Enhanced Text-to-Image Synthesis Conditional Generative Adversarial Networks

Yong Xuan Tan, Chin Poo Lee, *Member, IAENG,* Mai Neo, Kian Ming Lim, *Member, IAENG,* and Jit Yan Lim

***Abstract*—The text-to-image synthesis aims to synthesize an image based on a given text description, which is especially useful for applications in image editing, graphic design, etc. The main challenges of text-to-image synthesis are to generate images that are visually realistic and semantically consistent with the given text description. In this paper, we proposed some enhancements to the conditional generative model that is widely used for text-to-image synthesis. The enhancements include text conditioning augmentation, feature matching, and L1 distance loss function. The text conditioning augmentation expands the text embedding feature space to improve the semantic consistency of the model. The feature matching motivates the model to synthesize more photo-realistic images and enrich the image content variations. Apart from that, the L1 distance loss allows the model to generate images that have high visual resemblance to the real images. The empirical results on the CUB-200-2011 dataset demonstrate that the text-to-image synthesis conditional generative model with the proposed enhancements yield the highest Inception score and Structural Similarity Index.**

***Index Terms*—text-to-image-synthesis, generative adversarial network, conditional generative adversarial networks, GANs, cGANs**

I. Introduction

TEXT-TO-IMAGE synthesis is the task of producing images based on the given text descriptions. The core challenge of text-to-image synthesis lies in the creativity of the language and image. A text description can correspond to a large number of images. Likewise, an image can be described in many different text descriptions. In addition, given a text description, the synthesized images need to be semantically consistent and visually realistic. Semantic consistency means that the generated image contents should match the characteristics mentioned in the text descriptions. Not only that, the synthesized images should also contain both local and global details, specifically the object parts (local) and background (global) information.

Generative Adversarial Networks (GANs) [1] were initially proposed for generating complex real-world alike images. GANs were trained on a large image dataset with objects in the same class. After that, a variant of GANs known as conditional Generative Adversarial Networks (cGANs) [2] were introduced where the networks generate images conditioned on the class label. Later, the authors in [3] proposed a variant of cGANs, known as GAN-INT-CLS to make the networks generate images conditioned on text description instead of class label for text-to-image synthesis.

Inspired by the success of GAN-INT-CLS in generating images conditioned on text description, this paper adopts GAN-INT-CLS as the baseline model for text-to-image synthesis. The main challenges of text-to-image synthesis are to produce images that are visually realistic and semantically consistent with the text description. Despite the benefits offered by GANs in generating images, GANs are inherently prone to training instability. To deal with these challenges, this paper proposes some enhancements to GAN-INT-CLS. The enhancements include text conditioning augmentation, feature matching and L1 distance loss. The text conditioning augmentation function mitigates the discontinuity issue in the text latent data and improves the quality of the synthesized image to be more semantically consistent to the text description. The feature matching function improves the model's capability to synthesize more photo-realistic and richer variety images. The L1 distance loss function makes the visual information of fake images closer to the real images thus synthesizing the images that are more visually realistic. The proposed enhancements collectively improve the training stability of the text-to-image synthesis cGANs. The contributions of this paper are:

- The text conditioning augmentation is leveraged to alleviate the discontinuity in the text latent data and improve the semantic consistency of the model.
- The feature matching function is integrated to synthesize more visually realistic images with diversifying image contents to mitigate the mode collapse issue.
- The L1 distance loss is incorporated to enhance the visual similarity and semantic consistency of the synthesize and real images.
- The proposed enhancements improve the training stability of the text-to-image synthesis cGANs.

This paper is structured as follows: Section II reviews the current text-to-image synthesis techniques. Section III explains the proposed enhanced text-to-image GANs in detail. Section IV presents the experiments and analysis of the proposed enhanced text-to-image GANs. Lastly, Section V concludes the paper.

II. Related Work

Text-to-image synthesis has spurred great interest among the computer vision research community in recent years

Manuscript received October 9, 2021; revised February 8, 2022. This work was supported by the Fundamental Research Grant Scheme of the Ministry of Higher Education FRGS/1/2021/ICT02/MMU/02/4 and TM R&D grant RDTC/190995.

Y.X. Tan is a student in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: 1141124379@student.mmu.edu.my)

C.P. Lee, the corresponding author, is a Senior Lecturer in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: cplee@mmu.edu.my)

M. Neo is a Professor in the Faculty of Creative Multimedia, Multimedia University, Cyberjaya, 63100 Malaysia. (e-mail: neo.mai@mmu.edu.my)

K. M. Lim is a Senior Lecturer in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: kmlim@mmu.edu.my)

J.Y. Lim is a student in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: 1141124378@student.mmu.edu.my)

[4]–[6]. Many approaches have been proposed for text-to-image synthesis. In work [3], the authors proposed GAN-INT-CLS that is able to synthesize images based on the text description. The GAN-INT-CLS adopted the architecture of deep convolutional GANs [7], where it consists of a set of convolutional layers to produce the images. The text description is encoded as an embedding [8]and used to condition the generated images. Even though GAN-INT-CLS is able to produce the image with the correct object, it cannot control the object details such as location and pose. To deal with this issue, [9] proposed Generative Adversarial What-Where Network (GAWWN) which is able to generate images from the description with the desired location. GAWWN locates the object based on the bounding box or keypoint as input.

As the image produced by GAN-INT-CLS lacks some visual detail, [10] proposed StackGAN which is able to produce higher dimensional images with more complete visual details. StackGAN is divided into 2 stages: Stage-I produces small 64×64 pixels images with a rough object structure while Stage-II enhances those details and expands them into larger 256×256 pixels images. Furthermore, [11] proposed StackGAN++ as the improved StackGAN. This approach contains multiple generators and discriminators in a tree-like structure. Every generator produces a different scale of images and serves as the input to the next generator to finally construct the larger output.

To produce images containing the object mentioned in the text description, [12] proposed text-conditioned auxiliary classifier GANs (TAC-GAN) which uses the object labels as extra input during the synthesis process. The discriminator in TAC-GAN classifies whether the input image is real or fake and also predicts the class label. The use of class labels improves the structural coherence of the image. [13] proposed Attentional Generative Adversarial Network (AttnGAN) which uses attention techniques and multi-stage architecture to generate fine-grained visual information.

III. Enhanced Text-to-Image Synthesis Model

This section describes the proposed enhanced text-to-image synthesis model in detail. As GANs were initially used for generating images based on the training images and random noise, they are susceptible to training instability. This is due to the generator being too focused on defeating the discriminator, hence using the wrong features to generate images even though the generated images may not be visually realistic. Furthermore, the images generated by GANs are not associated with any text description. To that end, the conditional GANs (cGANs) were proposed where the images are synthesized based on the given class labels. Later, a variant of cGANs, known as GAN-INT-CLS were introduced where the generated images are conditioned on the text descriptions instead of class labels.

In view of this, to produce images that are semantically consistent to the given text description, this paper leverages GAN-INT-CLS that synthesize images conditioned on the text description. The proposed enhanced text-to-image synthesis cGANs consist of two networks: generator and discriminator. The generator receives random noise and text description as input and synthesizes images as output. On the other hand, the discriminator receives the real images, synthesized images and text description as input. The discriminator subsequently predicts whether the generated image is real or fake. The overall architecture of the enhanced text-to-image synthesis cGANs is illustrated in Fig. 1.

A. Preliminaries

The text-to-image synthesis cGANs comprises two networks: a generator and a discriminator. The generator is denoted as G and the discriminator is denoted as D. The input noise vector is randomly sampled from multivariate standard normal distribution as $z \in R^Z \sim N(0,1)$ where Z denotes the length of the vector, 0 represents the zeros vector, and 1 defines the identity matrix with the dimension of $Z \times Z$. The text description is denoted as t and is encoded by a pretrained character-level convolutional neural networks with recurrent neural network (char-CNN-RNN) text encoder [8] $f_\psi(t)$ to produce the 1024-dimensional embedding h where $h = f_\psi(t)$.

B. Network Architecture

This section presents the network architecture of the generator G and discriminator D.

1) Generator: Table I presents the architecture of the generator. The generator consists of two components, namely conditioning augmentation function f_{CA} and modeling f_ϵ.

The text conditioning augmentation function f_{CA} is introduced to generate more conditioning variables through the augmentation for generator learning. The text conditioning augmentation function f_{CA} is composed of a dense layer with 1024 nodes and activated by the Rectified Linear Unit function (ReLU). The input of the text conditioning augmentation function f_{CA} is the text embedding h and the output is the augmented text embeddings.

The modeling component f_ϵ acts as the text-to-image synthesizer in the generator. The modeling component f_ϵ consists of a dense layer and 4 deconvolutional layers. The first dense layer has 356 nodes and is activated by the ReLU function. The output of the dense layer is expanded to fit into the subsequent deconvolutional layer. The first three deconvolutional layers use 512, 256, and 128 filters of the size 4×4, with batch normalization and ReLU activation function, respectively. The fourth deconvolutional layer uses 64 filters of the size 4×4 and tanh activation function. The final output of the modeling component f_ϵ is the synthesized image with the resolution of 128×128 pixels.

TABLE I
THE CONFIGURATION AND NETWORK ARCHITECTURE OF THE GENERATOR.

Component	Configuration
Text Conditioning Augmentation f_{CA}	dense 1024, ReLU
Modeling f_ϵ	dense 356, ReLU convTranspose filter 4×4×512, batchnorm, ReLU convTranspose filter 4×4×256, batchnorm, ReLU convTranspose filter 4×4×128, batchnorm, ReLU convTranspose filter 4×4×64, tanh

2) Discriminator: Table II shows the architecture of the discriminator. The discriminator consists of two components: image embedding f_θ and classification f_{cls}.

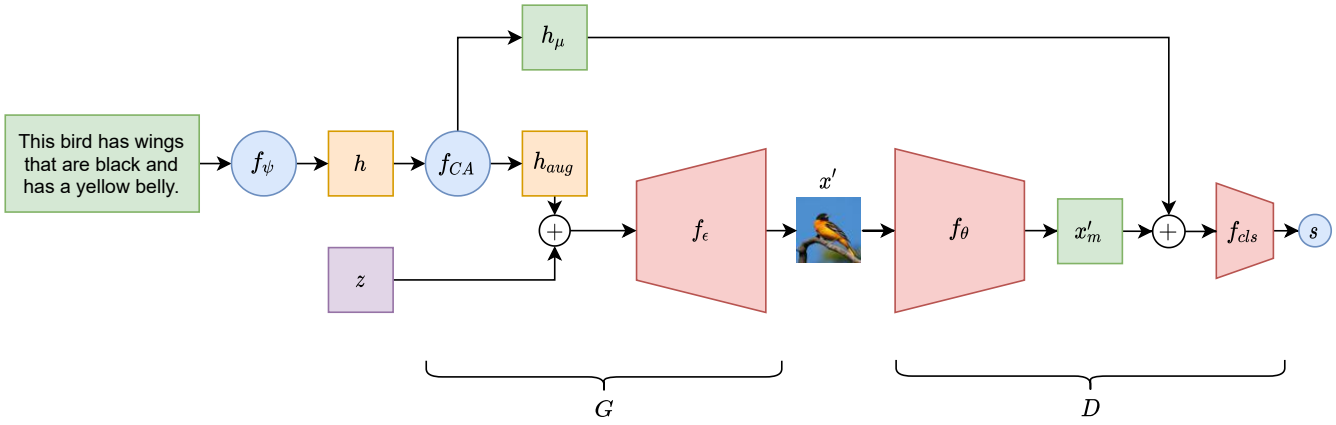

Fig. 1. The overall architecture of the proposed enhanced text-to-image synthesis cGANs. The text description is encoded through pre-trained text encoder f_ψ to obtain its embedding h. The h is augmented through f_{CA} to obtain augmented h_{aug} and h_μ. The h_{aug} is concatenated with random noise z before sending to f_ϵ to generate a fake image x'. The x' is encoded through f_θ to obtain the image embedding x'_m. After that, the x'_m is concatenated with augmented text embedding h_μ as the input for f_{cls} to return the prediction score s.

The image embedding f_θ comprises 4 convolutional layers with 3, 64, 128, and 256 filters, respectively. Each filter is of the size 4×4. Every convolutional layer is activated by the leaky ReLU function and followed by a batch normalization layer. The image embedding f_θ returns the extracted features from the input image.

The classification f_{cls} consists of one convolution layer with sigmoid function to classify whether the image is fake or real. The input to classification f_{cls} are the augmented text embeddings and the image features, while the output is the predicted class label.

TABLE II
THE CONFIGURATION AND NETWORK ARCHITECTURE OF THE DISCRIMINATOR.

Component	Configuration
Image Embedding f_θ	conv filter 4×4×3, batchnorm, leaky ReLU conv filter 4×4×64, batchnorm, leaky ReLU conv filter 4×4×128, batchnorm, leaky ReLU conv filter 4×4×256, batchnorm, leaky ReLU
Classification f_{cls}	conv filter 8×8×768, sigmoid

C. Process

This section describes the input, process and output of the generator G and discriminator D in the enhanced text-to-image synthesis cGANs.

1) Generator: The generator G in text-to-image synthesis cGANs aims to produce images that can deceive the discriminator into classifying the images as real images. The generator receives the text embedding h and a random latent variable z sampled from the Gaussian distribution $N(0, 1)$.

As the text embedding h is nonlinearly transformed in the generator, it might cause data discontinuity and negatively impact the generator's learning. In view of this, this paper proposes a text conditioning augmentation function. The text conditioning augmentation function f_{CA} addresses this problem by synthesizing more sample pairs from a given small set of input pairs. In the text conditioning augmentation function f_{CA}, the text embedding h is encoded through a linear layer with ReLU activation function f to obtain 256-dimensional augmented text embeddings h_μ and h_σ. Subsequently, the latent variable z multiplied with h_σ and added with h_μ to obtain the augmented text embedding h_{aug}. The process of the conditioning augmentation is shown in Fig. 2.

$$\begin{aligned} h_\mu, h_\sigma &= f(h) \\ h_{aug} &= z \times h_\sigma + h_\mu \end{aligned} \tag{1}$$

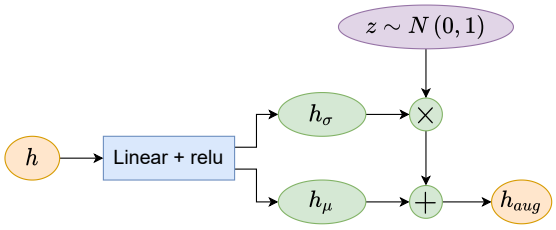

Fig. 2. The process of the text conditioning augmentation function f_{CA}.

The text conditioning augmentation incorporates an additional loss function $\mathcal{L}_{CA}$ into the generator which is determined by:

$$\mathcal{L}_{CA} = D_{KL}\left(N(h_\mu, h_\sigma) \parallel N(0, 1)\right) \tag{2}$$

where D_{KL} refers to the Kullback-Leibler divergence between the conditioned Gaussian distribution $N(h_\mu, h_\sigma)$ and standard Gaussian distribution $N(0, 1)$. With the text conditioning augmentation, the images are associated with more semantically similar text descriptions, thus improving the semantic consistency of the synthesized images. The output of the conditioning augmentation function f_{CA} are the augmented text embedding h_{aug}, h_μ, and h_σ, which serve as the input to the modeling component f_ϵ of the generator.

Given the latent variable z and augmented text embeddings h_{aug}, the modeling component f_ϵ synthesizes the images to fool the discriminator. The process of the modeling component f_ϵ is defined as:

$$x' = f_\epsilon(z, h_{aug}) \tag{3}$$

where x' represents the generated images. Given Equation 1 and 3, the overall process of the generator is defined as:

$$x', h_\mu, h_\sigma = G(z, h) \tag{4}$$

As the generator focuses on deceiving the discriminator into believing the synthesized images are true, the generator tends to synthesize similar images that had previously successfully deceived the discriminator. The scenario of synthesizing images with low varieties is known as mode collapse. In order to mitigate the mode collapse issue, we integrate an additional feature matching loss function into the generator. Instead of focusing on, The feature matching loss function shifts the generator's focus from synthesizing images to fool the discriminator to synthesizing images that resemble the real images. In doing so, the generator tends to synthesize more diverse images similar to the varieties of the real images. The feature matching loss function $\mathcal{L}_{fm}$ is defined as:

$$\mathcal{L}_{fm} = \|x_m - x'_m\|_2^2 \tag{5}$$

where x_m and x'_m denote the real and fake image features returned by the image embedding component f_θ of the discriminator. By integrating the feature matching loss function, the synthesized images are more visually realistic and semantically consistent to the text description.

In order to further reduce the dissimilarity between real and fake images, the L1 distance loss is introduced into the objective function of the generator. The L1 distance loss $\mathcal{L}_{L1}$ computes the distance between the real image x and fake image x' and is described as:

$$\mathcal{L}_{L1} = \|x - x'\|_1 \tag{6}$$

With the L1 distance loss, the generator strives to minimize the dissimilarity between the real and fake images. Thus, improving the synthesized images to be more photo-realistic and semantic consistent to the real images and associated text descriptions.

After integrating the additional loss functions introduced by text conditioning augmentation, feature matching and L1 distance, the final loss function of the generator is defined as:

$$\mathcal{L}_G = \log\left[D\left(x', h_\mu\right)\right] + \delta \cdot \mathcal{L}_{fm} + \eta \cdot \mathcal{L}_{L1} + \lambda \cdot \mathcal{L}_{CA} \tag{7}$$

where $\log\left[D\left(x', h_\mu\right)\right]$ is the fundamental loss function for the generator to defeat the discriminator. The δ, η, and λ are the coefficients for the feature matching loss $\mathcal{L}_{fm}$, the L1 distance loss $\mathcal{L}_{L1}$, and the conditioning augmentation loss $\mathcal{L}_{CA}$. Based on the loss function, the generator synthesizes image x' and passes it together with the augmented text embedding h_μ to the discriminator.

2) Discriminator: Upon receiving the real image x, generated image x', and the augmented text embedding h_μ, the discriminator D performs classification to determine whether the image is a real or fake image. The input image x' is encoded by the image embedding f_θ to obtain the image features x'_m. The process is defined as:

$$\begin{aligned} x_m &= f_\theta\left(x\right) \\ x'_m &= f_\theta\left(x'\right) \end{aligned} \tag{8}$$

The image features are then concatenated with the augmented text embedding h_μ, and serve as the input to the classifier f_{cls} in the discriminator. The classifier f_{cls} classifies the image features and augmented text embedding h_μ in pair to obtain a final decision score s. The general process in the classifier is defined as:

$$s = f_{cls}\left(X_m, h_\mu\right) \tag{9}$$

where X_m includes both real image features x_m and fake image features x'_m and s denotes the final decision score. Based on Equation 8 and 9, the overall discriminator process is shown as:

$$s = D\left(x, h_\mu\right) \tag{10}$$

In order to train the discriminator to distinguish whether the given image is real or fake and whether the image is semantically consistent with the text description, 3 types of input pairs are used:

- Real image x with matched text description t.
- Real image x with unmatched text description $t^{'}$.
- Fake image x' with any text description t.

The real and fake image pairs train the discriminator to distinguish the real image from the fake image while the matched and unmatched text description pairs train the discriminator to determine the semantic consistency. By utilizing these input pairs, the discriminator can better distinguish whether the image is real or fake as well as the relationship between the image and text description which is not available in the conventional discriminator.

With the different input pairs, the loss function of the discriminator is constructed as:

$$\mathcal{L}_r = \log\left[D\left(x, h_\mu\right)\right] \tag{11}$$

$$\mathcal{L}_f = \frac{\log\left[1 - D\left(x, h'_\mu\right)\right] + \log\left[1 - D\left(x', h_\mu\right)\right]}{2} \tag{12}$$

$$\mathcal{L}_D = \mathcal{L}_r + \mathcal{L}_f \tag{13}$$

where $\mathcal{L}_r$ denotes the computed loss from the real input pair. $\mathcal{L}_f$ represents the average total loss computed from two fake input pairs with fake images x' and unmatched text description t'. The input h'_μ is the augmented text embedding for t' using Equation 1. $\mathcal{L}_D$ denotes the final loss of the discriminator. The overall procedure of the enhanced text-to-image synthesis cGANs is presented in Algorithm 1.

IV. Experiments

This section describes the experimental details, including dataset, hyperparameter settings, evaluation metrics, and experimental results in comparison with existing methods.

A. Datasets

We have evaluated the enhanced text-to-image GANs on a widely used dataset: CUB-200-2011 [14]. The CUB-200-2011 dataset consists of 11788 images from 200 bird species. The dataset is split into 150 training classes (train+val) and 50 test classes. The splitting of the dataset is based on the zero-shot experimental settings where all the test data remain unseen during the training. In the training stage, each image is paired with 5 randomly selected captions out of 10 captions. While during the test stage, each image is paired with all captions available in the dataset. Throughout the experiments, all images are resized into 128×128 pixels and normalized into the range of $[-1, 1]$. Random horizontal flip is applied as the image augmentation technique in the training.

Algorithm 1 The training stage of our proposed enhanced text-to-image GANs.

Input: Mini batch b from the dataset T, real images x, matched text description t, unmatched text description t'.

random initialize D and G
for e iterations **do**
  for $b \sim T$ **do**
    $x, t, t' \leftarrow b$
    $h \leftarrow f_\psi(t)$
    $h' \leftarrow f_\psi(t')$
    $z \sim N(0,1)$
    $x', h_\mu, h_\sigma \leftarrow G(z, h)$
    $h'_\mu \leftarrow f_{CA}(h')$
    $\mathcal{L}_r \leftarrow \log[D(x, h_\mu)]$
    $\mathcal{L}_f \leftarrow \frac{\log\left[1-D\left(x,h'_\mu\right)\right]+\log\left[1-D\left(x',h_\mu\right)\right]}{2}$
    $\mathcal{L}_D \leftarrow \mathcal{L}_r + \mathcal{L}_f$
    $D \leftarrow D - \Delta\sigma\mathcal{L}_D/\sigma D$
    $\mathcal{L}_{fm} \leftarrow \|f_\theta(x) - f_\theta(x')\|_2^2$
    $\mathcal{L}_{L1} \leftarrow \|x - x'\|_1$
    $\mathcal{L}_{CA} \leftarrow D_{KL}(h_\mu, h_\sigma)$
    $\mathcal{L}_G \leftarrow \log[D(x', h_\mu)]$
    $\mathcal{L}_G \leftarrow \mathcal{L}_G + \delta \cdot \mathcal{L}_{fm} + \eta \cdot \mathcal{L}_{L1} + \lambda \cdot \mathcal{L}_{CA}$
    $G \leftarrow G - \Delta\sigma\mathcal{L}_G/\sigma G$
  end for
end for

B. Hyperparameter Settings

This section describes the hyperparameter settings of the enhanced text-to-image synthesis cGANs. For the loss function of the generator, δ and η are set to 1.0 while λ is set to 2.0. Both the generator and discriminator are trained using Adaptive Moment Estimation (Adam) optimizer [15] with learning rate 0.0002. The training epoch is set to 600 and the learning rate is decayed into half for every 100 epochs. The batch size is 64 and the dimension of the input noise is 100.

C. Evaluation

In this work, two evaluation metrics, i.e. Inception score [16] and Structural Similarity Index Matrix (SSIM) [17], are adopted to evaluate the performance of the enhanced text-to-image synthesis cGANs. Inception score is widely used for evaluating the quality of the generated image. The Inception score is computed as:

$$I = \exp\left(\mathbb{E}_{x'} D_{KL}\left(p\left(y \mid x'\right) \parallel p\left(y\right)\right)\right) \tag{14}$$

where x' is the image synthesized by the enhanced text-to-image synthesis cGANs. A pretrained Inception v3 model [18] is leveraged as the classification model to predict all images x' produced by the enhanced text-to-image synthesis cGANs based on all text descriptions in the test set. The Inception score evaluates the performance based on the quality and diversity of the synthesized images, the prediction accuracy, and the number of object types in the synthesized images.

SSIM is used to measure the similarity between the generated images and real images. It can also be used to measure the semantic consistency of the generated images as the real images and generated images are from the same text description. The generated images and real images with the same text description are evaluated as a pair in SSIM. The average score of all pairs are computed as the final SSIM result.

D. Ablation Study

An ablation study is performed to evaluate each component of the enhanced text-to-image synthesis cGANs. The experiments involve four parts: the baseline GAN-INT-CLS, feature matching loss $\mathcal{L}_{fm}$, L1 distance loss $\mathcal{L}_{L1}$, and text conditioning augmentation loss $\mathcal{L}_{CA}$. The results of the ablation study are presented in Table III.

TABLE III
THE ABLATION STUDY OF THE ENHANCED TEXT-TO-IMAGE SYNTHESIS CGANS ON CUB-200-2011 DATASET.

Model	IS
Baseline	4.64±0.05
Baseline + $\mathcal{L}_{L1}$	5.01±0.07
Baseline + $\mathcal{L}_{fm}$ + $\mathcal{L}_{L1}$	5.20±0.05
Baseline + $\mathcal{L}_{fm}$ + $\mathcal{L}_{L1}$ + $\mathcal{L}_{CA}$	**5.26±0.09**

It is observed that the baseline model without any enhancements obtained the Inception score of 4.64±0.05. When the L1 distance loss $\mathcal{L}_{L1}$ is added to the baseline model, the Inception score increases to 5.01±0.07. The improvement demonstrates that the $\mathcal{L}_{L1}$ is efficient in motivating the generator to synthesize images that are more similar to the real images. After incorporating the feature matching loss $\mathcal{L}_{fm}$, the Inception score increases to 5.20±0.05. The improvement is attributable to the $\mathcal{L}_{fm}$ that makes the generator synthesize more image varieties using the features similar to the real image features, which in turn mitigates the mode collapse issue and improves the training stability. The Inception score further escalates to 5.26±0.09 after integrating the text conditioning augmentation loss function $\mathcal{L}_{CA}$. The $\mathcal{L}_{CA}$ function produces more semantically similar text embeddings to improve the semantic consistency of the synthesized images.

E. Comparison with existing approaches

We have also included several existing state-of-the-art text-to-image synthesis methods in the performance evaluation. The experimental results on the CUB-200-2011 dataset are presented in Table IV. It is observed that the enhanced text-to-image GANs outshines the methods in comparison by yielding the highest Inception score of 5.26±0.09. The high Inception score corroborates the effectiveness of the proposed enhancements in improving the quality of the synthesized images, in terms of image variations, photographic realism and semantic consistency.

In addition to that, the enhanced text-to-image synthesis cGANs have recorded an outstanding SSIM of 0.7741. This is attributable to the feature matching loss and L1 distance loss that close the gap between the real and fake images in the visual feature space. Apart from that, the text conditioning augmentation also improves the semantic consistency between the synthesized images and text descriptions.

Fig. 3 presents some sample images produced by the enhanced text-to-image synthesis cGANs based on the text

TABLE IV
THE COMPARISON RESULTS ON CUB-200-2011 DATASET.

Methods	IS	SSIM
GAN-INT-CLS [3]	5.08±0.08	0.2934
GAWWN [9]	5.22±0.08	0.2370
StackGAN [10]	4.95±0.04	0.2812
AttnGAN [13]	4.43±0.06	0.3129
Enhanced Text-to-Image Synthesis GANs	**5.26±0.09**	**0.7741**

descriptions from CUB-200-2011 dataset. The samples show that the synthesized images are semantically consistent with the text description. This is ascribable to the L1 distance and feature matching that simulate the content of the generated images based on the real images to improve the diversity in the learned representation. Moreover, the text conditioning augmentation loss provides better conditioning in the text embedding space and enhances the semantic consistency of the synthesized images.

Fig. 4 shows some sample images generated by the enhanced text-to-image synthesis cGANs conditioned on the unseen text captions from the CUB-200-2011 dataset. The images exhibit diversifying and rich image contents from the feature matching function. The feature matching provides visual realistic features that are close to the real images to enrich the learned visual representation and stabilize the model training.

V. CONCLUSION

In this paper, we propose several enhancements to the text-to-image synthesis cGANs. Firstly, the text conditioning augmentation loss enriches the text embedding representation to diversify the visual features in the generator. It also standardizes the text feature information in both the discriminator and the generator to improve the semantic consistency of the synthesized images. Secondly, the feature matching loss gears the generator to generate visual contents closer to the real images thus diversifying the synthesized images and alleviates the mode collapse issue. Thirdly, the L1 distance loss helps ensure the synthesized visual contents to resemble the real images. Hence, the synthesized images are more photo-realistic and semantically consistent to the text description.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[3] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1060–1069.

[4] J. Liu and W. Wu, "Automatic image annotation using improved wasserstein generative adversarial networks." *IAENG International Journal of Computer Science*, vol. 48, no. 3, pp. 507–513, 2021.

[5] R. Yang, X.-Q. Chen, and T.-J. Cao, "APE-GAN++: An improved APE-GAN to eliminate adversarial perturbations." *IAENG International Journal of Computer Science*, vol. 48, no. 3, pp. 827–844, 2021.

[6] M. Wang and M. Wang, "Study on parameter correction of spring particle model based on generative adversarial network." *Engineering Letters*, vol. 29, no. 4, pp. 1494–1501, 2021.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[8] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[9] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Advances in Neural Information Processing Systems*, 2016, pp. 217–225.

[10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5907–5915.

[11] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[12] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Z. Afzal, and M. Liwicki, "TAC-GAN-Text conditioned auxiliary classifier generative adversarial network," *arXiv preprint arXiv:1703.06412*, 2017.

[13] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep., 2011.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *Advances in Neural Information Processing Systems*, pp. 2226–2234.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

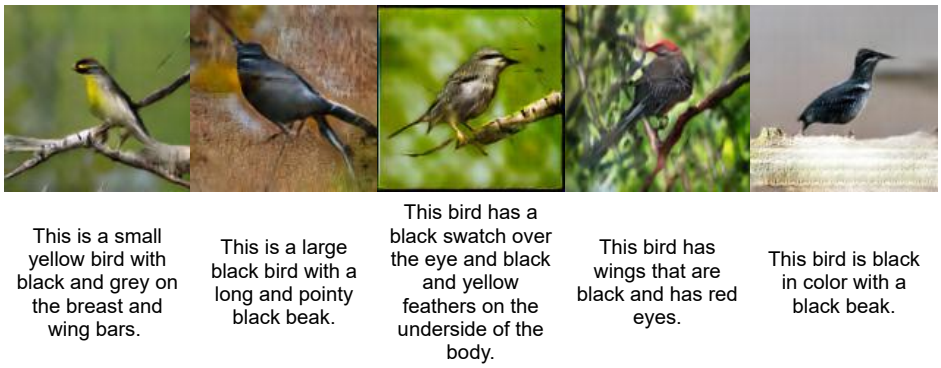

Fig. 3. Some sample bird images generated by the enhanced text-to-image synthesis cGANs. The images highly match with their corresponding text descriptions. All text descriptions are taken from the unseen classes (zero-shot experiments).

Fig. 4. Some sample images generated from the unseen classes by the enhanced text-to-image synthesis cGANs. The images demonstrate the ability to synthesize diverse image contents.