DOI: 10.54254/2755-2721/46/20241109

A comparative analysis and investigation of Attn-GAN and SSA-GAN for text-to-image generation

Jinjia Zhang

Department of Engineering School, Duke Kunshan University, Suzhou, 215316, Jiangsu

jz373@duke.edu

Abstract. Text-to-image generation has emerged as a captivating and intricate challenge within the field of artificial intelligence. This paper provided an extensive comparative evaluation of two cutting-edge Generative Adversarial Network (GAN) models, namely Attn-GAN and SSA-GAN, in the context of text-to-image generation. The significance of this problem extends to a multitude of applications, encompassing content generation, advertising, and virtual reality. Attn-GAN, an acronym for Attentional Generative Adversarial Network, leverages attention mechanisms to align textual descriptions with their corresponding image regions.This approach aims to detect feature details and ensure semantic consistency in the generated images. In contrast, SSA-GAN, or Semantic-Spatial Aware GAN, integrates spatial and semantic information into the image generation process to produce visually plausible and semantically meaningful images.This paper provides a detailed examination of the architectures and working principles of both Attn-GAN and SSA-GAN, followed by a comparative evaluation covering image quality, semantic fidelity, and computational efficiency. The results reveal that SSA-GAN excels in generating images with superior semantic consistency and fine-grained details, while Attn-GAN produces diverse and visually appealing images. Furthermore, the study includes practical recommendations for selecting the appropriate model based on specific project requirements. This study conducted experiments on the COCO and CUB datasets, showcasing the strengths and weaknesses of each model in different scenarios. The findings emphasize the importance of understanding these trade-offs when choosing between Attn-GAN and SSA-GAN for text-to-image generation tasks.

Keywords: Text2image, Attn-GAN, SSA-GAN, Computer Vision, Deep Learning.

1. Introduction

Text-to-image generation has emerged as a captivating and challenging problem in the field of artificial intelligence. The task involves generating realistic images from textual descriptions, a capability with profound implications for various applications, including content creation, advertising, and virtual reality [1].

Over the years, GANs have demonstrated remarkable success in various image generation tasks, ranging from style transfer to super-resolution [2]. Text-to-image generation presents a unique set of challenges due to the inherent semantic gap between text and visual data. Addressing this gap requires models that can effectively bridge the linguistic and visual modalities. Attention-Guided Generative

Adversarial Network (Attn-GAN) and Semantic-Spatial Aware GAN (SSA-GAN) represent two distinct approaches to tackling this problem.

Attn-GAN, short for Attention-Guided Generative Adversarial Network, is a GAN architecture that leverages attention mechanisms to align textual descriptions with corresponding image regions. This attention-guided approach is designed to capture fine-grained details and semantic consistency in generated images [3]. On the other hand, SSA-GAN, or Semantic-Spatial Aware GAN, focuses on incorporating spatial and semantic information into the image generation process. By encoding both spatial and semantic awareness, SSA-GAN aims to produce images that are not only visually plausible but also semantically meaningful [4].

The primary objective of this paper is to conduct a comprehensive comparative analysis of Attn-GAN and SSA-GAN in the context of text-to-image generation. This analysis will provide insights into the comparison of these two models, both from the perspective of the actual quality of the generated pictures and the results of evaluation metrics. To be more specific, this study is composed of the following parts: 1) A detailed examination of the architectures and working principles of Attn-GAN and SSA-GAN. 2) A comparative evaluation of these models in terms of image quality, semantic fidelity, and computational efficiency. 3) Insights into the trade-offs between the two models, highlighting their respective advantages and limitations. 4) Practical recommendations for selecting the appropriate model based on specific project requirements.

The remainder of this paper is structured as follows: In Section 2, this study provides an in-depth description of the methodologies employed by Attn-GAN and SSA-GAN, including their key components and training strategies as well as present the experimental setup and datasets used for this paper’s comparative analysis. Section 3 discusses the results of experiments, comparing the performance of Attn-GAN and SSA-GAN across various metrics. In Section 4,a comprehensive discussion of this paper’s research results was provided, emphasizing the implications and practical considerations of using these models in text-to-image generation applications. Finally, an overall conclusion for this paper’s research was given in section 5.

2. Methods

In this section, this paper delves into the methodologies of Attn-GAN and SSA-GAN, outlining their architectural components and training strategies. Any modifications or customizations made to these models for the experiments will be described.

2.1. Attn-GAN

Attn-GAN shown in Figure 1, introduced by Xu et al. [5], is an attention-based GAN designed explicitly for text-to-image generation. It comprises three primary components: a text encoder, an image generator, and a discriminator. The text encoder encodes textual descriptions into a fixed-length vector, while the image generator synthesizes images from these vectors. The discriminator evaluates the realism of generated images. A key innovation in Attn-GAN is the attention mechanism, which allows the model to focus on specific regions of the image that correspond to the words in the text description.

2.1.1. Architecture. The model follows a two-step process, consisting of text-to-image synthesis and image refinement: 1) GAN is employed to generate a low-resolution image based on the provided textual description. 2) Attention mechanisms are applied to refine the generated image by selectively focusing on specific regions according to the given description.

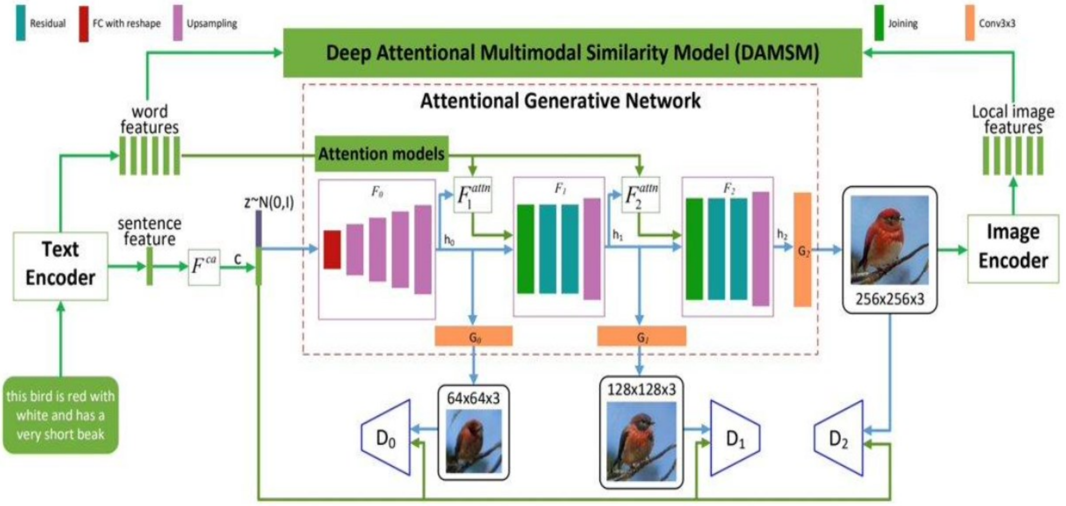

Figure 1. Architecture of Attn-GAN [6].

Attention mechanism: The attention mechanisms play a crucial role in computing attention weights or scores for individual elements within the input sequence [7]. These scores are subsequently employed to determine the significance of each element in shaping predictions or outputs. Through the selective focus on pertinent information and the exclusion of less important details, the model attains enhanced visual quality and coherence. DAMSM: DAMSM is a matching module that is used in conjunction with AttnGAN [8]. It enables the alignment of textual descriptions and visual features at multiple scales. By incorporating attention mechanisms, DAMSM computes matching scores between different levels of textual and visual features. This allows for more precise and fine-grained alignment, capturing the correspondence between textual details and visual elements in the generated images [9]. Together, AttnGAN and DAMSM form a powerful framework for generating images from textual descriptions. AttnGAN synthesizes visually coherent images, while DAMSM ensures the alignment between textual descriptions and visual features, enhancing the overall quality and fidelity of the generated images [10].

2.1.2. Implementation details. Training Attn-GAN involves adversarial training, where the generator and discriminator are trained simultaneously. The text encoder, image generator, and discriminator are optimized using gradient-based techniques, such as Stochastic Gradient Descent (SGD) or Adam. The loss function comprises adversarial loss, which encourages the generator to produce realistic images, and a text-image matching loss, which ensures that generated images are semantically consistent with the input text. This study chose 600 as the training epochs and 0.0001 as the initial learning rate.

2.2. SSA-GAN: Semantic-Spatial Aware GAN

SSA-GAN, proposed by Liao et al. [11], adopts a different approach to text-to-image generation. It focuses on integrating spatial and semantic awareness into the image generation process, with the goal of producing images that are not only visually plausible but also semantically meaningful.

2.2.1. Architecture. SSA-GAN contains a text encoder, a generator and a discriminator. The main difference between SSA-GAN and Attn-GAN is the generator. Figure 2 shows the architecture of the SSA-GAN model:

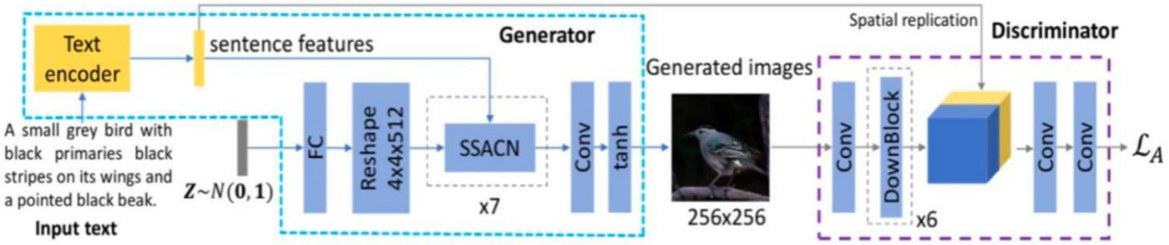

Figure 2. Architecture of SSA-GAN [12].

SSA-GAN applies a Semantic spatial-aware Convolutional network (SSACN) module that predicts a mask predictor according to currently generated image features [13]. This mask predictor shown in Figure 3 not only determines where to put texts, but also plays a weighting role in determining how much text information to reinforce on a certain section.

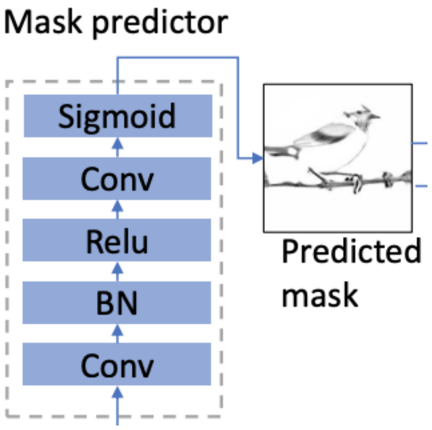

Figure 3. Mask Predictor in SSACN [14].

Additionally, SSACN includes a Semantic Condition Batch Normalization (SCBN) to train the text feature vectors [15]. The architecture of SCBN show in Figure 4 is similar to the DF-GAN [16], which also contains two MLPs to train the model. The creative point here is that it induced a new method for calculating affine parameters is proposed. Mask maps are added to SCBN as spatial conditions, then affine parameters are learned from encoded text vectors and the semantic spatial conditions are normalized in batches.

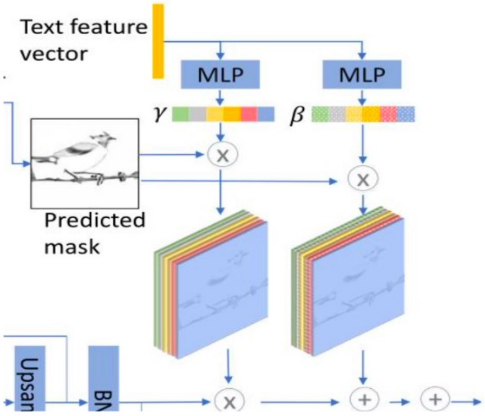

Figure 4. Logistics of Semantic Condition Batch Normalization (SCBN) [17].

2.2.2. Implementation details. Training SSA-GAN involves training the text embedding module, spatial generator, object generator, and discriminator jointly. The text embeddings are learned through backpropagation, while the generators and discriminator are optimized using gradient-based methods. The loss function includes adversarial loss, which encourages the generators to produce realistic objects and images, and a semantic matching loss, which enforces semantic consistency between the text and generated objects. This study chose 600 as the training epochs and 0.0005 as the initial learning rate.

2.3. Dataset

To evaluate the performance of Attn-GAN and SSA-GAN, this study utilized commonly used text-to-image generation datasets, including the COCO dataset and the CUB dataset. COCO Dataset: The COCO dataset [18] is a large-scale dataset with textual descriptions and corresponding images. It covers a wide range of objects and scenes, making it suitable for general text-to-image generation tasks. CUB Dataset: The CUB dataset [19] focuses on fine-grained bird species. It includes textual descriptions and images of bird species, making it a challenging dataset for generating detailed and semantically accurate images.

2.4. Evaluation metrics

Inception Score (IS): The Inception Score evaluates the diversity and quality of generated images. Higher IS values indicate better image quality and diversity.

Frechet Inception Distance (FID): FID measures the similarity between the distribution of generated images and real images. Lower FID values signify better image quality and realism.

3. Experimental results

Additional qualitative instances have been furnished from both the Bird (Figure5) and COCO (Figure6) datasets respectively. The Bird dataset primarily emphasizes the synthesis of ornithological detail, whereas the COCO dataset concentrates on the generation of multiple objects within diverse backgrounds.

In Figure 5, a discerning observation reveals that the avian entities generated by SSA-GAN methodology (up) exhibit a higher degree of vivacity and adhere more closely to the attributes outlined in the accompanying text, in comparison to Attn-GAN (bottom). Similarly, Figure 6 exemplifies the ability of SSA-GAN to generate images that are both more intricate and more true-to-life, featuring multiple objects set against varied backgrounds, as derived from textual inputs.

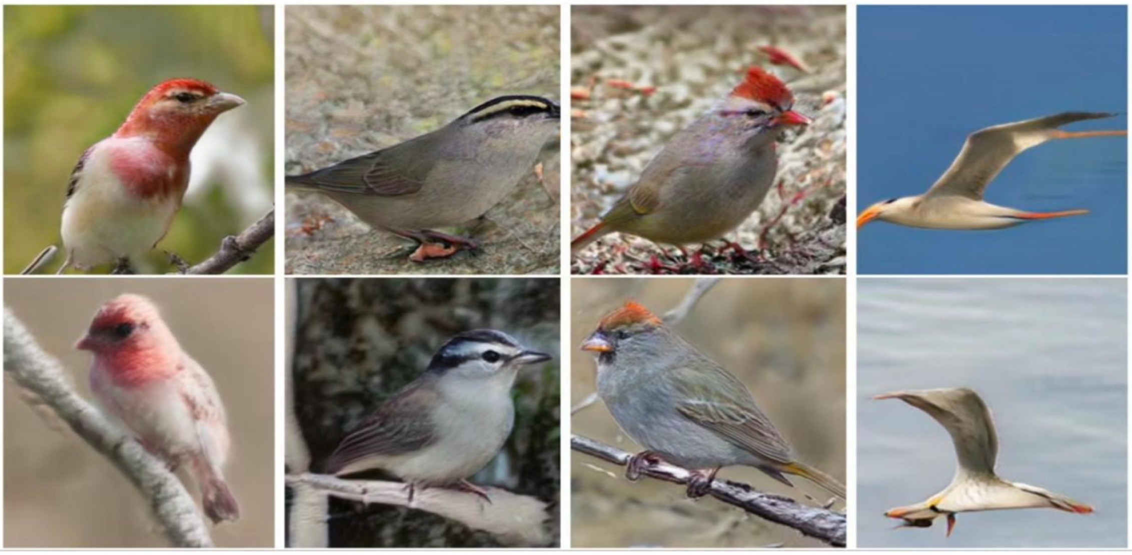

Figure 5. Qualitative comparison between SSA-GAN and Attn-GAN for bird dataset

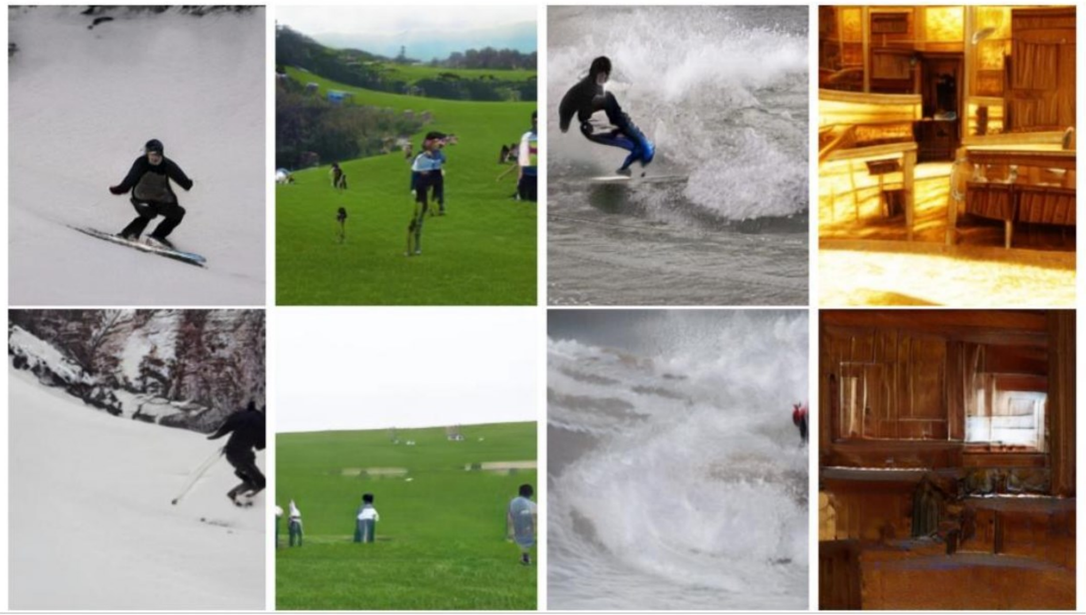

Figure 6. Qualitative comparison between SSA-GAN and Attn-GAN for Coco dataset

And the following Table 1 shows the values of evaluation metrics for different models with different datasets:

Table 1. Performance of IS and FID of At-tnGAN and the method on the Bird and Coco test set

Methods	IS		FID
	Bird	Coco	Coco
Attn-GAN	4.65±0.48	24.85±0.30	40.65
SSA-GAN	5.42±0.25	27.54±0.33	29.41

4. Discussion

SSA-GAN model instigates a significant reduction than Attn-GAN model in the FID score from 40.65 to 29.41 on the COCO dataset, a noteworthy improvement over the existing state-of-the-art performances. The complexity of the COCO dataset, as compared to the Bird dataset, arises from its consistent feature of multiple objects within images and the inherent intricacy of the backgrounds. This accentuates the superior results and clearly stipulates that SSA-GAN is adept at synthesizing high-quality, complex images.

The superiority and effectiveness of the employed SSA-GAN are underscored by comprehensive quantitative evaluations. It has been demonstrated that SSA-GAN can proficiently generate high-quality images that maintain superior semantic consistency. This pertains not only to images laden with detailed attributes but also extends to the more convoluted images comprising multiple objects. This underlines the adaptive and extensive applicability of SSA-GAN in challenging scenarios.

5. Conclusion

This paper conducted a comprehensive comparative analysis of Attn-GAN and SSA-GAN in the context of text-to-image generation. The experiments demonstrated that Attn-GAN excels in generating diverse and visually appealing images, while SSA-GAN focuses on semantic fidelity and fine-grained details. The results of both generated photos and evaluation metrics show that SSA-GAN has a better performance than Attn-GAN. However, due to the susceptible conditions, the choice between these models should be driven by the specific requirements of the task at hand, since the dataset might be diverse and sensitive to some outliers. Some datasets may exhibit a high degree of diversity and sensitivity to outliers, which can impact the suitability of Attn-GAN or SSA-GAN. Therefore, it is essential for practitioners to carefully consider the unique characteristics of their dataset and the desired outcomes of their text-to-image generation task when deciding between these two models. Ultimately, the choice should be made with a nuanced understanding of the strengths and limitations of each model and how they align with the specific needs of the project.

References

[1] Rassin R Ravfogel S Goldberg Y 2022 DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models arXiv preprint arXiv:2210 10606.

[2] He J Shi W Chen K Fu L Dong C 2022 Gcfsr: a generative and controllable face super resolution method without facial and gan priors In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 1889-1898.

[3] Zhong K Sun X Liu G Jiang Y Ouyang Y Wang Y 2023 Attention-based generative adversarial networks for aquaponics environment time series data imputation.

[4] Liao W Hu K Yang M Y Rosenhahn B 2022 Text to image generation with semantic-spatial aware gan In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 18187-18196.

[5] Xu T Zhang P Huang Q Zhang H Gan Z Huang X He X 2018 Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition pp 1316-1324.

[6] Xu T Zhang P Huang Q Zhang H Gan Z Huang X He X 2018 Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition pp 1316-1324.

[7] Zhong K Sun X Liu G Jiang Y Ouyang Y Wang Y 2023 Attention-based generative adversarial networks for aquaponics environment time series data imputation.

[8] Yang B Feng F Wang X 2022 July GR-GAN: Gradual Refinement Text-to-image Generation. In 2022 IEEE International Conference on Multimedia and Expo ICME pp 1-6 IEEE.

[9] Berrahal M Azizi M 2022 Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques. Indones J Electr Eng Comput Sci 25(2) 972-979.

[10] Cruz A P Jaiswal J 2021 Text-to-image classification using attngan with densenet architecture In Proceedings of International Conference on Innovations in Software Architecture and Computational Systems: ISACS 2021 pp 1-17 Springer Singapore.

[11] Liao W Hu K Yang M Y Rosenhahn B 2022 Text to image generation with semantic-spatial aware gan In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 18187-18196.

[12] Liao W Hu K Yang M Y Rosenhahn B 2022 Text to image generation with semantic-spatial aware gan In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 18187-18196.

[13] Mehmood R Bashir R Giri K J 2021 December Comparative Analysis of AttnGAN DF-GAN and SSA-GAN In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking ICAC3N pp 370-375 IEEE.

[14] Liao W Hu K Yang M Y Rosenhahn B 2022 Text to image generation with semantic-spatial aware gan In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 18187-18196.

[15] Zhou L Wu X J Xu T 2023 January COMIM-GAN: Improved Text-to-Image Generation via Condition Optimization and Mutual Information Maximization In International Conference on Multimedia Modeling pp 385-396 Cham: Springer International Publishing.

[16] Tao M Tang H Wu F Jing X Y Bao B K Xu C 2022 Df-gan: A simple and effective baseline for text-to-image synthesis In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 16515-16525.

[17] Liao W Hu K Yang M Y Rosenhahn B 2022 Text to image generation with semantic-spatial aware gan In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 18187-18196.

[18] Lin T Y Maire M Belongie S Hays J Perona P Ramanan D Zitnick C L 2014 Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference Zurich Switzerland September 6-12 2014 Proceedings Part V 13 pp. 740-755 Springer International Publishing.

[19] Wah C Branson S Welinder P Perona P Belongie S 2011 The caltech-ucsd birds-200-2011 dataset.