

Article

TextControlGAN: Text-to-Image Synthesis with Controllable Generative Adversarial Networks

Hyeon Ku and Minhyeok Lee * 

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Republic of Korea

* Correspondence: mlee@cau.ac.kr

Abstract: Generative adversarial networks (GANs) have demonstrated remarkable potential in the realm of text-to-image synthesis. Nevertheless, conventional GANs employing conditional latent space interpolation and manifold interpolation (GAN-CLS-INT) encounter challenges in generating images that accurately reflect the given text descriptions. To overcome these limitations, we introduce TextControlGAN, a controllable GAN-based model specifically designed for text-to-image synthesis tasks. In contrast to traditional GANs, TextControlGAN incorporates a neural network structure, known as a regressor, to effectively learn features from conditional texts. To further enhance the learning performance of the regressor, data augmentation techniques are employed. As a result, the generator within TextControlGAN can learn conditional texts more effectively, leading to the production of images that more closely adhere to the textual conditions. Furthermore, by concentrating the discriminator's training efforts on GAN training exclusively, the overall quality of the generated images is significantly improved. Evaluations conducted on the Caltech-UCSD Birds-200 (CUB) dataset demonstrate that TextControlGAN surpasses the performance of the cGAN-based GAN-INT-CLS model, achieving a 17.6% improvement in Inception Score (IS) and a 36.6% reduction in Fréchet Inception Distance (FID). In supplementary experiments utilizing 128×128 resolution images, TextControlGAN exhibits a remarkable ability to manipulate minor features of the generated bird images according to the given text descriptions. These findings highlight the potential of TextControlGAN as a powerful tool for generating high-quality, text-conditioned images, paving the way for future advancements in the field of text-to-image synthesis.

**Citation:** Ku, H.; Lee, M.TextControlGAN: Text-to-Image
Synthesis with Controllable

Generative Adversarial Networks.

Appl. Sci. **2023**, *13*, 5098. <https://doi.org/10.3390/app13085098>Academic Editors: Erik Kučera,
Oto Haffner and Danica Rosinová

Received: 6 April 2023

Revised: 14 April 2023

Accepted: 18 April 2023

Published: 19 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: generative adversarial networks; text-to-image synthesis; image generation; computer vision

1. Introduction

The rapid advancement of artificial intelligence (AI) technologies, particularly in the domains of machine learning and deep learning, has fostered the development of various AI models [1–3]. Generative models utilizing AI frameworks have garnered significant attention, as they learn from provided sample distributions and create samples that closely mimic the features of the training data [4–7]. These models have been successfully applied to numerous image processing and data analysis tasks due to their ability to generate interesting and realistic samples without requiring the learning of complex structural features [8–10].

Generative adversarial networks (GANs) are a prominent type of generative model, capable of producing realistic samples by learning the latent space of a dataset [11–14]. A GAN consists of two neural networks, namely the generator and the discriminator. The generator receives a random noise vector as input and aims to create fake samples that closely resemble real samples. The discriminator, on the other hand, learns to differentiate between real samples and fake samples generated by the generator. Through an iterative process of deception and detection, the generator and discriminator improve their performance, ultimately synthesizing a generated sample distribution that minimizes the difference from the real sample distribution.

In GAN, all data are considered random variables with their corresponding probability distributions. Each time a random variable is measured, it produces a different value, making it necessary to understand the probability distribution of the random variable to generate numbers that adhere to a specific distribution. By having knowledge of the probability distribution, the statistical properties of the data can be analyzed. GAN generates data randomly to conform to a given probability distribution, resulting in generated data with values comparable to the original data used to determine the probability distribution. Thus, the ultimate goal of GAN is to estimate the probability distribution of unprocessed data, enabling an artificial neural network to generate an infinite number of new datasets that share the exact same probability distribution as the original data. In summary, GAN learns the probability distribution and uses deep learning to generate this distribution.

GANs have been expanded in various domains [6,15,16], including image and video processing as well as image translation, voice signals, 3D rendering, and natural language processing [17–24]. One notable application is text-to-image synthesis, which involves generating synthetic images based on given conditional text inputs [25–29]. By using different input noise vectors, GANs can generate distinct synthetic images corresponding to the same input sentence.

However, conventional GANs face a significant limitation in text-to-image synthesis tasks: due to the use of a random distribution as the input noise vector, controlling the features of the generated samples based on input texts is challenging [30]. To overcome this limitation, conditional GANs (cGANs) have been introduced, allowing for the generation of text-conditional images by incorporating text-conditional encoding vectors in the generator and discriminator [31–33]. The cGAN-based GAN with conditional latent space interpolation and manifold interpolation (GAN-CLS-INT) has been proposed for text-to-image synthesis [25]. Although GAN-CLS-INT can generate natural-looking images from textual descriptions, it often fails to produce images that fully correspond to the given text, with generated images only partially reflecting the context of the input text [34,35].

In this study, we introduce TextControlGAN, a groundbreaking text-to-image synthesis model that expands upon the ControlGAN architecture [36]. The primary innovation of ControlGAN lies in the inclusion of a more advanced classifier, as opposed to the classification component within the cGAN discriminator. ControlGAN comprised three sub-networks: a generator, a discriminator, and a classifier. The generator capitalizes on both the discriminator and the classifier, where the classifier supplies conditional information, and the discriminator refines the authenticity of the generated images. Data augmentation (DA) techniques are employed to train the classifier, thereby improving its classification quality and reducing overfitting issues [37,38].

In TextControlGAN, we substitute the classifier with a regressor that learns to encode text conditional vectors. The regressor aims to estimate the corresponding encoding vector given an image input. DA techniques are applied in the training process of the regressor, paralleling the approach used with classifiers in traditional ControlGAN models. Consequently, the generator learns to generate images that can be accurately estimated by the regressor based on the feedback provided.

The primary objective of our study is to generate realistic images that adhere to the context of given texts using TextControlGAN. We evaluate the model using quantitative methods and conventional GAN metrics [39–41], comparing its performance to other text-to-image synthesis GANs based on the cGAN framework. This study's main contributions are fourfold: (1) the proposal of a GAN architecture capable of generating images conditioned on given text descriptions, (2) integrating neural network structures using independent regressors to train three neural network structures: in the regressor, it learns by estimating the text encoding vector for the given image, (3) the experimental validation of TextControlGAN's capacity to generate realistic images, and (4) the implementation of data augmentation techniques for the independent regressor without impacting the discriminator.

The rest of this paper is organized as follows: in Section 2, we provide the necessary background on GANs, conditional GANs, and controllable GANs, as well as related work in text-to-image synthesis. In Section 3, we describe our proposed TextControlGAN model and the training details. In Section 4, we present and discuss our experimental results. Finally, we conclude the paper and provide directions for future research in Section 5.

2. Background

2.1. Generative Adversarial Networks

The GAN, first proposed by Goodfellow et al. in 2014 [11], is a neural network architecture consisting of two adversarial networks: a generator and a discriminator. The generator network is responsible for generating synthetic data, while the discriminator network aims to distinguish between real and generated data. This competition between the two networks compels them to iteratively learn and enhance their abilities to generate and differentiate data, respectively. The GAN training process can be represented by the following equation, where D denotes the discriminator network and G represents the generator network:

$$\min_G \max_D O(D, G) = \mathbb{E}_{x \sim \mathbb{P}_x} [\log D(x)] + \mathbb{E}_{z \sim \mathbb{P}_z} [\log(1 - D(G(z)))], \quad (1)$$

where $O(D, G)$ represents the objective function of GAN training; x denotes image samples; and z denotes random noise vectors sampled from a specific distribution. $\mathbb{E}_{x \sim \mathbb{P}_x}$ and $\mathbb{E}_{z \sim \mathbb{P}_z}$ denote the original data distribution and noise distribution, respectively.

The primary goal of GAN is to maximize the $O(D, G)$ function from the perspective of D while minimizing it from the perspective of G . Consequently, if image x is used as the input of D , the discriminator aims to output one, i.e., $D(x) = 1$; otherwise, if a synthetic image becomes the input, it is expected to produce zero, i.e., $D(G(z)) = 0$. However, the generator has an opposite objective to the discriminator, in which $D(G(z)) = 1$.

This can be interpreted as the generator attempting to deceive the discriminator by generating increasingly realistic images that the discriminator cannot distinguish from real samples. This adversarial training process pushes both networks to improve, resulting in the generator producing high-quality synthetic data that closely resembles the distribution of the real data.

2.2. Conditional GANs

A cGAN [31] is a type of GAN that can generate data conditioned on given conditions. For example, a cGAN can generate images of faces given text descriptions of the desired facial features as conditions. cGANs can be used to generate images corresponding to a given input, such as generating an image of a face from its description. The cGAN training process is similar to the GAN training process, with the addition of conditions y on both the generator and discriminator networks:

$$\min_G \max_D O(D, G) = \mathbb{E}_{(x,y) \sim \mathbb{P}_{(x,y)}} [\log D(x|y)] + \mathbb{E}_{z \sim \mathbb{P}_z, y \sim \mathbb{P}_y} [\log(1 - D(G(z|y)))]. \quad (2)$$

where $O(D, G)$ represents the objective function of cGAN training; x denotes image samples; y denotes conditional vector; and z denotes random noise vectors sampled from a specific distribution. To implement this concept, the conditional vector, y , is generally concatenated or multiplied to the input noise vector, z , in the generator, whereas y is concatenated or multiplied to the input layer or penultimate layer in the discriminator. After the cGAN training, the generator can produce conditional samples with different y values.

2.3. Controllable GANs

Recent advancements have been made in conditional-based methods for generating realistic samples. For instance, the Auxiliary Classifier GAN (ACGAN) [35] is a conventional method for generating conditional samples that employs a classification layer in

the discriminator and exhibits satisfactory results. However, the auxiliary classifier was found to be insufficient for generator training [42–44], as it encountered issues of overfitting the classifier. Moreover, incorporating Data Augmentation (DA) techniques in ACGAN's training process proved challenging [45], given that the additional classifier was attached to the discriminator. The utilization of DA in GAN structures' learning has been identified as a key concern within this method.

To address this issue, ControlGAN [36] was proposed, highlighting a trade-off in performance when using DA for ACGAN. ControlGAN aims to separate the classifier from the discriminator, enabling the use of DA without impeding GAN training by leveraging DA solely for independent classifiers. Specifically, ControlGAN introduces three distinct neural network structures and a target function designed to maintain a balanced training process for a generator that is concurrently trained by two separate network modules.

By successfully separating the classifier from the discriminator, ControlGAN presents a promising solution for enhancing the performance of conditional GANs, thus contributing to the ongoing development and optimization of GAN architectures for a myriad of applications within the realm of generative models.

2.4. Text-to-Image Synthesis

Text-to-image synthesis refers to the process of producing images from textual descriptions, which presents a formidable challenge as it necessitates the comprehension of textual meaning and its subsequent conversion into corresponding visual concepts. This field offers numerous applications, such as the creation of photorealistic avatars based on textual descriptions. Multiple approaches have been proposed to address text-to-image synthesis, including retrieval-based methods and generation-based methods.

Retrieval-based methods involve selecting images from an extensive database that correspond to a given textual description. These selected images are subsequently combined to form a new image that adheres to the desired specifications. One notable advantage of retrieval-based methods lies in their capacity to generate high-quality images, provided that a database of high-quality images is available. Furthermore, these methods tend to be faster than generation-based methods since they do not necessitate the creation of images from scratch. However, retrieval-based methods are contingent upon the availability of a large image database, which may prove challenging to obtain. Additionally, these methods may struggle to retrieve all pertinent images from the database, particularly in cases where the textual description is lengthy or complex.

On the other hand, generation-based methods entail the creation of images from scratch, guided by the given textual description. These methods typically employ a generative model, such as a GAN, to generate images that align with the textual description. A primary advantage of generation-based methods is their ability to generate images even in the absence of an image database. Nonetheless, training generation-based methods can be arduous, and the resulting images may not exhibit the same level of realism as those produced by retrieval-based methods. As GANs have demonstrated outstanding performance in general image synthesis [35,46], GAN-based text-to-image synthesis has recently garnered significant attention. GAN-based models generally introduce the cGAN architecture, where encoding vectors of conditional texts are used as y in Equation (2). As one of the studies with this idea, GAN-INT-CLS [16] has demonstrated that this model can successfully generate synthetic images that follow the contexts in given sentences.

The GAN-INT-CLS model is a GAN architecture that accepts both image and text data as input, effectively generating images that correspond to specified textual descriptions. This model comprised two main components: a text encoder responsible for converting input text into a latent vector, and an image generator tasked with producing an image derived from the latent vector. The discriminator is trained to differentiate between authentic and counterfeit (text, image) pairs.

To provide contextual information during the discriminator's training process, pairs comprising a genuine image and mismatched text are incorporated as counterfeit examples.

Furthermore, GAN-INT-CLS introduces interpolation within the embedding vector, which facilitates training with continuous manifold data. Consequently, this enables the generation of images that adhere to the features delineated in corresponding textual descriptions, even in cases where the model has not previously encountered those precise combinations.

By employing such a sophisticated approach, the GAN-INT-CLS model demonstrates its potential for generating high-quality images in response to a diverse range of textual descriptions. Thus advancing the field of text-to-image synthesis and expanding its applicability in various domains.

There are several other existing studies related to the proposed method. Reed et al. [47], were able to generate 64×64 resolution images corresponding to textual descriptions by building on the Generative Adversarial What-Where Networks [48]. To improve the generative process, StackGAN [26] was proposed to divide the process into two stages, with the first stage generating low-resolution images with basic visual information and the second stage generating high-resolution images with more detailed features. Furthermore, the authors of [27] enhanced the StackGAN method to deal with both conditional and unconditional generative tasks while stabilizing the training of GANs by approximating multiple distributions jointly. To explore class information from text descriptions, TACGAN [49] was proposed, which employs a text-conditioned auxiliary classifier to diversify synthetic images and enhance their structural coherence.

3. Methods

3.1. Text-to-Image Synthesis with Controllable GAN Framework

In this paper, we present a text-to-image synthesis model based on the ControlGAN [36] framework as a viable alternative to cGAN. One primary advantage of employing a ControlGAN lies in its utilization of an independent classifier, while in cGAN, the discriminator assumes responsibility for classification. This approach enables us to train the classifier with DA methods, incorporating modified and slightly distorted images during the training process.

The proposed model, TextControlGAN, builds upon the concept of ControlGAN to condition texts. Given that text encoding vectors possess continuous values, TextControlGAN introduces a regressor, distinguishing it from conventional ControlGAN. The regressor fulfills a role analogous to the classifier in ControlGAN, as it seeks to estimate text encoding vectors from given images. By employing DA methods for training the regressor, which slightly distort input images during the process, the regressor model can be trained with reduced overfitting issues and exhibit enhanced performance.

Consequently, TextControlGAN comprises three distinct neural network structures: a generator, a discriminator, and a regressor. Figure 1 illustrates the comprehensive structure of TextControlGAN. The generator generates synthetic images using corresponding text embedding vectors and noise vectors as inputs. These images are then compared to real-world images by the discriminator. Deviating from cGAN and GAN-INT-CLS models, the discriminator is solely responsible for its primary task, which is the binary classification of authentic versus fake imitations. Instead of involving the discriminator in the conditional text learning, conditional texts are learned exclusively by another specialized structure called the regressor. Within the regressor, DA methods are employed to improve performance, enabling the generator to learn the conditional information more effectively than if it relied only on the discriminator in cGAN and GAN-INT-CLS. The objective of TextControlGAN learning can be represented as follows:

$$\min_G \max_D O(D, G) = \mathbb{E}_{(x,y) \sim \mathbb{P}_{(x,y)}} [\log D(x)] + \mathbb{E}_{z \sim \mathbb{P}_z(z), y \sim \mathbb{P}_y} [\log(1 - D(G(z|y)))], \quad (3)$$

$$\min_R O_R(R) = \mathbb{E}_{x \sim \mathbb{P}_x} \left[\sum (y - R(x))^2 \right], \quad (4)$$

$$\min_G O_R(G) = \mathbb{E}_{z \sim \mathbb{P}_z, y \sim \mathbb{P}_y} \left[\sum (y - R(G(z|y)))^2 \right] \tag{5}$$

where R represents the regressor and O_R is the objective function of the regressor.

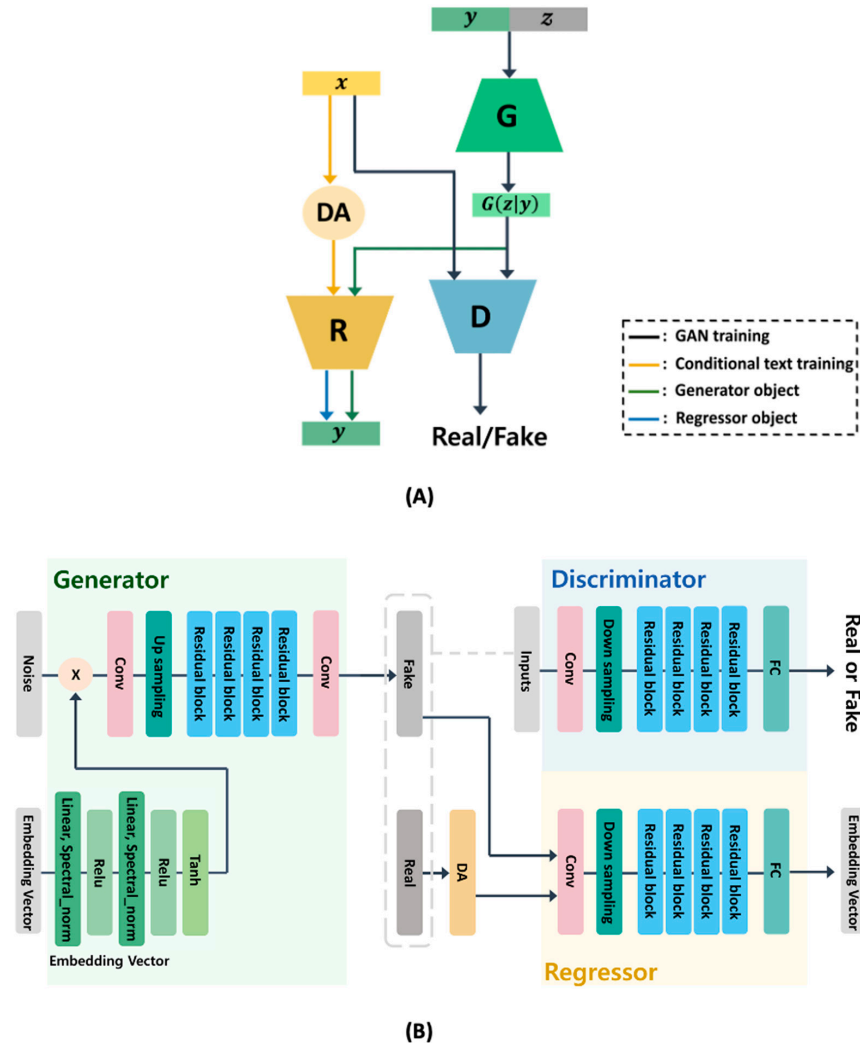


Figure 1. Structure of the proposed TextControlGAN. The G , D , and R denote the generator, discriminator and regressor, respectively in (A). z , x , and y denote noise variables for the generator, a real image dataset, and corresponding text embedding, respectively. (B) indicates the training process in terms of deep learning layers of each model.

The TextControlGAN objective as expressed in Equation (3) bears resemblance to that of cGAN; however, a key difference is that, in TextControlGAN, the discriminator (D) does not learn the conditional information with y . As demonstrated in Equations (3) and (5), the generator has dual objectives: it seeks to deceive the discriminator in order to be classified as real, while simultaneously aiming to have the correct text encoding vectors estimated by the regressor, in accordance with the generated images.

All structures in TextControlGAN are composed of residual blocks; the generator, discriminator, and regressor each contain four such blocks. Each block comprises two convolutional layers. In the generator, upsampling by a factor of two is executed using the nearest neighbor method within each residual block. The generator is composed of 10 convolutional layers with 96 kernels, where the first and the last two layers are convolutional layers and four residual blocks with two convolutional layers each. Similarly, downsampling is performed on the first three residual blocks of both the dis-

criminator and regressor, effectively reducing dimensionality by eight times. The regressor accepts augmented images as inputs with the purpose of estimating corresponding text encoding vectors.

The utilization of the TextControlGAN framework is anticipated to yield superior performance in conditional text learning. This expectation arises from the fact that the regressor is trained more effectively than the text-learning component of the discriminator in GAN-INT-CLS. Consequently, the generator learns from this more proficient regressor and, in turn, generates more realistic and text-conditioned images. Moreover, as the discriminator's training is specifically focused on GAN training, this also contributes to an overall enhancement in image quality.

3.2. Training Details of TextControlGAN

To train TextControlGAN, the Adam optimizer [50] was used with the parameters of $\beta_1 = 0$ and $\beta_2 = 0.999$; the learning rate of the generator and discriminator was set to 0.0001 [51]. In order to address the mode collapse issue in the GAN training, the Wasserstein loss [30] and hinge loss [52] functions were used in the generator and discriminator, respectively. For the regressor, the Adam optimizer was employed with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which are conventional values for deep learning models. The learning rate of the regressor was set at 0.0005. The model was trained on one RTX 3090GPU with 24 GB memory with a batch size of 128. The version of Python was 3.8.12 and the Pytorch version was 1.9.0 with CUDA 11.1.

4. Result and Discussion

4.1. Quantitative Evaluation with a Bird Image Dataset

In the evaluation of the proposed model, the Caltech-UCSD Birds-200 (CUB) [40] dataset was employed. The CUB dataset consists of 11,788 images for 200 different types of birds. Each image is annotated with attributes of birds, which correspond to the colors and shapes of the birds. By using this annotation, text captions to describe the images have been employed in GAN-INT-CLS, where 27,450 pairs of images and corresponding texts exist; whereas there are 11,788 images in the dataset, these pairs are composed of multiple texts for particular images. The text encoding vector y has a dimension of 1024 for each text. In this experiment, images with a 64×64 resolution were produced and evaluated; thus, real images in the dataset were down-scaled with the same size in the preprocessing step.

For the quantitative evaluation metrics, Inception score (IS) [39] and Frechet Inception distance (FID) [41] were used, where these metrics are conventional methods to evaluate generative models. These metrics commonly use a pre-trained Inception network, and specific layer outputs of the network are employed to assess the quality of generated images; a high IS and a low FID indicate superior performance. The IS was measured with ten different sets of 5000 samples, which is a conventional method for calculating IS; then, the average IS and standard deviation were computed.

In the evaluation, two baseline models were compared to the proposed model. First, cGAN-based GAN-INT-CLS was evaluated with IS and FID in order to demonstrate the performance of the DA methods integrated with the independent regressor. Additionally, TextControlGAN without DA was assessed to verify that the independent regressor cannot solely perform without DA methods.

Table 1 showcases the performance of various models, including recent models, such as AttnGAN and DM-GAN [53], highlighting that the proposed TextControlGAN outperforms other methods, including GAN-INT-CLS. TextControlGAN achieved an IS of 4.41 ± 0.02 , signifying a 17.6% improvement compared to GAN-INT-CLS in terms of IS. This demonstrates the effectiveness of the proposed framework, which employs an independent regressor in conjunction with DA techniques, in accurately learning conditional text. Moreover, the FID for TextControlGAN decreased by 36.6% relative to GAN-INT-CLS, further substantiating TextControlGAN's superior performance.

Table 1. Comparison of results in terms of Inception Score (IS) and Fréchet Inception Distance (FID). Note that IS and FID of AttnGAN and DM-GAN were measured with different resolutions, model structures, and hyperparameters, resulting in a direct comparison to TextControlGAN and those models being meaningless.

| Model | IS | FID |
|-----------------------|-------------|--------|
| TextControlGAN | 4.41 ± 0.02 | 57.92 |
| TextControlGAN w/o DA | 2.57 ± 0.02 | 105.21 |
| GAN-INT-CLS | 3.75 ± 0.02 | 91.37 |
| AttnGAN | 4.42 ± 0.07 | 20.85 |
| DM-GAN | 4.66 ± 0.06 | 15.10 |

However, when DA techniques were excluded from the proposed architecture (TextControlGAN w/o DA), the performance declined compared to that of TextControlGAN. This outcome is believed to stem from overfitting issues in the regressor, as DA techniques generally reduce overfitting in neural network models. As a result, this finding highlights the necessity of incorporating DA techniques when training the proposed architecture to ensure optimal performance.

An evaluation of the proposed model, TextControlGAN, was conducted to assess its computing time per epoch, in comparison to GAN-INT-CLS. The results indicated that the proposed model required approximately 447.56 s, while GAN-INT-CLS necessitated approximately 637.55 s per epoch. Despite being 29.8% less computationally efficient than GAN-INT-CLS, TextControlGAN surpassed GAN-INT-CLS in terms of performance, as evidenced by the presented results.

4.2. Quantitative Evaluation of Generated Images by TextControlGAN

In this section, we present several images generated by TextControlGAN alongside multiple textual examples. Figure 2 showcases the results obtained from TextControlGAN and other baseline models. Generally, when generating text-conditional images, TextControlGAN consistently outperformed the other methods, which occasionally failed to produce images that adhered to the given textual conditions.

For instance, in the first text example, one of the textual conditions specified that “The belly is white”. However, the alternative models were unable to accurately capture this condition in their generated images. In contrast, all images generated by TextControlGAN successfully depicted a bird with a white belly. Moreover, in the third example, where the objective was to generate images featuring an “orange beak”, TextControlGAN predominantly produced images with the desired attribute. Conversely, many images generated by the other models not only lacked the orange color in the beak but also failed to incorporate the color in other body parts.

These examples demonstrate the superior capability of TextControlGAN in capturing the nuances of textual descriptions and effectively translating them into accurate visual representations, thereby highlighting its potential for generating high-quality, text-conditional images in various applications.

4.3. Evaluation with Higher Resolution Images

In supplementary experiments, TextControlGAN was employed to generate 128×128 resolution images, which were then assessed using untrained text inputs. The CUB dataset, featuring 128×128 resolution images, served as the training set. The TextControlGAN architecture for this experiment remained identical to the previous experiment involving 64×64 images, with the exception of the number of layers. Additional residual modules, each comprising two convolutional layers and either downsampling or upsampling, were introduced. Specifically, the generator was equipped with an additional residual module featuring upsampling, while the discriminator and regressor each received an additional residual module with downsampling. The number of training iterations and hyperparameters for the model were set to match those of the previous experiment.



Figure 2. The result comparison of the proposed model and the baseline.

Figure 3 displays the generated images with random text inputs, revealing that the 128×128 resolution images corresponded to bird images and exhibited features characteristic of the CUB training set. Moreover, generated images utilizing untrained text inputs were evaluated. As ControlGAN-based models excel in the intricate modification of minor features, this strength was assessed through alterations in the shape of the bird images' bills. Figure 4 presents the generated images using untrained text inputs with modifications in minor features, demonstrating that the images produced by TextControlGAN adhered to the corresponding text inputs. In particular, it was observed that minor features, such as the shape of the bill, could be effectively controlled by TextControlGAN's input text. Consequently, TextControlGAN successfully generated bird images with both long-pointed and short-pointed bills, further illustrating the model's capacity for precise control over visual features.



Figure 3. Generated 128×128 resolution images by TextControlGAN.





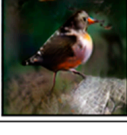
| Generated Images | Given Texts |
|--|--|
|  | a large bird with black body |
|  | a large bird with orange body |
|  | a large bird with orange body and a black head |
|  | a large bird with orange body and a black head with a very long pointed bill |
|  | a large bird with orange body and a black head with a very short pointed bill |

Figure 4. Generated 128×128 resolution images by TextControlGAN using untrained texts with minor modifications.

5. Conclusions

In this paper, we have introduced a novel GAN-based text-to-image synthesis model, termed TextControlGAN. While existing models have incorporated the conditional GAN (cGAN) framework in their training processes, TextControlGAN leverages the ControlGAN-based framework to enhance the model's text-conditioning capabilities. Within TextControlGAN, an independent regressor is implemented along with Data Augmentation (DA) techniques for its training.

Evaluations were conducted using a bird image dataset containing approximately 30,000 pairs of images and corresponding textual descriptions. The results revealed that TextControlGAN achieved a 17.6% improvement in Inception Score (IS) and a 36.6% reduction in Fréchet Inception Distance (FID) when compared to GAN-INT-CLS, a cGAN-based model. In the comparison of generated images, it was observed that those produced by TextControlGAN adhered to the conditional text inputs, while alternative models occasionally failed to accurately reflect the contextual information of the text inputs.

By incorporating an independent regressor and DA techniques, the proposed TextControlGAN learning method was applied to the GAN-INT-CLS model structure, leading to exceptional performance. The versatility of the proposed method allows for its easy adaptation to other model structures, thus contributing to the exploration and development of additional models in future research endeavors.

Author Contributions: Conceptualization, H.K. and M.L.; methodology, H.K. and M.L.; software, H.K.; validation, H.K.; formal analysis, H.K.; investigation, H.K. and M.L.; writing—original draft preparation, H.K. and M.L.; supervision, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2021R1F1A1050977).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Caltech-UCSD Birds-200-2011 Dataset used in this paper is available at http://www.vision.caltech.edu/datasets/cub_200_2011/ (accessed on 5 April 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Samek, W.; Wiegand, T.; Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv* **2017**, arXiv:1708.08296.
2. Lee, Y.-L.; Tsung, P.-K.; Wu, M. Technology trend of edge ai. In Proceedings of the 2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), Hsinchu, Taiwan, 16–19 April 2018; pp. 1–2.
3. Ongsulee, P. Artificial intelligence, machine learning and deep learning. In Proceedings of the 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 22–24 November 2017; pp. 1–6.
4. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. *arXiv* **2015**, arXiv:1511.05644.
5. Mescheder, L.; Nowozin, S.; Geiger, A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2391–2400.
6. Wang, Z.; She, Q.; Ward, T.E. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–38. [[CrossRef](#)]
7. Chen, X.; Li, Y.; Yao, L.; Adeli, E.; Zhang, Y.; Wang, X. Generative adversarial u-net for domain-free few-shot medical diagnosis. *Pattern Recognit. Lett.* **2022**, *157*, 112–118. [[CrossRef](#)]
8. Wang, F.; Ma, Z.; Zhang, X.; Li, Q.; Wang, C. Ddsg-gan: Generative adversarial network with dual discriminators and single generator for black-box attacks. *Mathematics* **2023**, *11*, 1016. [[CrossRef](#)]
9. Kim, M.; Song, M.H. High performing facial skin problem diagnosis with enhanced mask r-cnn and super resolution gan. *Appl. Sci.* **2023**, *13*, 989. [[CrossRef](#)]
10. Wang, Y.; Zhang, S. Prediction of tumor lymph node metastasis using wasserstein distance-based generative adversarial networks combing with neural architecture search for predicting. *Mathematics* **2023**, *11*, 729. [[CrossRef](#)]
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [[CrossRef](#)]
12. Hitawala, S. Comparative study on generative adversarial networks. *arXiv* **2018**, arXiv:1801.04271.
13. Donahue, J.; Krähenbühl, P.; Darrell, T. Adversarial feature learning. *arXiv* **2016**, arXiv:1605.09782.
14. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
15. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3313–3332. [[CrossRef](#)]
16. Aggarwal, A.; Mittal, M.; Battineni, G. Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100004. [[CrossRef](#)]
17. Tulyakov, S.; Liu, M.-Y.; Yang, X.; Kautz, J. Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1526–1535.
18. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
19. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
20. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8821–8831.
21. Kim, I.; Lee, M.; Seok, J. Icegan: Inverse covariance estimating generative adversarial network. *Mach. Learn. Sci. Technol.* **2023**, *4*, 025008. [[CrossRef](#)]
22. Ko, K.; Yeom, T.; Lee, M. Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains. *Neural Netw.* **2023**, *162*, 330–339. [[CrossRef](#)]
23. Lee, M.; Seok, J. Score-guided generative adversarial networks. *Axioms* **2022**, *11*, 701. [[CrossRef](#)]
24. Kim, W.; Kim, S.; Lee, M.; Seok, J. Inverse design of nanophotonic devices using generative adversarial networks. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105259. [[CrossRef](#)]
25. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1060–1069.
26. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
27. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1947–1962. [[CrossRef](#)]

28. Qi, Z.; Fan, C.; Xu, L.; Li, X.; Zhan, S. Mrp-gan: Multi-resolution parallel generative adversarial networks for text-to-image synthesis. *Pattern Recognit. Lett.* **2021**, *147*, 1–7. [[CrossRef](#)]
29. Tan, Y.X.; Lee, C.P.; Neo, M.; Lim, K.M. Text-to-image synthesis with self-supervised learning. *Pattern Recognit. Lett.* **2022**, *157*, 119–126. [[CrossRef](#)]
30. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5769–5779.
31. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
32. Shin, Y.; Qadir, H.A.; Balasingham, I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access* **2018**, *6*, 56007–56017. [[CrossRef](#)]
33. Gauthier, J. *Conditional Generative Adversarial Nets for Convolutional Face Generation*; Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition; Winter Semester; University of Stanford: Stanford, CA, USA, 2014; p. 2.
34. Miyato, T.; Koyama, M. Cgans with projection discriminator. *arXiv* **2018**, arXiv:1802.05637.
35. Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.
36. Lee, M.; Seok, J. Controllable generative adversarial network. *IEEE Access* **2019**, *7*, 28158–28169. [[CrossRef](#)]
37. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
38. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
39. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 2226–2234.
40. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-UCSD Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
41. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
42. Akbarizadeh, G. A new statistical-based kurtosis wavelet energy feature for texture recognition of sar images. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 4358–4368. [[CrossRef](#)]
43. Karimi, D.; Akbarizadeh, G.; Rangzan, K.; Kabolizadeh, M. Effective supervised multiple-feature learning for fused radar and optical data classification. *IET Radar Sonar Navig.* **2017**, *11*, 768–777. [[CrossRef](#)]
44. Raeisi, A.; Akbarizadeh, G.; Mahmoudi, A. Combined method of an efficient cuckoo search algorithm and nonnegative matrix factorization of different zernike moment features for discrimination between oil spills and lookalikes in sar images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4193–4205. [[CrossRef](#)]
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
46. He, J.; Zheng, J.; Shen, Y.; Guo, Y.; Zhou, H. Facial image synthesis and super-resolution with stacked generative adversarial network. *Neurocomputing* **2020**, *402*, 359–365. [[CrossRef](#)]
47. Yan, F.; Mikolajczyk, K. Deep correlation for matching images and text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3441–3450.
48. Chi, J.; Peng, Y. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1173–1187. [[CrossRef](#)]
49. Dash, A.; Gamboa, J.C.B.; Ahmed, S.; Liwicki, M.; Afzal, M.Z. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv* **2017**, arXiv:1703.06412.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv* **2018**, arXiv:1802.05957.
52. Moore, R.C.; DeNero, J. L1 and l2 regularization for multiclass hinge loss models. In Proceedings of the Symposium on Machine Learning in Speech and Natural Language Processing, Bellevue, WA, USA, 21 June 2011.
53. Ye, H.; Yang, X.; Takac, M.; Sunderraman, R.; Ji, S. Improving text-to-image synthesis using contrastive learning. *arXiv* **2021**, arXiv:2107.02423.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.