use a convolutional neural network with three convolutional layers as an image encoder to extract contour features. In addition, in specific experiments, we also try to use VGG16 or VGG19 as a pre-trained image encoder [67] to extract the corresponding contour feature (See Section 3.3.4 for details). The text description is encoded as a text vector by the pre-trained text encoder [68], and then its dimension is changed to 128 through a fully connection (FC). Referring to the work of Zhang *et al.* [4], conditional augmentation (CA) has also been added to increase the number of text embeddings. The corresponding content of the conditional augmentation is shown near Eq. 2.10.

In order to combine text embeddings with feature extraction from the contour, spatial replication is performed to expand the dimension of text embeddings. Finally, the dimension of the contour extraction feature is  $16 \times 16 \times 512$ , and the dimension of text embeddings is  $16 \times 16 \times 128$ . After the connection, it will pass through two residual transformation units, which are composed of residual blocks [37]. Accordingly, the employment of residual blocks is to make the connection features more effective through deeper layer processing. On the other hand, it can better learn the feature representations to ensure the contour of the generated image is consistent with the input

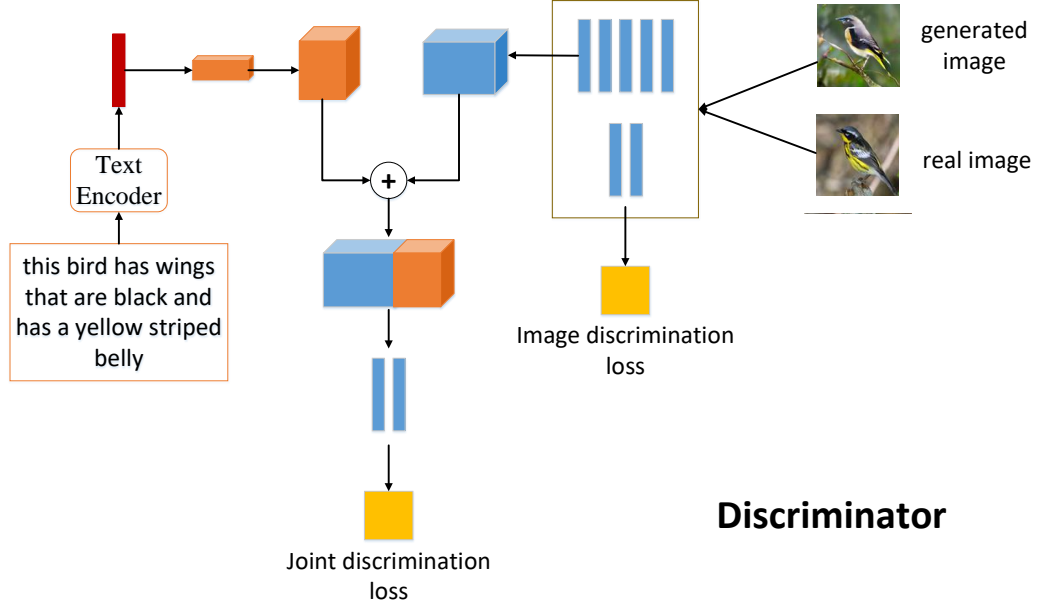


Figure 3.3: The discriminator structure of the model. The discriminator judges whether the received image itself is real or generated and the matching degree between the image and text.

contour. Finally, the generator synthesizes the corresponding result by up-sampling.

The discrimination process consists of two parts: the discrimination of real or generated image and of the consistency of image and text description. Fig. 3.3 shows the network structure of the discriminator. In the discriminator, there is feature extraction of the input image through down-sampling. There are two kinds of down-sampling. One is used to distinguish the real or generated image. The other is to distinguish the consistency of the image and text. The down-sampling for the real or generated image discrimination consists of two convolution layers: the first layer is followed by BN [39] and leaky-ReLU [41], the second layer is directly followed by the sigmoid function. For the discrimination of the consistency of the image and text, the image features are first extracted through five convolution layers, then combined with the text vector of extended dimension, and finally identified by two convolution layers. Each convolution layer is followed by BN and leaky-ReLU, except for the last layer for discrimination. Unlike the generator, the features extraction dimension in the discriminator is  $4 \times 4 \times 512$ , and the text embeddings dimension is  $4 \times 4 \times 128$ .

## Training algorithm

Customizable image synthesis determines the shape through the contour and defines the specific content through the text description. This indicates that the result of the synthesis should match the basic shape of the input contour as well as the text description. We utilize the method of adversarial learning to train the whole process, as shown in algorithm 2.

---

### Algorithm 2 CustomizedGAN training algrothim

---

```

1: Input: matched text  $T$ , mismatched text  $T_{mis}$ ,
2:         relevant text  $T_{rel}$ , real image  $I_{real}$ ,
3:         contour  $con$ , number of epochs  $N$ 
4: for  $n = 1$  to  $N$  do
5:    $s = \phi(T)$ 
6:    $s_{mis} = \phi(T_{mis})$ 
7:    $s_{rel} = \phi(T_{rel})$ 
8:    $I_{gen} = G(con, s)$ 
9:    $d, d_{ucond} = D(I_{real}, s)$ 
10:   $d_{gen}, d_{gen\_ucond} = D(I_{gen}, s_{rel})$ 
11:   $d_{mis}, d_{mis\_ucond} = D(I_{real}, s_{mis})$ 
12:   $L_{D\_real} = \log(d) + \log(d_{ucond})$ 
13:   $L_{D\_mis} = (\log(1 - d_{mis}) + \log(d_{mis\_ucond}))/2$ 
14:   $L_{D\_gen} = (\log(1 - d_{gen}) + \log(1 - d_{gen\_ucond}))/2$ 
15:   $L_D = L_{D\_real} + L_{D\_mis} + L_{D\_gen}$ 
16:   $D = D - ss * \partial L_D / \partial D$ 
17:   $L_G = \log(d_{gen}) + \log(d_{gen\_ucond})$ 
18:   $G = G - ss * \partial L_G / \partial G$ 
19: end

```

---

There are three types of text input in the training process, that is, the matched text  $T$ , the mismatched text  $T_{mis}$ , and the relevant text  $T_{rel}$ .  $T$  represents the text that matches the real image, and  $T_{mis}$  represents the text that does not match the real image.  $T_{rel}$  represents the text related to the semantics of the generated image. The three types of texts and the corresponding images form three types of image-text pairs for the discriminator to discriminate to improve the discriminator's discrimination ability in the image-text consistency so that the final image generated by the generator

is more match the input text. In the algorithm,  $\phi$  is a pre-trained text encoder [68] used to encode text into vectors. The generator synthesizes generated images based on the input contour and text. The discriminator distinguishes three situations: the real image with the matched text, the generated image with the relevant text, and the real image with the mismatched text. The purpose of introducing text that does not match the real image is to make the whole network learn the situation of mismatch so that it can improve the final matching degree. Unlike the general GAN, the discriminator in our algorithm returns two outputs: first is the degree of the image-text matching, and second is the judgment of the authenticity of the image. The advantage of this method is to distinguish the results from many aspects to improve the discrimination ability of the discriminator.  $G$  and  $D$  are updated by the SGD method, where  $ss$  is the step size.

### 3.3.2 Loss Function

According to the training algorithm, the specific loss functions are as follows:

$$\begin{aligned}
L_D = \sum_{I_{real} \sim p_{data}, I_{gen} \sim p_G} \{ & \log D_0(I_{real}, T) + [\log(1 - D_0 \\
& (I_{real}, T_{mis})) + \log(1 - D_0(I_{gen}, T_{rel}))]/2\} \\
& + \{\log D_1(I_{real}, T) + [\log D_1(I_{real}, T_{mis}) + \\
& \log(1 - D_1(I_{gen}, T_{rel}))]/2\}
\end{aligned} \tag{3.1}$$

$$L_G = \sum_{I_{gen} \sim p_G} \log D_0(I_{gen}, T_{rel}) + \log D_1(I_{gen}, T_{rel}) \tag{3.2}$$

where  $D_0$  represents the first output of the discriminator and  $D_1$  represents the second.  $I_{real}$  represents the real image conforming to the image distribution of the original dataset, and  $I_{gen}$  represents the generated image conforming to the distribution of the generated image. In  $L_D$ , the content of the first brace represents the unconditional loss (discriminating the authenticity of the image), and the second brace content represents

the conditional loss (determining whether the image and text match).  $\log$  represents log means logarithmic function.

### 3.3.3 Implementation Details

In the training process, the initial learning rate is set to 0.0002, and it decays to half of the original every 100 epochs. Adam optimization [38] with a momentum of 0.5 is used to optimize and update parameters. A total of 600 epochs are trained iteratively in the network, of which the batch size is 10. In leaky-ReLU, the leaky value is 0.2.

For the image encoder used to extract contour features in Figure 3.2, there are three cases in total: 1) use three convolutional layers directly (denoted as ‘without VGG’); 2) use the pre-trained VGG16 model (denoted as ‘with VGG16’); 3) use the pre-trained VGG19 model (denoted as ‘with VGG19’). In the following content, ‘Customizable GAN’ and ‘ours’ both represent the first case; ‘Customizable GAN(with VGG16)’ and ‘ours+VGG16’ both represent the second case; ‘Customizable GAN(with VGG19)’, ‘ours+VGG19’ both represent the third case.

### 3.3.4 Experiments

#### Qualitative results

Firstly, we compare the existing text-to-image synthesis model. The existing T2I model has two main directions. One is based on the multi-stage synthesis, and the other is based on the attention mechanism. AttnGAN [6] not only uses multi-stage synthesis but also is based on the attention mechanism, so we choose AttnGAN as the representative model for comparison. The specific results are shown in Fig. 3.4. From the comparison results, it can be seen that for the same text description, our model not only conforms to the text description but also can control the shape of the final synthesized object through simple contour. For AttnGAN, multiple results can be synthesized, but the shape, size, and position of the synthesized object are different,

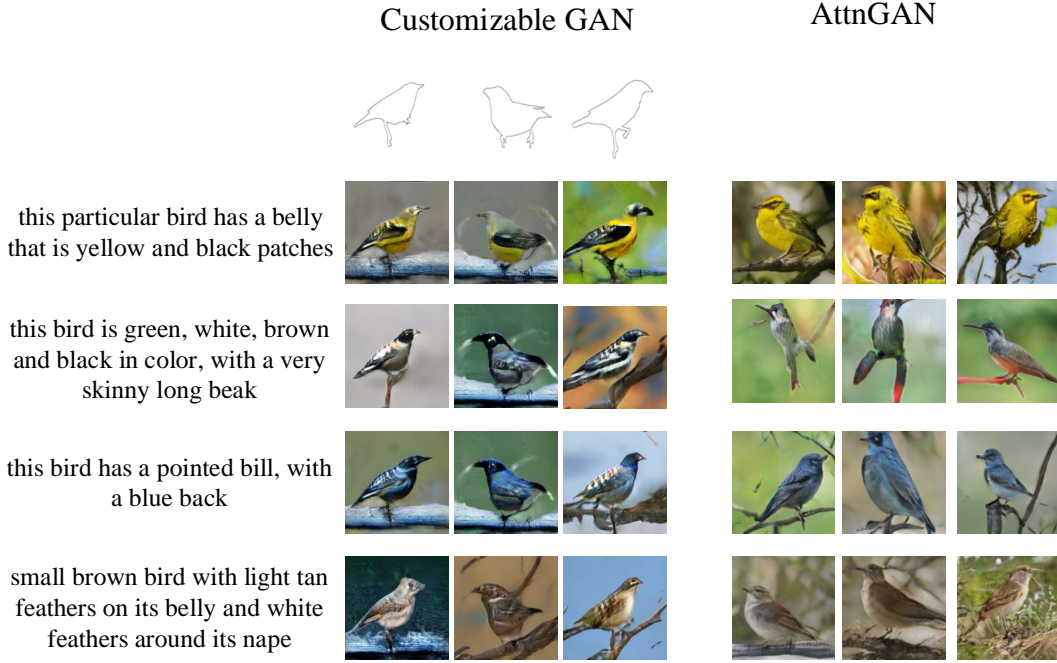


Figure 3.4: The comparison between our method and the existing text-to-image synthesis method. The text-to-image synthesis model can not control the contour information of the synthesized object, and we can control the specific contour of the object while conforming to the basic text description information.

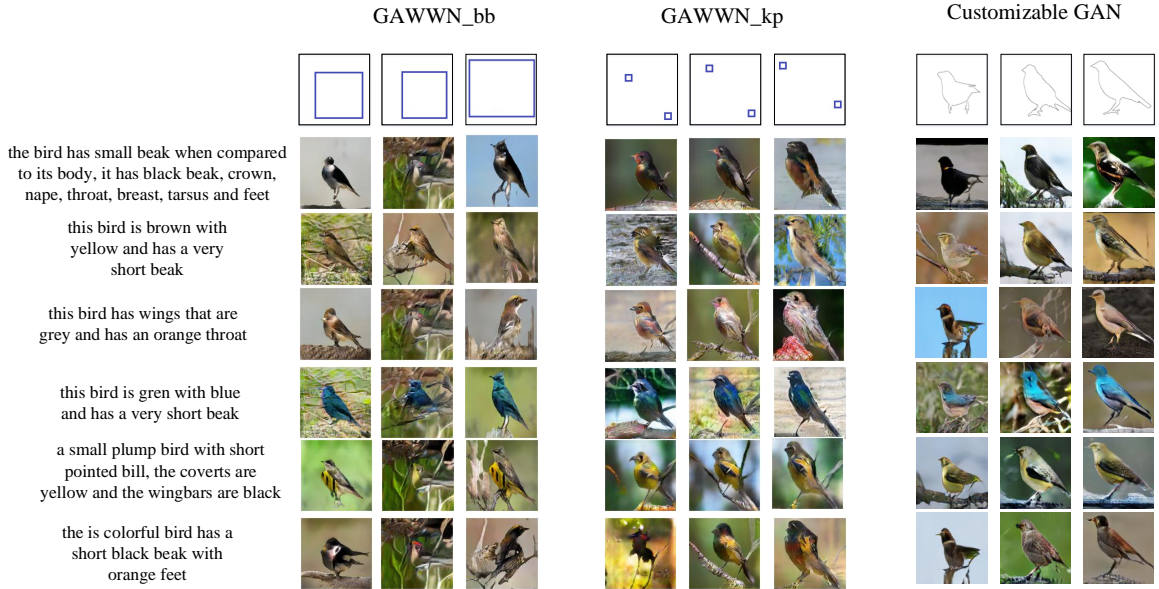


Figure 3.5: The comparison between our method and GAWWN (including two results based on bounding box and key points). It can be seen from the comparison results that our results are obviously superior to GAWWN and have a better degree of control than GAWWN.

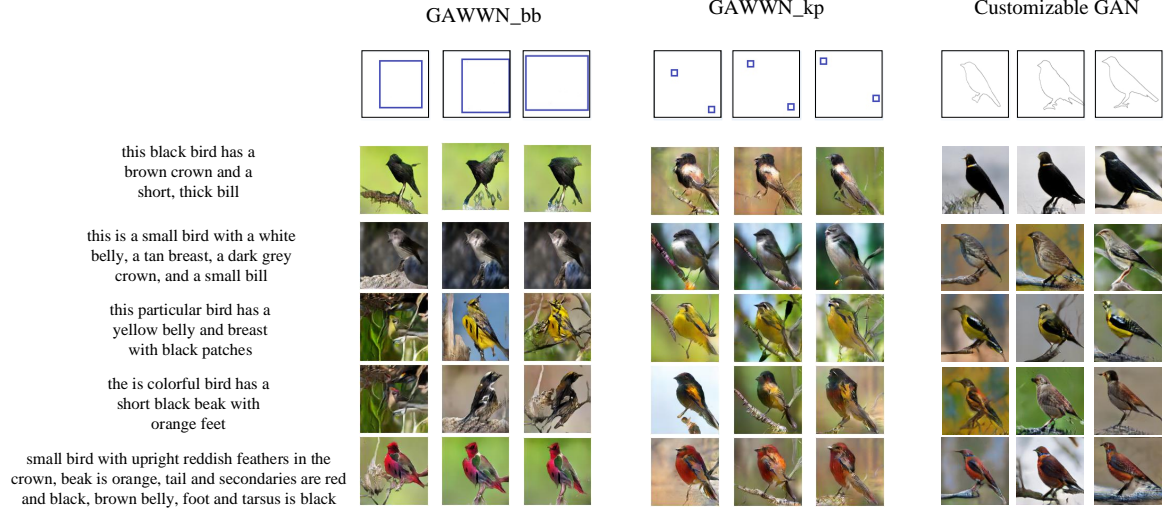


Figure 3.6: The comparison results of the second group between our method and GAWWN. These results also reflect the roughness control and poor authenticity of GAWWN. In contrast, our results are more realistic, and the degree of control is also more refined.

which indicates that the existing T2I model can not control the specific style. This reflects the low practicability of the existing T2I model. Compared with the existing T2I model, GAWWN [42] and our method are all studying in the direction of more effective image synthesis control. Meanwhile, the input of GAWWN can also be artificially controllable. Therefore, we choose to compare our method with GAWWN.

Compare our method with the existing controllable image synthesis based on text and annotations (GAWWN), as shown in Fig. 3.5 and 3.6. There are two kinds of comments in GAWWN: the bounding box, and the key point information. In the figure, GAWWN\_bb represents the GAWWN result based on the bounding box. The input bounding box can only control the generated area, which is powerless for the specific shape and orientation. GAWWN\_kp represents the corresponding result based on the key points. Key points control the basic position and orientation of the generation, but the specific shape cannot be determined. Simultaneously, the synthesis results based on the bounding box and key points generally have poor authenticity. All these shows that indicate although GAWWN has high flexibility in image synthesis, its overall control is relatively poor and rough, and the results of synthesis are not satisfactory.

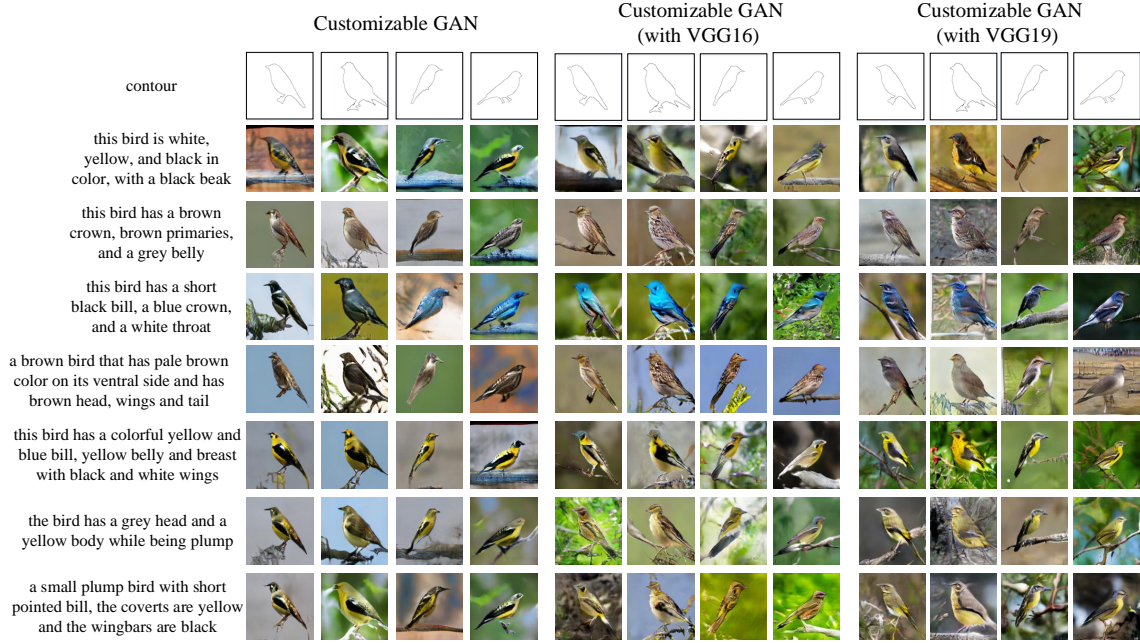


Figure 3.7: The comparison bird results of our method without VGG and with VGG16, with VGG19. It can be seen that the results of using VGG are better in details (such as eyes, pecking).

Compared with GAWWN, our method has a higher control ability, evidenced by the specific shape, position, and orientation of the synthesized results. This shows more fine-grained control than GAWWN’s rough control and realizes the genuinely customized image synthesis. For the generated results, on the one hand, our method maintains consistency with the input contour and text description. On the other hand, it is better than GAWWN in authenticity. This demonstrates the superiority of our method in controlling the generation of authentic results.

In addition to the comparison with GAWWN, we also made an internal comparison. In this paper, we compared the three methods of contour feature extraction without VGG, with VGG16, and with VGG19, as shown in Fig. 3.7. The results of the three methods have a high degree of authenticity. They maintain both the shape of the input contour and match the content of the text description. From a more detailed level (eye, pecking, texture) of comparison, the results obtained by using VGG are better than those not applicable to VGG, which makes the results of using VGG have

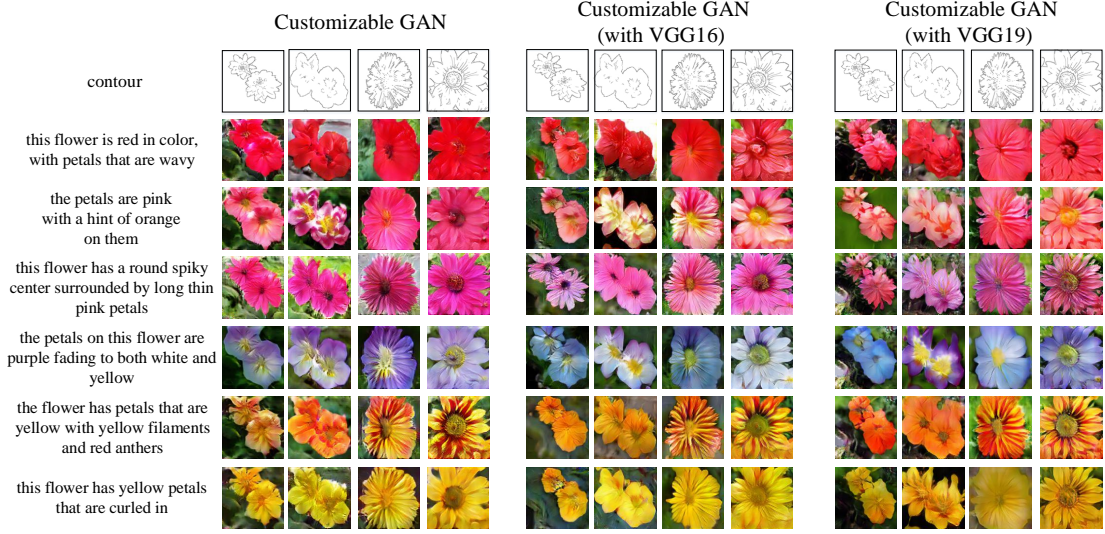


Figure 3.8: The comparison flower results of our method without VGG and with VGG16, with VGG19.

pretty authenticity. Compared to VGG16, VGG19 handles the details of the texture better to make the results more realistic.

We extended our method on the flower dataset and made the internal comparison, as shown in Fig. 3.8. The results of the three methods are authentic. Overall, all results maintain the shape of the contour and conform to the text description. In comparison, VGG16 has a higher degree of agreement with the contour because it reflects better the overall details of the contour, which makes its results have higher authenticity.

## Quantitative results

For the evaluation of the generation model, Human Rank (HR) is used to quantify the comparison models. HR can be used to evaluate whether the synthesized image conforms to subjective effects (such as authenticity, matching degree with text, etc.), and it is widely used in various image synthesis works, such as [24], [4], [5].

In this work, we employed 10 subjects to rank the quality of synthetic images by different methods. The text descriptions and contours corresponding to these results are all from the test set and are divided into 10 groups for use by 10 subjects. For

the bird datasets, we established two ways for quantitative comparison. One of them contains three results: 1) GAWWN\_bb, 2) GAWWN\_kp, and 3) ours without VGG. The other includes five synthetic results: 1) GAWWN\_bb, 2) GAWWN\_kp, 3) ours without VGG, 4) ours with VGG16, and 5) ours with VGG19. In this way, the comparison of the bird is three tuples (bird\_1) and five tuples (bird\_2), respectively. The employers were not informed of the method corresponding to the result, but only knew the text description and contour, bounding box, and key points corresponding to the current result. The subjects were asked to rank the results (bird\_1: 1 is best, 3 is worst; bird\_2: 1 is best, 5 is worst) in the following ways:

- Whether the result is highly consistent with control information (the contour or bounding box or key points);
- Whether the result matches the text description;
- The level of the authenticity of all results.

The average score will be calculated for the ranking results of all subjects, as shown in Tables 1 and 2. The comparative results show the following points:

Table 3.1: The results of quantitative comparison between our three methods and GAWWN. It includes three aspects of comparison: one is the consistency with the control information (consistency), the other is the matching with the text content (text), and the third is the authenticity of the results (authenticity).

	GAWWN_bb	GAWWN_kp	ours	ours+VGG16	ours+VGG19
consistency	4.64	4.18	2.12	2.00	2.02
text	4.10	3.80	2.40	2.34	2.30
authenticity	4.46	3.90	2.37	2.20	2.07

**More authenticity and better text matching.** Compared with GAWWN in Tables 3.1 and 3.2, it is obvious that our method has higher authenticity and degree of text matching. In comparison, the results of using key points (kp) in GAWWN are

Table 3.2: The quantitative comparison results between our method with GAWWN in CUB dataset.

	GAWWN_bb	GAWWN_kp	ours
consistency	2.78	2.51	1.27
text	2.46	2.28	1.44
authenticity	2.67	2.34	1.42

better than those of using the bounding box (bb). However, compared with our results, the overall authenticity and matching of GAWWN\_kp are still worse than ours.

**More effective control.** In the process of image synthesis, the control of our method is more effective since it shows better consistency with the control information. The control degree of GAWWN\_kp is better than that of GAWWN\_bb. This is consistent with the subjective comparison. In subjective results, GAWWN\_kp can control the basic direction of synthesis, but GAWWN\_bb cannot. Compared with GAWWN\_kp, our method has more excellent control. The reason for this circumstance is that our results can not only control the synthesis direction but also control the specific shape, while GAWWN\_kp can not control the shape.

**Better performance when using VGG.** Table 1 shows that the results obtained by our three methods (without VGG, with VGG16, with VGG19) are not significantly different. In close comparison, the results of using the VGG model are better than those of not using VGG. This reflects that VGG can extract better contour features and promote the synthesis of final results.

Table 3.3: The internal quantitative comparison results of our methods in CUB dataset.

	ours	ours+VGG16	ours+VGG19
consistency	1.27	1.20	1.21
text	1.44	1.40	1.38
authenticity	1.42	1.32	1.24

Table 3.4: The internal quantitative comparison results of our methods in Oxford-102 flower dataset.

	ours	ours+VGG16	ours+VGG19
consistency	1.25	1.12	1.23
text	1.23	1.15	1.23
authenticity	1.28	1.15	1.17

### Ablation Study

It can be found in Table 3.1 that among the results of birds, VGG19 is better than VGG16, and VGG16 is better than not using VGG. Does this phenomenon also apply to flower results? What are the differences between not using VGG and using VGG 16 and VGG 19 and the reasons behind the differences? To solve these problems, we conducted an ablation study.

For the internal comparison of our three methods, it can be seen from Table 1 that there is no obvious difference. In the separate comparison, the result of using VGG is better than that of not using VGG. In Tables 3.3 and 3.4, among the results of birds, the overall authenticity of VGG19 is better than that of VGG16, while that of flowers is the opposite. The reason for this is that the proportion of birds in the image is relatively small (generally less than 50%), so the judgment of the authenticity of bird image is more dependent on the generation of bird details. VGG19 performs the best authenticity in generating bird results, which shows that it does best in detail generation. Compared with bird images, the proportion of flowers in the image is generally more than 80%, so its authenticity depends on the overall structure. In the authenticity of flower results, VGG16 is better than VGG 19, which indicates that VGG 16 does the best performance in structural consistency. Although VGG19 can obtain pretty detailed information in flower results, the authenticity of VGG16 results is better because flowers pay more attention to integrity. VGG16 also showed the best structural consistency in birds results, indicating that VGG16 is indeed better than

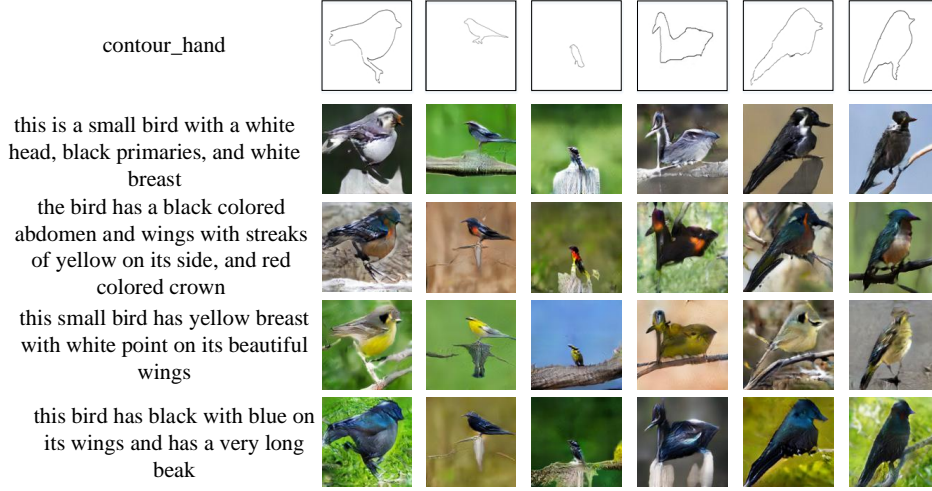


Figure 3.9: The text descriptions on the left are all artificial descriptions that do not exist in the dataset. The contours are also drawn manually. The results show the effectiveness of our method in generating high-quality results and high flexibility in image control generation.

VGG19 in terms of structural consistency.

On the whole, VGG19 is better than VGG16 in detail synthesis, and VGG16 is better than VGG19 in overall structure synthesis. This is reasonable because VGG19 is deeper than VGG16, so it can extract more detail-oriented feature information. The number of layers of VGG16 is relatively small, so it pays more attention to the overall feature information. VGG is a network structure specially designed for feature extraction, which performs well in classification, segmentation, and other tasks. Therefore, the use of VGG is better than the simple use of convolution operation (without VGG) to extract features, so the final performance is better.

### Controllable image synthesis

The most important feature of our work is to realize fine-grained controllable image synthesis based on artificial hand drawing and manual description. The relevant results are shown in Fig. 3.9 and 3.10. Both the contour and the text description in the figure are artificial and do not exist in the dataset. Besides, it can also be seen from the results that our model can generate corresponding high-quality results for the different

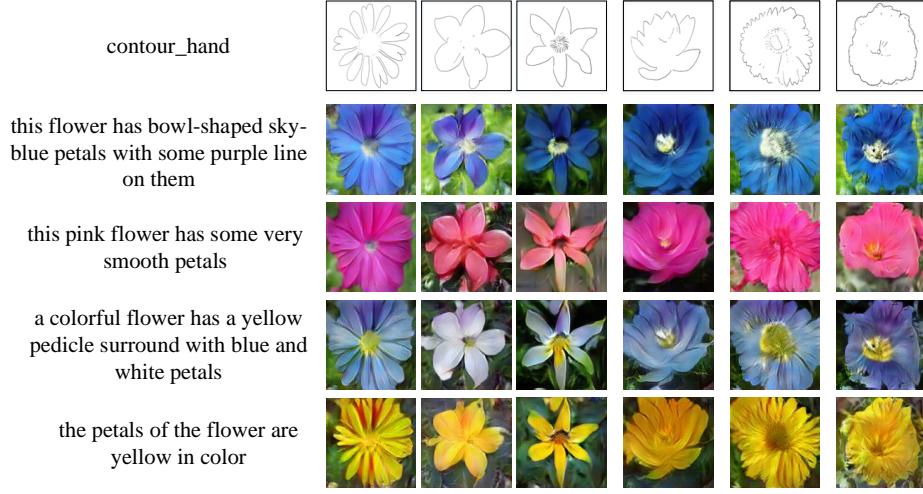


Figure 3.10: The text descriptions on the left are all artificial descriptions that do not exist in the dataset. The contours are also drawn manually. The results show the effectiveness of our method in generating high-quality results and high flexibility in image control generation.

contours of shapes, sizes, positions, and orientations. Such as shown in the bird results in Fig. 3.9, the first, second, and third columns well show that the model can synthesize high-quality results based on different contour sizes and positions. At the same time, the fourth, fifth, and sixth columns also show that the model can adapt to different contour orientations and generate high-quality results. The flower results in Fig. 3.10 also reflect that the model can adapt to different contour shapes, sizes, and orientations and generate high-quality flower results. These results not only reflect well the hand-drawn contour and artificial text description content but also have a high degree of authenticity. This demonstrates the effectiveness of our method in synthesizing high-quality authentic images and shows the high flexibility and controllability of our method because all inputs can be controlled artificially.

### Complex image synthesis

In addition to generating images of birds and flowers, we also test the synthesis performance of our proposed method on complex images. Specifically, we use the pre-trained

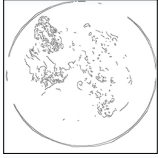

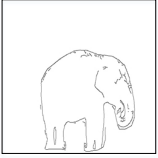





Input Text	A bowl with rice, broccoli and a purple relish	Two zebras who are grazing on some grass	An elephant picks up an object with its trunk	A very large commercial plane flying in blue skies
Input Contour				
Generated results				

Figure 3.11: Based on text and contour information, the complex image synthesis results are shown above.

VGG19 model for contour feature extraction and train our designed network structure 200 epochs in the MS COCO dataset. Fig. 3.11 shows the generation results of our proposed method on complex images. Overall, the generated results are terrible. We can find that the content of generated zebras, elephant and airplane is very abstract, and the content of generated food is also very mediocre. Such poor results on complex images show that our proposed method still has a lot of room for improvement.

### 3.4 Method 2 — TCGIS: Text and Contour Guided artificially controllable Image Synthesis

Our proposed CustomizableGAN achieves a more controllable image synthesis effect by using text and contour information, where the text information is used to control the component content, and the contour information is used to control the basic shape, size, and position information of the synthetic object. Although CustomizableGAN achieves a more controllable image synthesis effect, however, according to the results in Figures 3.9, 3.10, and 3.11, it can be found that CustomizableGAN still has a large room for improvement in image synthesis quality, especially in complex image synthesis.

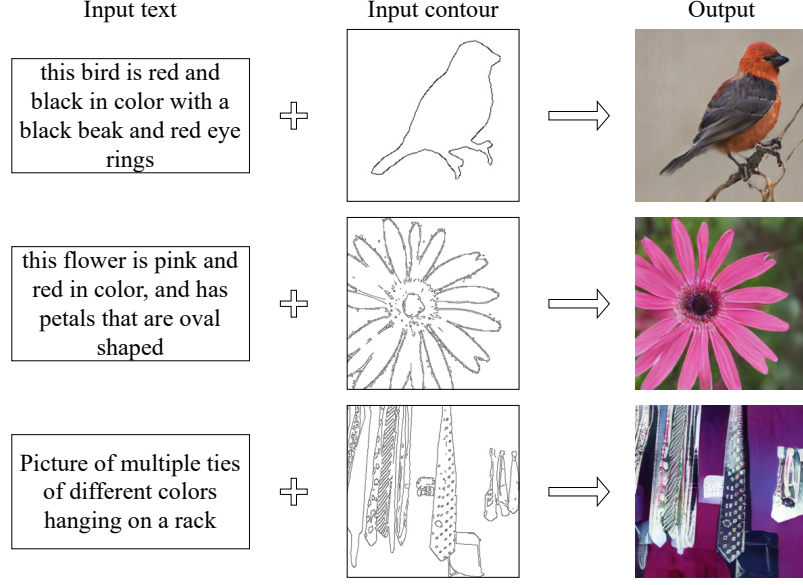


Figure 3.12: The basic structure of our proposed method is shown in the figure above. By deeply fusing textual and contour information, our proposed structure finally synthesizes high-quality and satisfactory image results.

In order to solve the problem of insufficient synthesis quality in CustomizableGAN and achieve a controllable and high-quality image synthesis effect, we refer to the structure design of T2I and design a more complicated network structure to achieve higher quality image synthesis. In addition, our designed network structure still accepts text and contour information as input to make it highly controllable. Figure 3.12 shows some synthetic results achieved by our designed structure. The results demonstrate that our designed architecture achieves a highly controllable and higher-quality image synthesis effect.

### 3.4.1 Network Structure

The structure of our proposed method is shown in Figure 3.13. For the input text description and contour information, they are encoded as corresponding features by the text encoder [30] and image encoder [31], respectively. The text features include global sentence features ( $s$ ) and local word features ( $w$ ). The encoded contour features ( $c$ ) and sentence features are first fused together, and then the initial fusion features

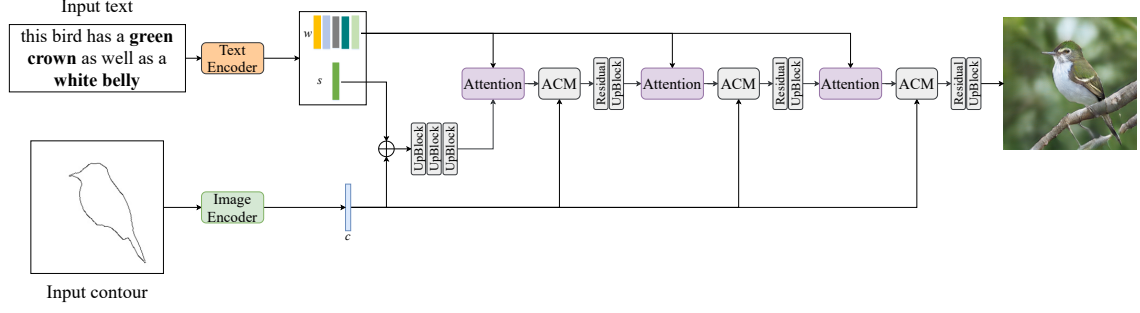


Figure 3.13: The basic structure of our proposed method is shown in the figure above. By deeply fusing textual and contour information, our proposed structure finally synthesizes high-quality and satisfactory image results.  $w$ ,  $s$ , and  $c$  represent word features, sentence features, and contour features, respectively.

are obtained through multiple consecutive upsampling operations. After that, the word features first perform attention fine-tuning on the initially fused features, and then perform in-depth fusion with the fused features through the affine combination module (ACM) [69]. After the processing of residual block [37] and upsampling, two consecutive attention, ACM, residual block, and upsampling operations are repeated and the corresponding image result is finally generated. The reason for repeating the above operations is to deeply fuse text and contour information and gradually increase the resolution of the synthesized result. Finally, the size of the image synthesized by our proposed method is  $256 \times 256$ .

In the structure, the core modules are the attention mechanism and affine combination module (ACM), and their corresponding details are as follows:

**Attention mechanism.** To synthesize higher-quality image results, we employ the spatial and channel attention mechanism [26] to fine-tune the fused features using word features. The equations in the processing are as follows:

$$f = Up(s, c) \quad (3.3)$$

$$m = DT(w) \cdot f \quad (3.4)$$

$$c_{i,j} = \frac{\exp(m_{i,j})}{\sum_{l=0}^{H*W-1} \exp(m_{i,l})} \quad (3.5)$$

$$f'_j = \sum_{i=1}^N c_{i,j} * f_j \quad (3.6)$$

where  $s$ ,  $c$ , and  $w$  represent sentence features, contour features, and word features, respectively.  $Up$  represents the upsampling operation.  $DT$  represents the dimension transformation operation, it can make the dimension of  $w$  match the dimension of  $f$  so that they can perform matrix operation. ‘ $\cdot$ ’ stands for the matrix multiplication operation.  $c_{i,j}$  represents the semantic consistency between the  $i^{th}$  word feature and the  $j^{th}$  region feature in  $f$ .  $\exp$  means exponential function.  $f'$  represents the fusion feature after fine-tuning of the attention.  $H$  and  $W$  represent the height and width of the feature.  $N$  represents the number of the word in the input text.

The above process is the processing process of the first attention in Fig. 3.13. The processing of the last two attentions in the structure is as follows:

$$f'_{acm} = ACM(f', c) \quad (3.7)$$

$$f'_{res} = Res(f'_{acm}) \quad (3.8)$$

$$f'_{up} = Up(f'_{res}) \quad (3.9)$$

$$m' = DT(w) \cdot f'_{up} \quad (3.10)$$

$$c'_{i,j} = \frac{\exp(m'_{i,j})}{\sum_{l=0}^{H*W-1} \exp(m'_{i,l})} \quad (3.11)$$

$$f'_j = \sum_{i=1}^N c'_{i,j} * f'_{up} \quad (3.12)$$

where  $ACM$  represents the affine combination module.  $Res$  denotes the residual block processing operation.  $f'$  represents the updated features.

**Affine combination module.**  $ACM$  takes the fused features after attention and the contour features input at the beginning. It will further fuse these two features, and the

specific equation is as follows:

$$f'_{acm} = f' \otimes W(c) + b(c) \quad (3.13)$$

where  $W(c)$  and  $b(c)$  are the learned weights and bias based on the  $c$ , and  $\otimes$  represents the Hadamard element-wise product.

### 3.4.2 Loss Function

The loss of the whole structure includes the generator and the discriminator's loss. The generator's loss function consists of two parts: adversarial loss and perceptual loss. The equation of adversarial loss in the generator is as follows:

$$L_{G_{adv}} = -\frac{1}{2} \sum_{I_{gen} \sim P_G} \log D(I_{gen}) - \frac{1}{2} \sum_{I_{gen} \sim P_G} \log D(I_{gen}, s) \quad (3.14)$$

where  $I_{gen}$  represents the generated image,  $s$  represents textual information.  $P_G$  denotes the distribution of the generated images.  $\log$  represents log means logarithmic function. The first term refers to the discriminator to distinguish whether the generated image is real or fake. The second term refers to the discriminator to determine whether the generated image matches the input text.

The specific perceptual loss calculation is as follows:

$$L_{per} = \frac{1}{CHW} \|\theta(I_{ori}) - \theta(I_{gen})\|_2^2 \quad (3.15)$$

where  $CHW$  represents the channel, height, and width of the generated image, and  $\theta$  denotes the VGG [67] feature extractor.  $I_{ori}$  and  $I_{gen}$  represent the original image and the generated image, respectively.

In summary, the final loss function of the generator is:

$$L_G = L_{G_{adv}} + L_{per} \quad (3.16)$$

The loss function of the discriminator only includes the adversarial loss, and the specific equation is as follows:

$$L_D = -\frac{1}{2}[\sum_{I_{ori} \sim P_{data}} \log D(I_{ori}) + \sum_{I_{gen} \sim P_G} \log(1 - D(I_{gen}))] \\ -\frac{1}{2}[\sum_{I_{ori} \sim P_{data}} \log D(I_{ori}, s) + \sum_{I_{gen} \sim P_G} \log(1 - D(I_{gen}, s))] \quad (3.17)$$

$P_{data}$  and  $P_G$  denote the distribution of the original images and generated images, respectively. The first term is used to judge the authenticity of the original image ( $I_{ori}$ ) and the generated image ( $I_{gen}$ ), and the second term is used to evaluate whether the original and generated images and text match.

### 3.4.3 Implementation Details

During training, we use ADAM optimization [38] to train 600 epochs for the CUB and Oxford-102 datasets, and 120 for the MS COCO dataset. The initial learning rate is set to 0.0002, and the batch size is set to 10. For image and text encoders, we use a pre-trained image encoder [31] and text encoder [30] to extract corresponding image and text features.

Besides, referring to the work of LightweightGAN [70], we lightweight our designed model. Specifically, we remove the second group of attention, ACM, and residual modules in Figure 3.13 and only keep the upsampling operation.

### 3.4.4 Experiments

#### Qualitative results

Figure 3.14 presents the qualitative comparison results between our method and existing T2I methods. It can be found that the bird results and the complex image results synthesized by the existing T2I methods are poor. The flower results they synthesized are pretty on the whole, but they are still insufficient in detail synthesis, and they cannot effectively control the shape and position information of the synthesized results.

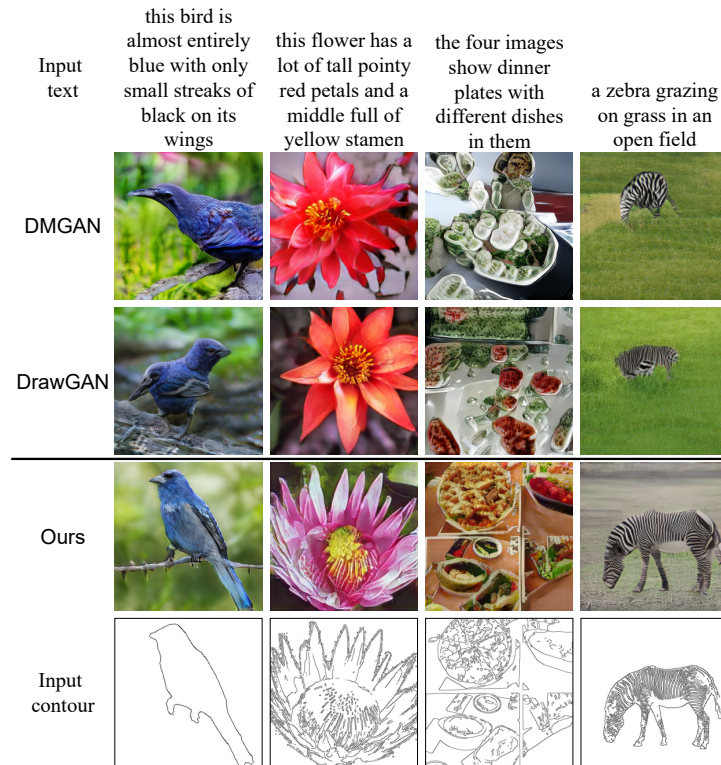


Figure 3.14: The subjective comparison results between our method with existing T2I methods are shown above.

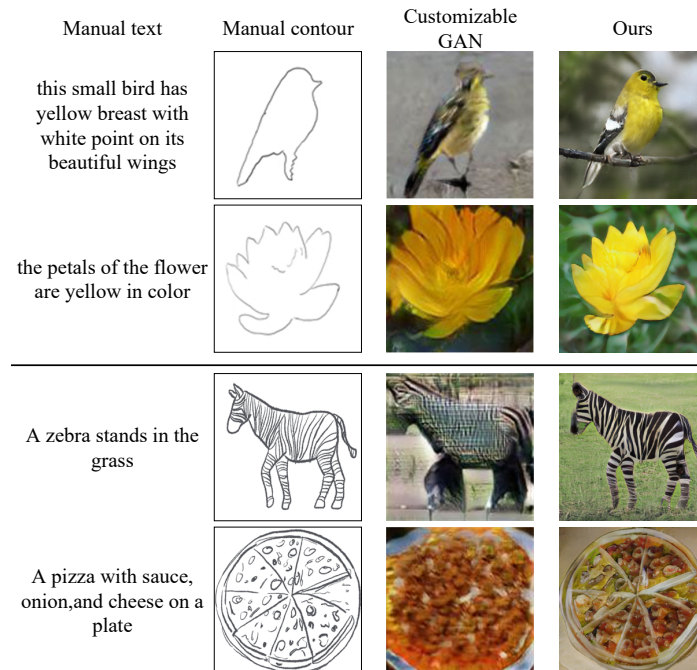


Figure 3.15: The subjective comparison results between our method with CustomizableGAN are shown above.

Table 3.5: The MS-SSIM, SSIM, and FSIM comparison results of our method and existing T2I methods are shown below. ‘lw’ in the table means our lightweight structure.

Datasets	Metrics	AttnGAN [6]	DMGAN [10]	DrawGAN [62]	INS_fore [71]	Ours	Ours (lw)
CUB	MS-SSIM	0.10	0.09	0.09	0.10	0.19	<b>0.21</b>
	SSIM	0.26	0.21	0.23	0.25	0.28	<b>0.31</b>
	FSIM	0.59	0.56	0.58	0.58	0.60	<b>0.62</b>
Oxford-102	MS-SSIM	0.10	0.09	0.09	0.08	0.26	<b>0.27</b>
	SSIM	0.20	0.15	0.16	0.15	<b>0.25</b>	0.21
	FSIM	0.61	0.58	0.59	0.57	<b>0.66</b>	0.63
MS COCO	MS-SSIM	0.08	0.08	0.08	0.09	<b>0.14</b>	0.13
	SSIM	0.18	0.17	0.16	0.16	<b>0.20</b>	0.18
	FSIM	0.56	0.56	0.56	0.56	<b>0.59</b>	0.57

In contrast, our method is able to synthesize highly realistic image results while being consistent with the basic information of text and contour. Especially in complex image synthesis, the results synthesized by our method are subjectively very realistic, while the authenticity of the synthesis results of existing T2I methods is very poor.

Figure 3.15 shows the comparison results of our method and CustomizableGAN. The text and contour information in the figures are inputted manually. Among the synthetic results of birds and flowers, the results synthesized by our method have significantly better clarity and realism, and our synthesized results are more conform to the input contour information. Furthermore, our method is able to synthesize corresponding high-quality complex image results based on the manual input text and complex contour, which demonstrates the generality of our proposed method.

## Quantitative results

**Evaluation Method.** In addition to traditional evaluation methods (IS and FID), Multi-Scale Structural SIMilarity (MS-SSIM) [72], Structural SIMilarity (SSIM), Feature Similarity Index Measure (FSIM) [73] are also employed to evaluate our method quantitatively. The lower the value of FID, the better the result. Other methods are that the higher the value, the better the result.

Table 3.5 shows the comparison of our method with existing T2I methods on MS-SSIM, SSIM, and FSIM. The comparative results show that our method performs the best. This shows that the structural similarity between the results of our method and

Table 3.6: The IS and FID comparison results of our method, existing T2I methods, and CustomizableGAN are shown below.

Model	CUB		Oxford-102		MS COCO	
	IS $\uparrow$	FID $\downarrow$	IS $\uparrow$	FID $\downarrow$	IS $\uparrow$	FID $\downarrow$
AttnGAN [6]	4.36	23.98	3.75	37.94	25.89	35.49
DMGAN [10]	4.75	16.09	4.03	21.36	30.49	32.64
TIME [74]	4.91	14.30	-	-	30.85	31.14
DFGAN [75]	<b>5.10</b>	14.81	-	-	-	19.31
DrawGAN	4.76	9.87	4.07	20.24	31.11	31.51
INS_fore	4.79	<b>9.29</b>	4.19	18.98	30.70	29.55
EruditeGAN	4.69	9.58	4.07	<b>17.69</b>	<b>31.94</b>	28.79
CustomizableGAN	3.12	65.36	2.76	78.84	7.83	91.26
Ours	4.64	10.98	<b>5.03</b>	27.21	24.48	<b>19.07</b>
Ours (lw)	4.77	11.95	5.00	29.88	21.44	25.82

real images is the closest, indicating that our synthetic results have the best realism.

Table 3.6 shows the quantitative comparison results on IS and FID. Compared to CustomizableGAN, our results are better, which shows that our method can synthesize better image results based on text and contour information. Moreover, the performance of our method is also competitive compared to the T2I methods.

Tables 3.5 and 3.6 also show the comparison between our designed structure and the lightweight structure. The results between them are relatively close. In terms of parameters, the model parameters are reduced by 50.3% (from 31.8M to 15.8M) in the CUB and Oxford-102 datasets. In the MS COCO dataset, it is reduced by 36.8% (from 47.6M to 30.1M), which reflects the effectiveness of our lightweight work.

### 3.5 Internal Comparison of Proposed Methods

In order to improve the controllability of image synthesis, we have proposed two methods: Customizable GAN: a method for image synthesis of human controllable; TCGIS: Text and Contour Guided Artificially Controllable Image Synthesis, respectively expressed as CustomizableGAN, TCGIS.

The basic idea of the two methods we propose is to synthesize corresponding image results based on text and contour information, and both text and contour can be

manually input. Therefore, our proposed methods make the whole image synthesis process well human-controllable and achieve the objective of improving controllability.

The qualitative and quantitative comparison results of CustomizableGAN and TCGIS are shown in Figure 3.15 and Table 3.6. Judging from the synthesis results, the two methods we proposed both achieve better human-controllable image synthesis effects. In contrast, the synthesis quality of TCGIS is significantly better than that of CustomizableGAN, which indicates that the structure design of TCGIS is more efficient. In addition, TCGIS has demonstrated excellent performance in complex image synthesis, and the synthesis effect is better than the current T2I methods, which shows that TCGIS initially has good applicability.

## 3.6 Chapter Conclusion

In this chapter, we propose two methods (CustomizableGAN, and TCGIS) to improve the controllability of image synthesis. We conduct a detailed introduction for each proposed method, including network structure, loss function, implementation details, and experiments.

Extensive experimental results show that our proposed method can well promote the controllability of image synthesis. Besides, the TCGIS method shows excellent performance in image synthesis quality. Especially for complex image synthesis, it can synthesize very realistic image results. Combined with its own human controllability, it can be said that this method already has certain applicability.

In the next chapter, we will introduce the high practicality oriented image synthesis methods.

# Chapter 4

## High Practicality Oriented Image Synthesis Methods

### 4.1 Introduction

In Chapters 2 and 3, we introduce our proposed high quality oriented and high controllability oriented image synthesis methods, respectively. These methods have better solved the problems of insufficient synthesis quality and controllability existing in the current T2I method. However, the practicability of these proposed methods is still relatively modest. Although the TCGIS method proposed in Chapter 3 is practical to a certain extent, it is far from enough.

To further improve the practicality of image synthesis methods, we introduce text-guided image manipulation methods. Specifically, for previously synthesized image results, its content can be modified using text information. In this way, by combining the previously proposed image synthesis method with the proposed text-guided image manipulation method, a high practicality image synthesis approach is obtained.

Therefore, in this Chapter, we first introduce our proposed text-guided image manipulation (TGIM) method. And then show the experimental results of the high practicality image synthesis formed by introducing the proposed TGIM method into the previously proposed image synthesis methods. Specifically, it includes two parts: text-guided image synthesis and manipulation (text-guided image synthesis method combined with text-guided image manipulation method); text-guided controllable image synthesis and manipulation (text- and contour-guided image synthesis method combined with text-guided image manipulation method).

## 4.2 Related Works

Image manipulation has attracted much attention in the computer vision community because of its many potential applications, such as computer-aided design, image editing, and computer games.

It aims to modify the content of the given image with different granularity, including low-level color [76] and texture [77] information and high-level semantic information [78] to meet the various preferences of users. In recent years, with the rise of artificial intelligence, especially the introduction of deep learning, the research of image manipulation has entered the automatic image manipulation stage, including style conversion [79][80], image inpainting [81][82], image translation [83][84], and image colorization [76][85].

Despite the fact that many recent studies on image manipulation have produced remarkable outcomes, the majority of them are narrowly focused and lack flexibility. In response to this problem, a new type of image manipulation method—text-guided image manipulation—is developed. In this way, natural language description is used to modify the image content so that the entire image manipulation task has good flexibility. On the other hand, natural language description conforms to human input habits, which can accelerate the development of image manipulation toward user-friendly applications. The existing text-based image manipulation methods [24][86][69][70][87] have been able to modify some parts of the image according to the input text description and have achieved encouraging results. Nevertheless, some methods [24][86] are not satisfactory in terms of modification effect. Meanwhile, some other methods [69][70][87] have certain flaws in the whole modification process so that the modification effect still has a large room for improvement. Specifically, they only use sentence information in the initial modification stage and only use fixed word information to fine-tune the modification content in the later modification stage.

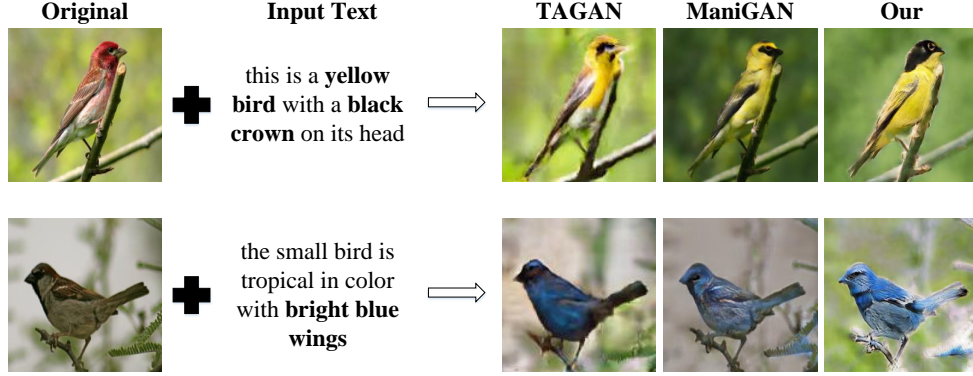


Figure 4.1: The comparison results with other text-guided image manipulation methods are shown above. In contrast, the image results manipulated by our method are clearer and more conform to the semantic information of the input text.

### 4.3 Text-guided Image Manipulation based on Sentence-aware and Word-aware Network

In order to achieve a better text-guided image manipulation effect, we propose a sentence-aware and word-aware network (SWN). In our proposed SWN, there is a sentence-aware (SA) and a word-aware (WA) approach. The sentence-aware method refers to the fusion of sentence features during feature processing so that the final manipulated image result is more consistent with the input text information. The word-aware method refers to utilizing the attention mechanism (specifically, we employ the dynamic selection method [10]) to use word information to fine-tune the results of manipulating images to further improve image quality. Fig. 4.1 shows that our method has made a significant improvement in image manipulation.

#### 4.3.1 Network Structure

Fig. 4.2 shows the specific network structure of our proposed method. For the input image and text description, the corresponding image features and text features are extracted through a pre-trained image encoder [31] and a pre-trained text encoder [30], respectively. The text features include global sentence features and local word

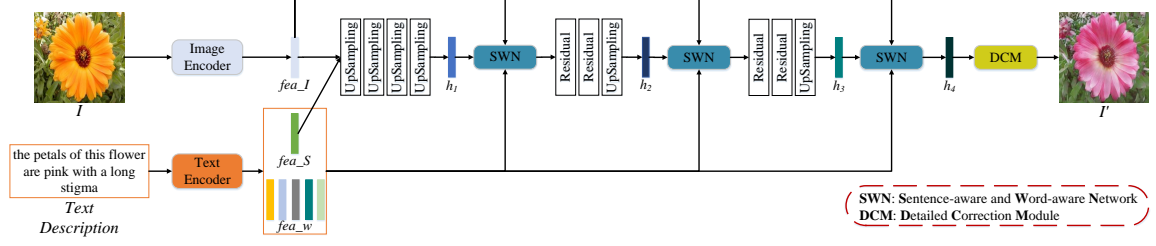


Figure 4.2: The network architecture of our proposed method is shown above. By using global sentence information and dynamically adjusted word information in SWN, we finally obtain high-quality image manipulation result that conforms to the semantics of the input text.

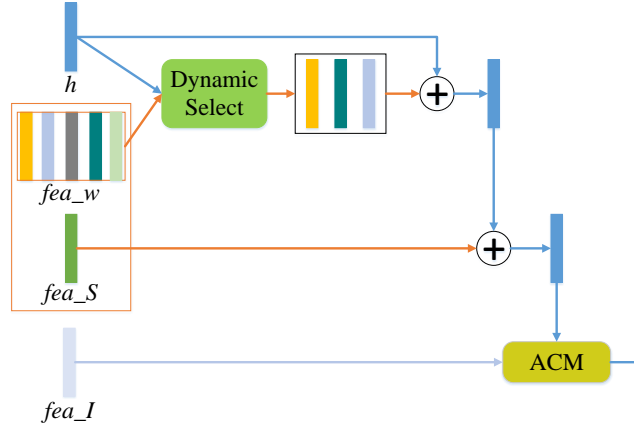


Figure 4.3: The specific processing process of the sentence-aware and word-aware network (SWN) is shown above. ‘ $\oplus$ ’ represents the element-wise addition operation.

features. At first, the image features are combined with the global sentence features, and then the initially hidden features are generated through continuous up-sampling operations. After that, the hidden features will be further fused with image features and text features (including global sentence features and local word features) through the sentence-aware and word-aware network (SWN). Then, new hidden features will be generated through residual blocks [37] and up-sampling operations. This process continues until the third time when the features generated by SWN are directly input into the detail correction module (DCM) to obtain the final hidden features and synthesize the corresponding image manipulation result.

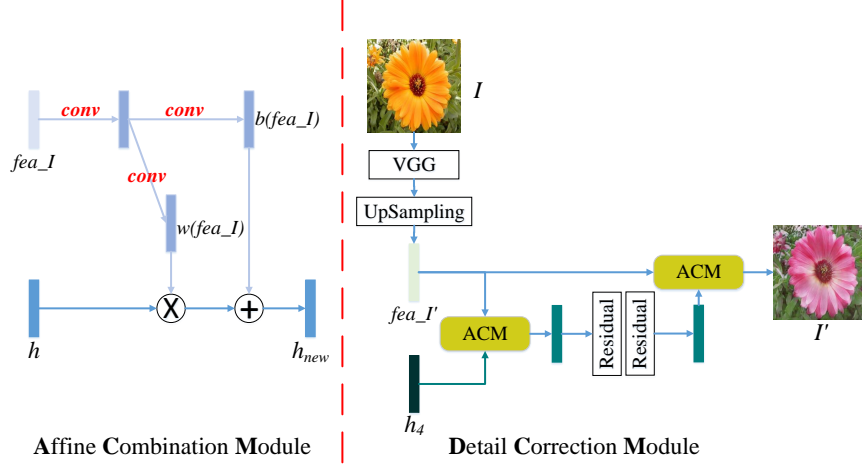


Figure 4.4: The basic processing procedures of affine combination module (ACM) and detail correction module (DCM) are shown in the figure above.  $\otimes$  represents the Hadamard product operation.

### Proposed SWN

The specific structure of SWN is shown in Fig. 4.3. Hidden features and local word features are first used by the dynamic selection method [10] to select the most relevant word features. The specific content of the dynamic selection method has shown in Eq. 2.18-2.23. After obtaining the selected word features, they will be fused with the input hidden features and global sentence features in turn. Then the fused features and the input image features are further fused through an affine combination module (ACM) operation and then input into the subsequent network. Different from the feature fusion design of existing works [69][70], the core of our proposed SWN is to use the dynamic selection method to fuse word features and further fuse sentence features, where word information can fine-tune the manipulation results to further improve the quality, while sentence information provides global text semantic information so that it can improve the semantic consistency between the manipulation result and the input text.

## ACM & DCM

ACM and DCM are proposed by ManiGAN [69], they can provide better feature processing results to achieve higher-quality image manipulation results. Therefore, referring to ManiGAN, we have followed the ACM and DCM proposed by ManiGAN in our designed structure. The specific contents of ACM and DCM are shown as follows:

**The affine combination module (ACM)** is used to associate the cross-modal representations. It contains two inputs, one is the image regional features  $fea\_I \in \mathbb{R}^{256 \times 17 \times 17}$ , and the other is the hidden features  $h \in \mathbb{R}^{C \times H \times W}$ , where  $C$  is the number of channels,  $H$  and  $W$  are the height and width of feature map, respectively. The regional features are sampled and input into two convolutional layers to generate a weight matrix  $W$  and a bias matrix  $b$ , which are then combined with hidden features to form the final new hidden features. The specific equation is as follows:

$$h_{new} = h \otimes W(fea\_I) + b(fea\_I) \quad (4.1)$$

where  $\otimes$  represents the Hadamard product operation.

**The detail correction module (DCM)** is used to realize more fine-grained image modification. As shown on the right side of Fig. 4.4, its input includes three parts: the previously hidden features, the input text features (including sentence features and word features), and the input image features (obtained after the VGG network [67] and the up-sampling operation). Firstly, the hidden features and text features are combined through SWN to form new hidden features. Afterward, the new hidden features and image features are combined through ACM, then processed through residual operations. The processed features will be merged with the image features through ACM to form the final hidden features.

### 4.3.2 Loss Function

We follow the objective function of ControlGAN [26] for training the generator and discriminator. Besides, we add a reconstruction term in the generator as follows:

$$L_{rec} = 1 - \frac{1}{CHW} \|I' - I\| \quad (4.2)$$

where  $C$ ,  $H$ , and  $W$  respectively represent the number of channels, height, and width of the image.  $I$  denotes the real image, and  $I'$  denotes the image result modified by our proposed method. Reconstruction loss can be used to improve the diversity of modification results. This item will produce a significant penalty value when the modification result is the same as the input image.

### 4.3.3 Implementation Details

During training, one convolutional layer and one InstanceNorm operation [88] are used in up-sampling, and two convolutional layers and two InstanceNorm operations are used in the residual block. We train our models using Adam optimizer [38] with an initial learning rate of 0.0002 and a batch size of 10. Empirically, we train 600 epochs before the DCM stage and 800 epochs in the DCM stage for the CUB, Oxford-102 flower datasets, and 200 epochs before the DCM stage and 200 epochs in the DCM stage for the MS COCO dataset. Besides, we use a pre-trained image encoder [31] and text encoder [30] to obtain the corresponding image and text features.

### 4.3.4 Experiments

#### Qualitative comparison

The result of the qualitative comparison between our proposed method and the current existing text-guided image manipulation methods on the CUB dataset is shown in Fig.

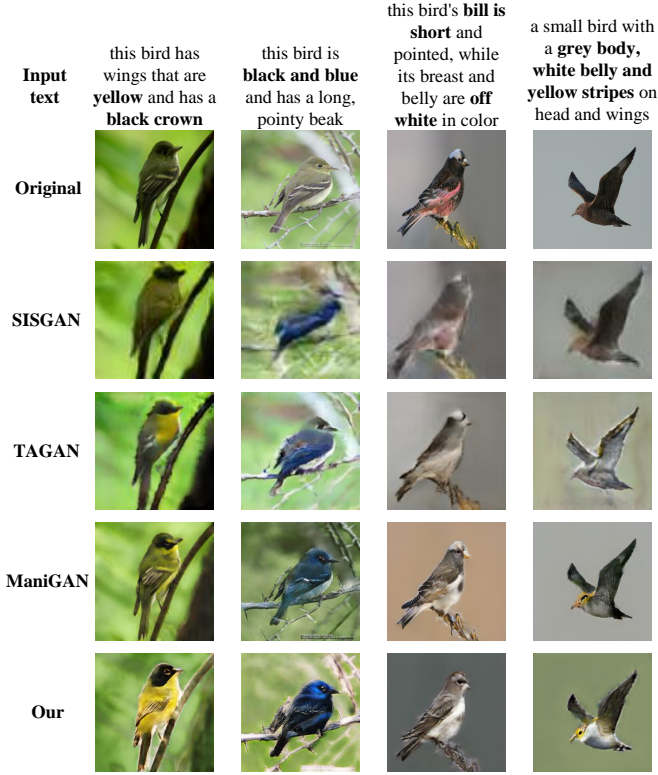


Figure 4.5: The comparison between our method and other existing methods on the CUB dataset is shown above. It shows that our modified details are more refined than other methods.

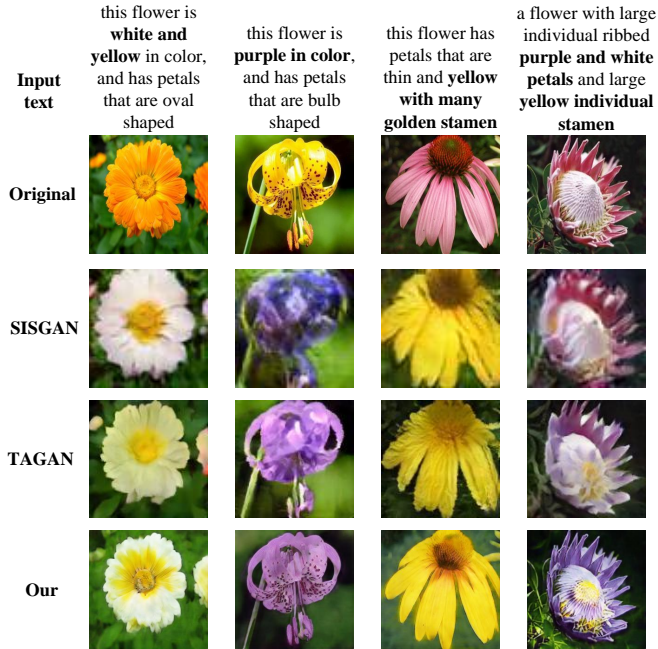


Figure 4.6: The comparison between our method and other existing methods on the Oxford-102 flower dataset is shown above. The comparison results show that our method has achieved an excellent modification effect for flower images.

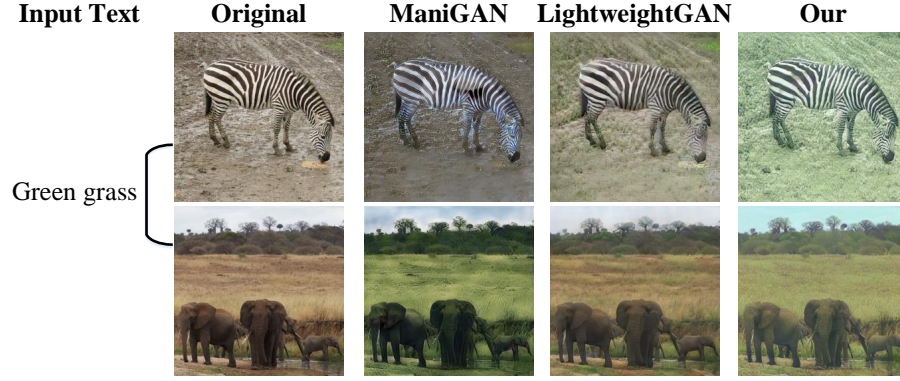


Figure 4.7: The comparison between our method and other existing methods on the MS COCO dataset is shown above.

4.5. In contrast, the overall quality of the results of SISGAN [24] and TAGAN [86] is relatively poor. Especially for SISGAN, the subjective authenticity of its manipulation results is lacking. By contrast, TAGAN’s results are more realistic subjectively, but the overall resolution and specific detail processing are not surprisingly good. The manipulation results corresponding to ManiGAN [69] shown in the figure are better than TAGAN in terms of authenticity and detail processing. At the same time, the results of ManiGAN are more conform with the semantic information of the input text than the results of TAGAN. Compared with the above methods, the manipulation results generated by our method are the best in terms of authenticity, detail processing, and semantic consistency with the text. Compared with ManiGAN, which has better performance, our results are better in detail processing and more conform to the semantic information of the text (such as the results in the first and second columns).

The comparison results on the Oxford-102 flower dataset are shown in Fig. 4.6. It is obvious from the comparison results that our method has achieved excellent manipulation effects in flower images. Our method is far superior in terms of subjective authenticity, detail processing, and semantic consistency than SISGAN and TAGAN.

The comparison results on the MS COCO dataset are shown in Fig. 4.7. It is obvious from the comparison results that our method has achieved excellent manipulation

effects in flower images. Our method is far superior in terms of subjective authenticity, detail processing, and semantic consistency than SISGAN and TAGAN.

It can be found that the manipulation effect of ManiGAN is sometimes very mediocre (no manipulation effect is achieved), and sometimes excessive (the foreground content is also greatly manipulated). And the manipulation effect of LightweightGAN is relatively slight overall. In contrast, the manipulation effects of our proposed method are most appropriate.

Table 4.1: The quantitative comparison results of our method and other existing methods on the CUB dataset are shown below.

Method	IS	FID	NIMA
SISGAN [24]	2.24	104.27	2.81
TAGAN [86]	3.32	57.20	3.48
SAGAN [87]	4.55	-	4.26
ManiGAN [69]	8.47	13.85	4.84
LightweightGAN [70]	9.02	11.02	4.95
Our	<b>11.69</b>	<b>10.75</b>	<b>5.05</b>

## Quantitative comparison

We quantify the performance of the proposed method according to Inception Score (IS) [13], FID [35], and Neural Image Assessment (NIMA) [89]. NIMA is a quantitative evaluation material that is more conforms to human subjective evaluation, that is, the scoring of images by this metric conforms to the subjective perception of humans. The higher value of NIMA, the better the human perception of the image, that is, the higher image authenticity.

Table 4.2: The quantitative comparison results of our method and other existing methods on the Oxford-102 flower dataset are shown below.

Method	IS	FID	NIMA
SISGAN [24]	3.33	108.35	3.77
TAGAN [86]	3.88	55.16	4.74
Our	<b>9.56</b>	<b>30.33</b>	<b>4.96</b>

Table 4.3: The quantitative comparison results of our method and other existing methods on the MS COCO dataset are shown below.

Method	IS	FID	NIMA
ManiGAN [69]	15.09	25.09	5.09
LightweightgAN [70]	28.58	12.39	5.90
Our	<b>29.69</b>	<b>8.62</b>	<b>5.93</b>

The quantitative comparison results on the CUB dataset are shown in Table 4.1. In terms of Inception Score, our method is 421% higher than SISGAN [24], 252% higher than TAGAN [86], 156% higher than SAGAN [87], 38.0% higher than ManiGAN [69], 29.6% higher than LightweightGAN [70], which demonstrates the image manipulation performance by our method is the best according to image quality and diversity. In terms of NIMA, the scores of [69]-[87] all exceed 4, which reflects that these methods are more realistic in human subjective evaluation.

The quantitative comparison results on the Oxford-102 flower dataset are shown in Table 4.2. In terms of Inception Score, our method is 187% higher than SISGAN and 146% higher than TAGAN, which echoes the qualitative results, indicating that our method has an excellent performance in flower image manipulation. In terms of NIMA, our method is also the best.

The quantitative comparison results on the MS COCO dataset are shown in Table 4.3. In terms of Inception Score, our method is 96.7% higher than ManiGAN and 3.9% higher than LightweightGAN, indicating that our method performs better in complex image manipulation. In terms of NIMA, our method is also the best than ManiGAN and LightweightGAN.

### Ablation study

In this section, we conduct an ablation study to verify the effectiveness of Sentence-Aware (SA) and Word-Aware (WA) in our proposed SWN. Ablation models with/without SA and WA are trained and evaluated in the same condition. The results are shown in Table 4.4. Experimental results show the effectiveness of each

Table 4.4: The ablation comparison results of our proposed SWN are shown below. SA indicates the use of the sentence-aware method, and WA indicates the use of the word-aware method. ✓ and ‘-’ indicate that the corresponding methods are used and not used.

		CUB			Oxford-102			MS COCO		
SA	WA	IS	FID	NIMA	IS	FID	NIMA	IS	FID	NIMA
✓	-	10.49	12.23	4.94	9.17	32.84	4.88	28.91	9.68	5.85
-	✓	11.27	11.31	5.03	9.41	32.60	4.94	29.08	8.98	5.90
✓	✓	<b>11.69</b>	<b>10.75</b>	<b>5.05</b>	<b>9.56</b>	<b>30.33</b>	<b>4.96</b>	<b>29.68</b>	<b>8.62</b>	<b>5.93</b>

technique in improving IS, FID, and NIMA scores. Eventually, the best performance is obtained when they are applied together, proving the effectiveness and superiority after integrating two improvements.

## 4.4 Image Synthesis Methods with High Practicality

In order to achieve a highly practical image synthesis method, we introduced the proposed text-guided image manipulation method into the previously proposed T2I method and TCGIS method to form two highly practical image synthesis methods: Text-guided Image Synthesis and Manipulation; Text-guided Controllable Image Synthesis and Manipulation.

### 4.4.1 Text-guided Image Synthesis and Manipulation

The basic way of text-guided image synthesis and manipulation is to first synthesize the corresponding image results based on text information, and then continue to input new text to modify the previously synthesized image until a satisfactory result is generated. The corresponding results are shown in Fig. 4.8. The figure shows that based on the input text, the corresponding image results are first synthesized. Then, for the synthesized image result, the new text can continue to enter to modify the content of

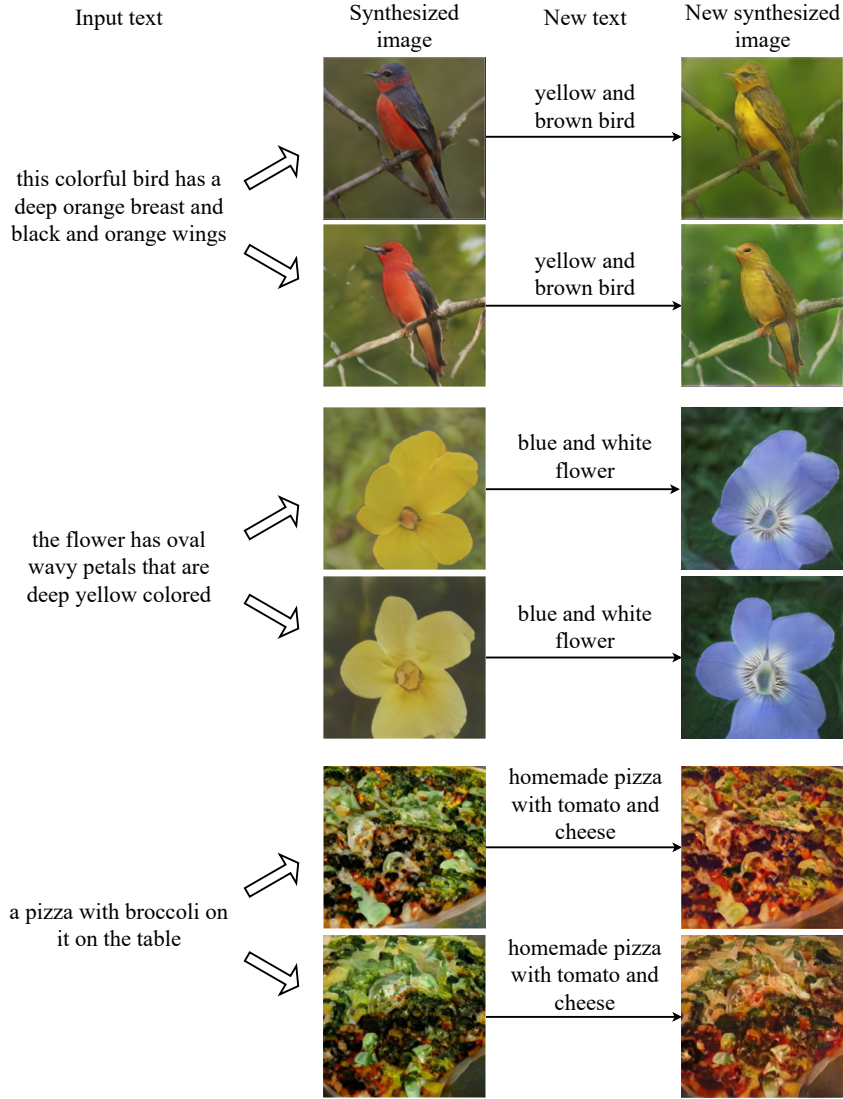


Figure 4.8: The corresponding results of text-guided image synthesis and manipulation are shown above.

the image.

We can find that in the initial stage of text-to-image synthesis, multiple corresponding image results can be synthesized based on the text description. This is mainly because the text information can only determine the basic content of the synthesized image but cannot determine the shape and position information of the synthesized object. This situation makes the method of text-guided image synthesis and manipulation still have room for improvement in terms of human controllability and practicability of the synthesis method.

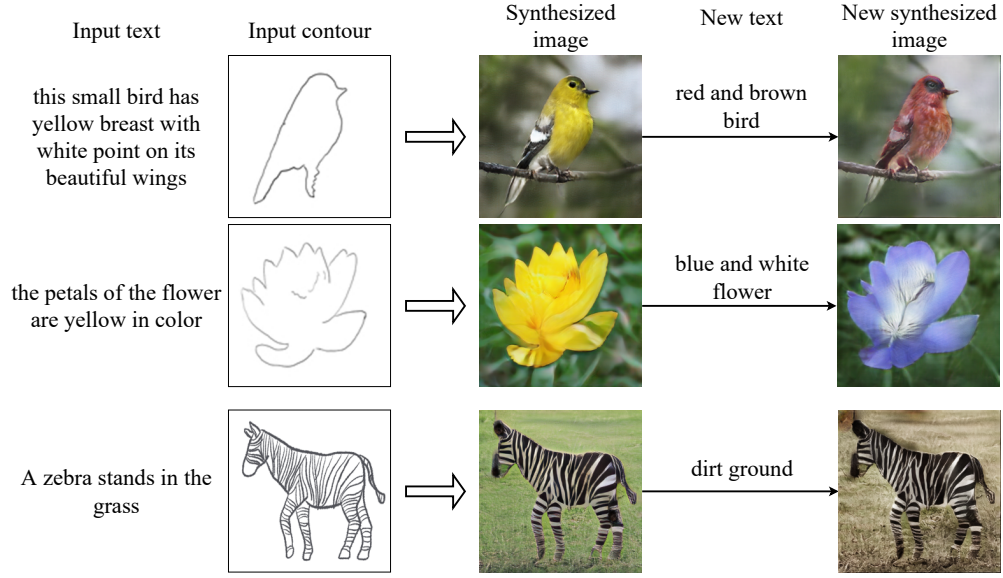


Figure 4.9: The corresponding results of text-guided controllable image synthesis and manipulation are shown above.

#### 4.4.2 Text-guided Controllable Image Synthesis and Manipulation

The basic way of text-guided image synthesis and manipulation is to first synthesize the corresponding image results based on text and contour information, and then continue to input new text to modify the previously synthesized image until a satisfactory result is generated. The corresponding results are shown in Fig. 4.9. The figure shows that based on the input text and contour information, the corresponding image result is synthesized first. Afterward, new text can continue to be entered to manipulate the previously synthesized image content.

We can find that in the initial stage, the text information can control the basic synthetic content, the contour information can control the shape and position information of the synthetic object, and the new text can be used to modify the content of the synthetic image in the later stage. The whole process is all artificially controllable and has better practicability.

## 4.5 Chapter Conclusion

In this chapter, we first propose a TGIM method — Text-guided Image Manipulation based on Sentence-aware and Word-aware Network. We conduct a detailed introduction for our proposed method, including network structure, loss function, implementation details, and experiments. Experimental results show that our proposed method can achieve the image content manipulation effect well.

After that, we combined the proposed TGIM method with the previously proposed T2I and TCGIS methods to form two image synthesis methods with high practicality: text-guided image synthesis and manipulation and text-guided controllable image synthesis and manipulation. Judging from the synthesis process and results, the text-guided image synthesis and manipulation method has achieved good practicability. However, since it cannot control the shape and position information of the synthesized object at the beginning, there is still room for improvement in the practicality of this method. In contrast, the text-guided controllable image synthesis and manipulation method can control the basic content of the synthesized image and the position and shape information of the synthesized object at the beginning, making the method realize the content controllability of the whole process and thus has better practicality.

In the next section, we will summarize the research work and give conclusions.

# Chapter 5

## Conclusion

In this work, deep learning approaches for text-guided artificially controllable image synthesis have been introduced. This dissertation is mainly dedicated to solving the three problems (insufficient quality, insufficient controllability, and insufficient practicality) existing in the current text-to-image synthesis methods. The main proposal consists of three parts: High quality oriented text-to-image synthesis, high controllability oriented image synthesis, and high practicality oriented image synthesis.

In the part of high quality oriented text-to-image synthesis, we propose three text-to-image synthesis methods in total, specifically DrawGAN: Text to Image Synthesis with Drawing Generative Adversarial Networks (aka. DrawGAN), Text-to-Image Synthesis: Starting Composite from the Foreground Content (aka. INS\_fore), and ext to Image Synthesis with Erudite Generative Adversarial Networks (aka. EruditeGAN). Among them, DrawGAN achieves higher-quality image results by simulating the painting process. Specifically, it first synthesizes simple contour result based on text, then synthesizes foreground content, and then synthesizes the final image result. The whole synthesis process is from simple to complex, just like painting step by step. The basic idea of INS\_fore is to first synthesize the corresponding foreground content based on the text, and then synthesize the final image result. Compared with DrawGAN, directly synthesizing the foreground content that matches the text information is simpler and more effective than first synthesizing the contour information that is not closely related to the text information. Therefore, INS\_fore can achieve higher-quality image results. For EruditeGAN, it takes a roundabout way to achieve higher-quality image synthesis results. Specifically, it promotes the generation ability of the generator by improving the discrimination ability of the discriminator. To improve the discriminative ability

of discriminator, we introduce various additional discriminative image types into the discriminator to improve its discriminative ability. Overall, all three of our proposed methods achieve better quality image synthesis results. In contrast, the synthesis quality of INS\_fore and EruditeGAN is better than DrawGAN, and EruditeGAN performs better in complex image synthesis.

In the part of high controllability oriented image synthesis, we propose two methods in total, specifically Customizable GAN: a method for image synthesis of human controllable (aka. CustomizableGAN), and TCGIS: Text and Contour Guided artificially controllable Image Synthesis (aka. TCGIS). The basic idea of CustomizableGAN is to synthesize the corresponding image result based on text and contour information, where text information is used to determine the basic content of synthesis, and contour information is used to determine the shape and position of the synthesized object. The specific implementation method is first to use the encoder to extract the corresponding text features and contour features, then combine the text and contour features, and then generate the corresponding image result after the residual and upsampling operations. CustomizableGAN initially realizes the controllable image synthesis effect, which allows artificial input of text and contour, so it has good artificial controllability. For TCGIS, it designs a more effective network structure based on the realization idea of CustomizableGAN. Specifically, the attention mechanism is introduced to fine-tune the synthesis result to improve its quality. In contrast, TCGIS achieves higher-quality controllable image synthesis results, and it demonstrates unparalleled performance in artificially controllable complex image synthesis.

In the part of high practicality oriented image synthesis, we first propose a text-guided image manipulation method, and then compare this method with the previously proposed T2I and TCGIS methods to form the high practicality image synthesis methods. Specifically, the proposed text-guided image manipulation method is Text-guided Image Manipulation based on Sentence-aware and Word-aware Network (aka. SWN).

The basic idea of SWN is to deeply integrate the hidden features of synthetic content, image features, sentence features, and word features to achieve a better image manipulation effect. The proposed TGIM method is combined with the previously proposed T2I and TCGIS methods to form a total of two highly practical methods: Text-guided image synthesis and manipulation, and Text-guided controllable image synthesis and manipulation. Text-guided Image Synthesis and Manipulation can first synthesize the corresponding image based on the text description, then modify the image’s content based on the new text. For text-guided controllable image synthesis and manipulation, it can first synthesize the image result based on the text and contour information, then modify the generated image’s content based on the new text. Both methods have good practicability. In contrast, the second method achieves the content controllable for the whole process, so it has better practicability.

In summary, this paper realizes the research on image synthesis methods with high quality, high controllability, and high practicability, which can well promote the development of image synthesis towards industrial application.

Although our proposed methods achieve high quality, high controllability, and high practicability image synthesis, from the results shown, our proposed methods still have some limitations. Firstly, in terms of image synthesis quality, our proposed methods achieve higher-quality image synthesis results. However, on complex image synthesis, we found that the performance of the proposed methods is relatively general, which indicates that there is still room for improvement in the synthesis quality of our proposed methods. Secondly, the high controllability image synthesis methods we proposed achieve better control degree over the synthetic content, but from the results shown (Figures 3.4-3.10), our proposed methods cannot control the background content, which makes the overall controllability still have room for improvement. Furthermore, our proposed methods cannot synthesize the realistic image result when the input contour content is not real. Thirdly, in terms of practicability, the text-guided image manipula-

tion method we proposed is still lacking in manipulation functions (such as the ability to manipulate image style), which indicates that there is still room for improvement in the practicability of our proposed method.

For the above limitations, in the future, we will design more effective network structures for complex image synthesis to improve the quality of complex image synthesis. Besides, we will add a separate background generation module and contour content modification module to the proposed high-controllability image synthesis methods to further improve the overall controllability. Furthermore, we will design a more effective and practicability text-guided image manipulation method, which can achieve diverse manipulation functions, such as converting the image content into cartoons, oil paintings, etc., adding related objects to the image content, and so on. In this way, the image synthesis method formed by fusing the new TGIM method with the T2I and TCGIS methods will have higher practicability. In addition to the above research content, we will also be committed to researching lightweight network structures in the future, and lightening the models of our proposed method so that the models we develop can be applied to specific devices, such as cameras, mobile phones, etc. In this way, the overall practicality can be further improved.

# Bibliography

- [1] J. G. Ian, P. Jean, M. Mehdi, X. Bing, W. David, O. Sherjil, C. C. Asron, and B. Yoshua, “Generative adversarial nets,” in *In Proc. NIPS Conf.*, 2014, pp. 2672–2680.
- [2] M. Mehdi and O. Simon, “Conditional generative adversarial nets,” *CoRR*, vol. abs/1411.1784, 2014.
- [3] E. R. Scott, A. Zeynep, Y. Xinchun, L. Lajanugen, S. Bernt, and L. Honglak, “Generative adversarial text to image synthesis,” in *In Proc. ICML Conf.*, 2016, pp. 1060–1069.
- [4] Z. Han, X. Tao, L. Hongsheng, Z. Shaoting, H. Xiaolei, W. Xiaogang, and N. M. Dimitris, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *In Proc. ICML Conf.*, 2017, pp. 5908–5916.
- [5] Z. Han, X. Tao, L. Hongsheng, Z. Shaoting, W. Xiaogang, H. Xiaolei, and N. M. Dimitris, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 41, pp. 1947–1962, 2019.
- [6] X. Tao, Z. Pengchuan, H. Qiuyuan, Z. Han, G. Zhe, H. Xiaolei, and H. Xiaodong, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,” in *In Proc. CVPR Conf.*, 2018, pp. 1316–1324.
- [7] Z. Zizhao, X. Yuanpu, and Y. Lin, “Photographic text-to-image synthesis with a hierarchically-nested adversarial network,” in *In Proc. CVPR Conf.*, 2018, pp. 6199–6208.
- [8] Q. Tingting, Z. Jing, X. Duanqing, and T. Dacheng, “Mirrorgan: Learning text-to-image generation by redescription,” in *In Proc. CVPR Conf.*, 2019, pp. 1505–1514.
- [9] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Learn, imagine and create: Text-to-image generation from prior knowledge,” pp. 885–895, 2019.
- [10] Z. Minfeng, P. Pingbo, C. Wei, and Y. Yi, “Dm-gan: dynamic memory generative adversarial networks for text-to-image synthesis,” in *In Proc. CVPR Conf.*, 2019, pp. 5802–5810.
- [11] S. Hong, D. Yang, J. Choi, and H. Lee, “Inferring semantic layout for hierarchical text-to-image synthesis,” in *In Proc. CVPR Conf.*, 2018, pp. 7986–7994.
- [12] P. K. Diederik and W. Max, “Auto-encoding variational bayes,” in *In Proc. ICLR Conf.*, 2014.

- [13] S. Tim, J. G. Ian, Z. Wojciech, C. Vicki, R. Alec, and C. Xi, “Improved techniques for training gans,” in *In Proc. NIPS Conf.*, 2016, pp. 2226–2234.
- [14] R. Alec, M. Luke, and C. Soumith, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *In Proc. ICLR Conf.*, 2016.
- [15] T. G. Aurele, C. Wenming, M. Xudong, W. Si, W. Hau-San, and L. Qing, “ $\alpha\beta$ -gan: Robust generative adversarial networks,” *Information Sciences.*, vol. 593, pp. 177–200, 2022.
- [16] C. Wengling and H. James, “Sketchygan: Towards diverse and realistic sketch to image synthesis,” in *In Proc. CVPR Conf.*, 2018, pp. 9416–9425.
- [17] G. Chengying, L. Qi, X. Qi, W. Limin, L. Jianzhuang, and Z. Changqing, “Sketchycoco: Image generation from freehand scene sketches,” in *In Proc. CVPR Conf.*, 2020, pp. 5173–5182.
- [18] J. Yongcheng, L. Xiao, D. Yukang, W. Xinchao, D. Errui, S. Mingli, and W. Shilei, “Dynamic instance normalization for arbitrary style transfer,” in *In Proc. AAAI Conf.*, 2020, pp. 4369–4376.
- [19] H. Hua, L. Xinxin, and Y. Rong, “Image style transfer for autonomous multi-robot systems,” *Information Sciences.*, vol. 576, pp. 274–287, 2021.
- [20] D. Hong, F. Gang, Y. Qinan, J. Caoqing, C. Tuo, L. Wenjie, H. Shenghong, and X. Chunxia, “Deep attentive style transfer for images with wavelet decomposition,” *Information Sciences.*, vol. 587, pp. 63–81, 2022.
- [21] H. Zheng, L. Jie, G. Xinbo, and W. Xiumei, “Progressive perception-oriented network for single image super-resolution,” *Information Sciences.*, vol. 546, pp. 769–786, 2021.
- [22] X. Xiangyu, M. Yongrui, S. Wenxiu, and Y. Ming-Hsuan, “Exploiting raw images for real-scene super-resolution,” *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 44, no. 4, pp. 1905–1921, 2022.
- [23] M. Qing, J. Junjun, L. Xianming, and M. Jiayi, “Multi-task interaction learning for spatsiospectral image super-resolution,” *IEEE Transactions on Image Processing*, vol. 31, pp. 2950–2961, 2022.
- [24] D. Hao, Y. Simiao, W. Chao, and G. Yike, “Semantic image synthesis via adversarial learning,” in *In Proc. ICCV Conf.*, 2017, pp. 5707–5715.
- [25] Z. Yulan, D. Feng, and G. Z. Sam, K., “Feature pyramid network for diffusion-based image inpainting detection,” *Information Sciences.*, vol. 572, pp. 29–42, 2021.

- [26] L. Bowen, Q. Xiaojuan, L. Thomas, and H. S. T. Philip, “Controllable text-to-image generation,” in *In Proc. NeurIPS Conf.*, 2019, pp. 2063–2073.
- [27] G. Lianli, C. Daiyuan, S. Jingkuan, X. Xing, Z. Dongxiang, and S. Hengtao, “Perceptual pyramid adversarial networks for text-to-image synthesis,” in *In Proc. AAAI Conf.*, 2019, pp. 8312–8319.
- [28] T. Hongchen, L. Xiuping, L. Xin, Z. Yi, and Y. Baocai, “Semantics-enhanced adversarial nets for text-to-image synthesis,” in *In Proc. ICCV Conf.*, 2019, pp. 10 500–10 509.
- [29] N. Anh, C. Jeff, B. Yoshua, D. Alexey, and Y. Jason, “Plug & play generative networks: Conditional iterative generation of images in latent space,” in *In Proc. CVPR Conf.*, 2017, p. 4467–4477.
- [30] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transaction on Signal Processing.*, vol. 45, p. 2673–2681, 1997.
- [31] S. Christian, V. Vincent, I. Sergey, S. Jonathon, and W. Zbigniew, “Rethinking the inception architecture for computer vision,” in *In Proc. CVPR Conf.*, 2016, pp. 2818–2826.
- [32] W. Catherine, B. Steve, W. Peter, P. Pietro, and B. Serge, “The caltech-ucsd birds-200-2011 dataset,” in *Tech. Rep. CNS-TR-2010-001*, 2011.
- [33] N. Maria-Elena and Z. Andrew, “Automated flower classification over a large number of classes,” in *In Proc. ICVGIP Conf.*, 2008, pp. 722–729.
- [34] L. Tsung-Yi, M. Michael, J. B. Serge, H. James, P. Pietro, R. Deva, D. Piotr, and L. Z. C, “Microsoft coco: Common objects in context,” in *In Proc. ECCV Conf.*, 2015, pp. 740–755.
- [35] H. Martin, R. Hubert, U. Thomas, N. Bernhard, and H. Sepp, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *In Proc. NIPS Conf.*, 2017, pp. 6626–6637.
- [36] R. F. Matthew, D. Z., “Visualizing and understanding convolutional networks,” in *In Proc. ECCV Conf.*, 2014, pp. 818–833.
- [37] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, “Deep residual learning for image recognition,” in *In Proc. CVPR Conf.*, 2016, pp. 770–778.
- [38] P. K. Diederik and B. Jimmy, “Adam: A method for stochastic optimization,” in *In Proc. ICLR Conf.*, 2015.

- [39] I. Sergey and S. Christian, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *In Proc. ICML Conf.*, 2015, pp. 448–456.
- [40] M. Takeru, K. Toshiki, K. Masanori, and Y. Yuichi, “Spectral normalization for generative adversarial networks,” in *In Proc. ICLR Conf.*, 2018.
- [41] X. Bing, W. Naiyan, C. Tianqi, and L. Mu, “Empirical evaluation of rectified activations in convolutional network,” *CoRR*, vol. abs/1505.00853, 2015.
- [42] E. R. Scott, A. Zeynep, M. Santosh, T. Samuel, S. Bernt, and L. Honglak, “Learning what and where to draw,” in *In Proc. NIPS Conf.*, 2016, pp. 217–225.
- [43] K. J. Joseph, P. Arghya, R. Sailaja, and N. B. Vineeth, “C4synth: Cross-caption cycle-consistent text-to-image synthesis,” in *In Proc. WACV Conf.*, 2019, pp. 358–366.
- [44] G. Lianli, C. Daiyuan, Z. Zhou, S. Jie, and S. HengTao, “Lightweight dynamic conditional gan with pyramid attention for text-to-image synthesis,” *Pattern Recognition.*, vol. 110, pp. 107384–10795, 2021.
- [45] P. Dunlu, Y. Wuchen, L. Cong, and L. Shuairui, “Sam-gan: Self-attention supporting multi-stage generative adversarial networks for text-to-image synthesis,” *Neural Networks.*, vol. 138, pp. 57–67, 2021.
- [46] M. Fengling, M. Bingpeng, C. Hong, S. Shiguang, and C. Xilin, “Learning efficient text-to-image synthesis via interstage cross-sample similarity distillation,” *Science China Information Science.*, vol. 64, no. 2, 2021.
- [47] W. Zixu, Q. Zhe, W. Zhi-Jie, H. Xinjian, and C. Yangyang, “Text to image synthesis with bidirectional generative adversarial network,” in *In Proc. ICME Conf.*, 2020, pp. 1–6.
- [48] Y. Yanhua, W. Lei, X. De, and T. Cheng, D. Dacheng, “Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis,” *IEEE Transaction on Image Processing.*, vol. 30, pp. 2798–2809, 2021.
- [49] S. Shikhar, S. Dendi, M. Vincent, E. K. Samira, and B. Yoshua, “Chatpainter: Improving text to image generation using dialogue,” in *In Proc. ICLR Conf.*, 2018.
- [50] H. Kaiming, G. Georgia, D. Piotr, and B. G. Ross, “Mask R-CNN,” in *In Proc. ICCV Conf.*, 2017, pp. 2980–2988.

- [51] I. Phillip, Z. Jun-Yan, Z. Tinghui, and A. E. Alexei, “Image-to-image translation with conditional adversarial networks,” in *In Proc. CVPR Conf.*, 2017, pp. 5967–5976.
- [52] L. Ming-Yu, B. Thomas, and K. Jan, “Unsupervised image-to-image translation networks,” in *In Proc. NIPS Conf.*, 2017, p. 700–708.
- [53] C. Xi, D. Yan, H. Rein, S. John, S. Ilya, and A. Pieter, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *In Proc. NIPS Conf.*, 2016, pp. 2172–2180.
- [54] M. Elman, P. Emilio, J. B. Lei, and S. Ruslan, “Generating images from captions with attention,” in *In Proc. ICLR Conf.*, 2016.
- [55] G. Karol, D. Ivo, G. Alex, J. R. Danilo, and W. Daan, “Draw: A recurrent neural network for image generation,” in *In Proc. ICML Conf.*, 2015, pp. 1462–1471.
- [56] V. D. O. Aäron, K. Nal, E. Lasse, K. Koray, V. Oriol, and G. Alex, “Conditional image generation with pixelcnn decoders,” in *In Proc. NIPS Conf.*, 2015, pp. 1252–1260.
- [57] V. D. O. Aäron, K. Nal, and K. Koray, “Pixel recurrent neural networks,” in *In Proc. ICML Conf.*, 2016, pp. 1747–1756.
- [58] D. Alexey, T. S. Jost, and B. Thomas, “Learning to generate chairs with convolutional neural networks,” in *In Proc. CVPR Conf.*, 2015, pp. 1538–1546.
- [59] E. R. Scott, Z. Yi, Z. Yuting, and L. Honglak, “Deep visual analogy making,” in *In Proc. NIPS Conf.*, 2015, pp. 1252–1260.
- [60] J. R. Danilo, M. Shakir, and W. Daan, “Stochastic backpropagation and approximate inference in deep generative models,” in *In Proc. ICML Conf.*, 2014, pp. 1278–1286.
- [61] A. Martín, C. Soumith, and B. Léon, “Wasserstein gan,” *CoRR*, vol. abs/1701.07875, 2017.
- [62] Z. Zhiqiang, Z. Jinjia, Y. Wenxin, and J. Ning, “Drawgan: Text to image synthesis with drawing generative adversarial networks,” in *In Proc. ICASSP Conf.*, 2021, pp. 4195–4199.
- [63] M. Luke, P. Ben, P. David, and S. Jascha, “Unrolled generative adversarial networks,” in *In Proc. ICLR Conf.*, 2017.

- [64] L. Christian, T. Lucas, H. Ferenc, C. Jose, C. Andrew, A. Alejandro, A. P. A., T. Alykhan, T. Johannes, W. Zehan, and W. S., “Photo-realistic single image super-resolution using a generative adversarial network,” in *In Proc. CVPR Conf.*, 2017, pp. 105–114.
- [65] B. Andrew, L. Theodore, M. R. James, and W. Nick, “Neural photo editing with introspective adversarial networks,” in *In Proc. ICLR Conf.*, 2017.
- [66] T. Yaniv, P. Adam, and W. Lior, “Unsupervised cross-domain image generation,” in *In Proc. ICLR Conf.*, 2017.
- [67] S. Karen and Z. Andrew, “Very deep convolutional networks for large-scale image recognition,” in *In Proc. ICLR Conf.*, 2015.
- [68] E. R. Scott, A. Zeynep, L. Honglak, and S. Bernt, “Learning deep representations of fine-grained visual description,” in *In Proc. CVPR Conf.*, 2016, pp. 49–58.
- [69] L. Bowen, Q. Xiaojuan, L. Thomas, and H. S. T. Philip, “Manigan: Text-guided image manipulation,” in *In Proc. CVPR Conf.*, 2020, pp. 7877–7886.
- [70] L. Bowen, Q. Xiaojuan, H. S. T. Philip, and L. Thomas, “Lightweight generative adversarial networks for text-guided image manipulation,” in *In Proc. NeurIPS Conf.*, 2020.
- [71] Z. Zhiqiang, Z. Jinjia, Y. Wenxin, and J. Ning, “Text-to-image synthesis: Starting composite from the foreground content,” *Inf. Sci.*, vol. 607, pp. 1265–1285, 2022.
- [72] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in *In Proc. ACSSC Conf.*, 2003, pp. 1398–1402.
- [73] Z. Lin, Z. Lei, M. Xuanqin, and Z. David, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.
- [74] L. Bingchen, S. Kunpeng, Z. Yizhe, D. M. Gerard, and E. Ahmed, “TIME: text and image mutual-translation adversarial networks,” in *In Proc. AAAI Conf.*, 2021, pp. 2082–2090.
- [75] T. Ming, T. Hao, W. Fei, J. Xiaoyuan, B. Bing-Kun, and X. Changsheng, “DF-GAN: A simple and effective baseline for text-to-image synthesis,” in *In Proc. CVPR Conf.*, 2022, pp. 16 494–16 504.
- [76] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *In Proc. ECCV Conf.*, 2016, pp. 649–666.

- [77] Z. Jun-Yan, P. Taesung, I. Phillip, and A. E. Alexei, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *In Proc. ICCV Conf.*, 2017, pp. 2242–2251.
- [78] J. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, “Generative visual manipulation on the natural image manifold,” in *In Proc. ECCV Conf.*, 2016, pp. 597–613.
- [79] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *In Proc. ECCV Conf.*, 2016, pp. 694–711.
- [80] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *In Proc. CVPR Conf.*, 2016, pp. 2414–2423.
- [81] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *In Proc. CVPR Conf.*, 2016, pp. 2536–2544.
- [82] Y. Jo and J. Park, “Sc-fegan: Face editing generative adversarial network with user’s sketch and color,” in *In Proc. CVPR Conf.*, 2019, pp. 1745–1753.
- [83] X. Qi, Q. Chen, J. Jia, and V. Koltun, “Semi-parametric image synthesis,” in *In Proc. CVPR Conf.*, 2018, pp. 8808–8816.
- [84] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *In Proc. CVPR Conf.*, 2018, pp. 8798–8807.
- [85] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification,” *ACM Transactions on Graphics (ToG)*, vol. 35, no. 4, pp. 1–11, 2016.
- [86] N. Seonghyeon, K. Yunji, and J. K. Seon, “Text-adaptive generative adversarial networks: Manipulating images with natural language,” in *In Proc. NeurIPS Conf.*, 2018, pp. 42–51.
- [87] T. Haruyama, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, “Segmentation-aware text-guided image manipulation,” in *In Proc. ICIP Conf.*, 2021, pp. 2433–2437.
- [88] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” *arXiv preprint arXiv:1607.08022*, 2016.
- [89] H. Talebi and P. Milanfar, “Nima: Neural image assessment,” *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

# List of Abbreviations

<b>GAN</b>	Generative Adversarial Netowrks
<b>T2I</b>	Text-to-Image Synthesis
<b>TGIM</b>	Text-Guided Image Manipulation
<b>cGAN</b>	Conditional Generative Adversarial Networks
<b>VAE</b>	Variational Autoencoders
<b>DCGAN</b>	Deep Convolution Generative Adversarial Netowrks
<b>CNN</b>	Convolutional Neural Networks
<b>DAMSM</b>	Deep Attentional Multimodal Similarity Model
<b>BiLSTM</b>	Bidirectional Long Shore-Term Memor
<b>IS</b>	Inception Score
<b>CS</b>	Fréchet Inception Distance
<b>DrawGAN:</b>	Drawing Generative Adversarial Networks
<b>BN</b>	Batch Normalization
<b>SpectralNorm</b>	Spectral Normalization
<b>leaky ReLU</b>	Rectified Linear Unit
<b>FC</b>	Fully Connected
<b>EruditeGAN</b>	Erudite Generative Adversarial Network
<b>CA</b>	Conditional Augmentation
<b>GAWWN</b>	Generative Adversarial What-Where Network
<b>DRAW</b>	Deep Recurrent Attention Writer
<b>VGG</b>	Visual Geometry Group
<b>HR</b>	Human Rank
<b>kp</b>	key point
<b>bb</b>	bounding box
<b>ACM</b>	Affine Combination Module

**MS-SSIM** Multi-Scale Structural SIMilarity

**SSIM** Structural SIMilarity

**FSIM** Feature Similarity Index Measure

**TCGIS** Text and Contour Guided Image Synthesis

**SWN** Sentence-aware and Word-aware Network

**RNN** Recurrent Neural Networks

**DCM** Detail Correction Module

**NIMA** Neural Image Assessmen