iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59397

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Enhancing Image Realism Through FineGrained Text to Image Synthesis

Priyavarshini. S[1], Sriranganayaki. S[2], Steffy. D[3]
Department Of Artificial Intelligence and Data Science, Meenakshi Sundararajan Engineering College, Chennai -21

Abstract: ***We propose a more effective Deep Fusion Generative Adversarial Networks (DF- GAN) for synthesizing high-quality realistic images from text descriptions. The main challenges in this task are the entanglements between generators of different image scales, the reliance on extra networks for text-image semantic consistency, and the computational cost of cross-modal attention-based fusion. Ourproposed approach addresses these challenges as follows:***

1) ***We introduce a novel one-stage text-to- image backbone that directly synthesizes high-resolution images without entanglements between different generators. This simplifies the architectureand improves efficiency.***
2) ***We propose a novel Target-Aware Discriminator composed of Matching- Aware Gradient Penalty and One-Way Output. This discriminator enhances text- image semantic consistency without introducing extra networks, allowing for better supervision capability.***
3) ***We introduce a novel deep text-image fusion block, which deepens the fusion process to make a full fusion between textand visual features. This enhances the quality of synthesized images.***

To evaluate the performance of our proposed DF- GAN, we compare it with current state-of-the-art methods on widely used datasets. Ourexperimental results show that our proposedapproach is simpler but more efficient in synthesizing realistic and text-matching images and achieves better performance.

I. INTRODUCTION

We propose a novel text-to-image generation method named Deep Fusion Generative Adversarial Network (DF-GAN). DF-GAN generates high-quality images directly and fuses the text and image features deeply by our deep text-image fusion blocks. This approach results in realistic and text-consistent images from given natural language descriptions. Text-to-image synthesis has become an active research area due to its practical value.

However, two major challenges remain: the authenticity of thegenerated image and the semantic consistency between the given text and the generated image.

To address these challenges, most recent models adopt a stacked architecture to generate high- resolution images. These models employ cross-modal attention to fuse text and image features and introduce extra networks, such as DAMSM, cycle consistency, or Siamese networks, to ensure text-image semantic consistency.

Despite impressive results from previous works, there are still three problems. First, the stacked architecture introduces entanglements between different generators, resulting in final refined images that look like a simple combination of fuzzy shapes and some details.Second, existing studies usually fix the extra networks during adversarial training, making themeasily fooled by the generator and weakening theirsupervision power on semantic consistency.

Third, cross-modal attention can only be applied two times on 64×64 and 128×128 image features due to its high computational cost, limiting itseffectiveness in the text-image fusion process and making it difficult to extend to higher-resolution image synthesis.

We propose a novel one-stage text-to-image backbone that can synthesize high-resolution images directly without entanglements between different generators. Our proposed backbone iscomposed of hinge loss and residual networks, which stabilize the GAN training process. We alsodesign a Target-Aware Discriminator composedof Matching-Aware Gradient Penalty (MA-GP)and One-Way Output to enhance the text-image semantic consistency. Finally, we propose a Deep text-image Fusion Block (DFBlock) to fuse the text information into image features more effectively.

Through experimental results, we demonstrate the effectiveness of our proposed method in synthesizing high-resolution images with high- quality text-matching features. Our methodoutperforms previous state-of-the-art methods in terms of synthesis quality and semantic consistency.

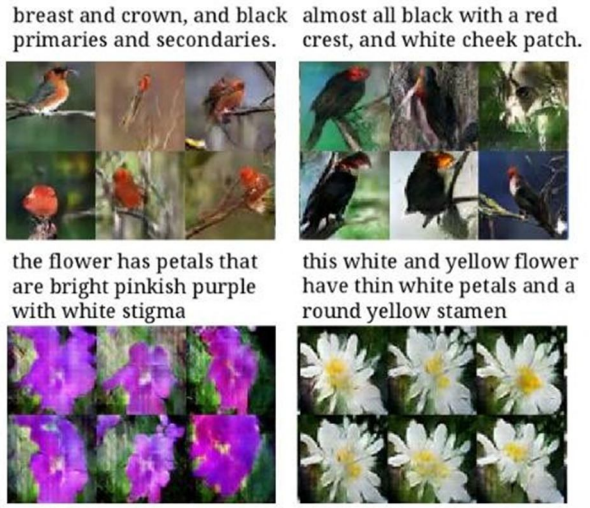


Figure 1. Examples of generated images from text descriptions

The main contributions of our paper can besummarized as follows:

1) We propose a novel one-stage text-to- image backbone that can synthesize high- resolution images directly without entanglements between different generators.
2) We design a Target-Aware Discriminator composed of Matching-Aware Gradient Penalty (MA-GP) and One-Way Output to enhance the text-image semantic consistency.
3) We propose a Deep text-image Fusion Block (DFBlock) to fuse the text information into image features more effectively.
4) We address the issue of entanglementsbetween different generators in GANs by replacing the stacked backbone with a one-stage backbone. This approach stabilizes the GAN training process and allows the generator to synthesize high- resolution images directly.
5) We provide experimental results todemonstrate the effectiveness of our proposed method in synthesizing high- resolution images with high-quality text- matching features. These results show that our method outperforms previous state-of-the-art methods in terms of synthesis quality and semanticconsistency.
6) We conclude by highlighting the potential applications of our proposed DF-GAN method in various image synthesis tasks, such as image editing, style transfer, and image generation from natural language descriptions.

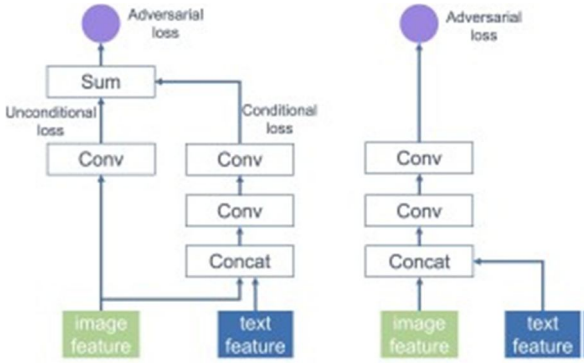
We propose a Deep text-image Fusion Block (DF Block) to fuse the text information into imagefeatures more effectively. The DF Block consistsof several Affine Transformations. These lightweight modules manipulate the visual feature maps through channel-wise scaling and shifting operations. Stacking multiple DF Blocks atall image scales deepens the text-image fusion process and makes a full fusion between text and visual features.

A. The Proposed DF GAN

DF-GAN generates high-resolution imagesdirectly by one pair of generator and discriminatorand fuses the text information and visual feature maps through multiple Deep text-image Fusion Blocks (DF Block) in UP Blocks. Armed with Matching-Aware Gradient Penalty (MA-GP) and One-Way Output, our model can synthesize more realistic and text-matching images.

The main contributions of our paper can be summarized as follows:

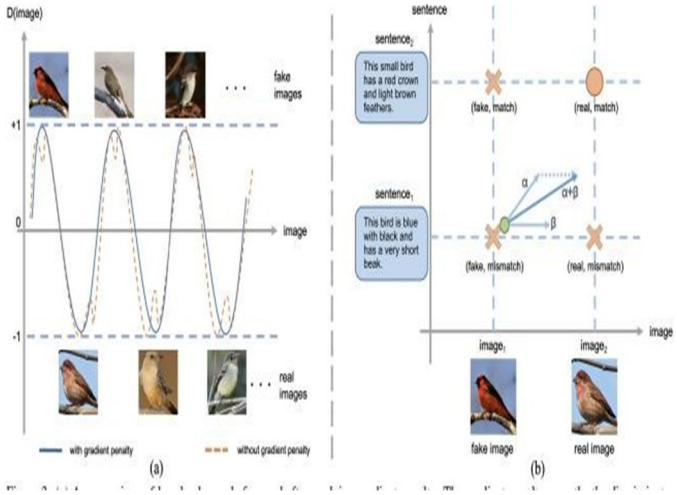
1) We propose a novel one-stage text-to- image backbone that can synthesize high- resolution images directly without visual feature entanglements.
2) We design a Target-Aware Discriminator composed of Matching-Aware Gradient Penalty (MA-GP) and One-Way Output to enhance the text-image semantic consistency without introducing extra networks.



3) We propose a novel Deep text-image Fusion Block (DF Block) to more fullyfuse text and visual features.

The generator consists of multiple UP Blocks, each of which contains several Residual-in- Residual Dense Blocks (RRDB) [18] and a DF Block. The DF Block fuses the text information and visual feature maps through channel-wise attention and spatial attention mechanisms.

The discriminator consists of several Down Blocks, each of which contains several convolutional layers and a leaky ReLU activation function. The discriminator also includes a MA- GP and One-Way Output to enhance the text- image semantic consistency.



We evaluate our proposed DF-GAN model on theCUB dataset [30] and compare it with several state-of-the-art text-to-image synthesis models.

The experimental results show that our model cansynthesize more realistic and text-matching images than the compared models.

The proposed DF-GAN model consists of a generator and a discriminator. The generator takesa random noise vector and a text embedding as inputs and generates an image. The discriminator takes an image and a text embedding as inputs and outputs a probability that the image is real or fake given the text.

The generator consists of multiple UP Blocks, each of which contains several Residual-in- Residual Dense Blocks (RRDB) [18] and a DF Block. The DF Block fuses the text information and visual feature maps through channel-wise attention and spatial attention mechanisms.

The discriminator consists of several Down Blocks, each of which contains several convolutional layers and a leaky ReLU activation function. The discriminator also includes a MA- GP and One-Way Output to enhance the text- image semantic consistency.

The architecture of the proposed DF-GAN for text-to-image synthesis is shown in Figure 2. DF- GAN generates high-resolution images directly byone pair of generator and discriminator and fuses the text information and visual feature maps through multiple

Deep text-image Fusion Blocks (DF Block) in UP Blocks. Armed with Matching- Aware Gradient Penalty (MA-GP) and One-Way Output, our model can synthesize more realistic and text-matching images.

We evaluate our proposed DF-GAN model on theCUB dataset [30] and compare it with several state-of-the-art text-to-image synthesis models.

The experimental results show that our cansynthesize more realistic and text-matchingimages than the compared models.

This pink flower has overlapping petals and yellow florets growing in the middle.

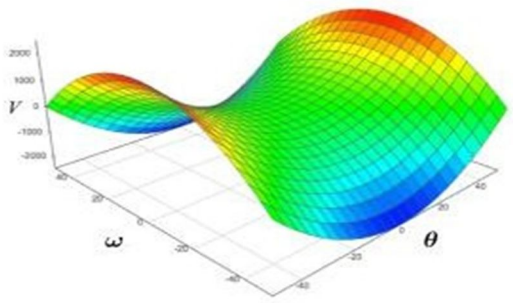
This bird is white with blue on its back and has a long, pointy beak.

In summary, our paper presents a novel text-to- image generation method that addresses the limitations of existing studies. By replacing the stacked backbone with a one-stage backbone, we avoid entanglements between different generators.We enhance the text-image semantic consistency by employing a Target-Aware Discriminator composed of Matching-Aware Gradient Penalty(MA-GP) and One-Way Output. Finally, wepropose a Deep text-image Fusion Block (DF Block) to fuse the text information into image features more effectively. These contributions make our proposed method a promising approach for future research in text-to-image generation.

II. OVERVIEW

Deep Fusion Generative Adversarial Network (DF-GAN) for text-to-image synthesis. Our model is composed of three main components: a generator, a discriminator, and a pre-trained text encoder

The generator takes two inputs, a sentence vector encoded by the text encoder and a noise vector sampled from a Gaussian distribution, to ensure the diversity of the generated images. The noise vector is first fed into a fully connected layer and reshaped. Then, a series of UP-Blocks up sample the image features, where each UP-Block is composed of an up sample layer, a residual block, and DF-Blocks to fuse the text and image features during the image generation process. Finally, aconvolution layer converts image features into images. The discriminator converts images into image features using a series of Down Blocks. Then, the sentence vector is replicated and concatenated with the image features. Anadversarial loss is predicted to evaluate the visual realism and semantic consistency of the inputs. By distinguishing generated images from real samples, the discriminator promotes the generator to synthesize images with higher quality and text- image semantic consistency. The text encoder is a bi-directional Long Short-Term Memory (LSTM) that extracts semantic vectors from the textdescription. We use the pre-trained model provided by AttnGAN.



Our DF-GAN model differs from existingapproaches in the following ways:

1) We propose a novel one-stage text-to- image backbone that generates high- resolution images directly without visual feature entanglements.

2) We introduce a Target-Aware Discriminator with Matching-Aware Gradient Penalty (MA-GP) and One-Way Output to enhance text-image semantic consistency without introducing extranetworks.

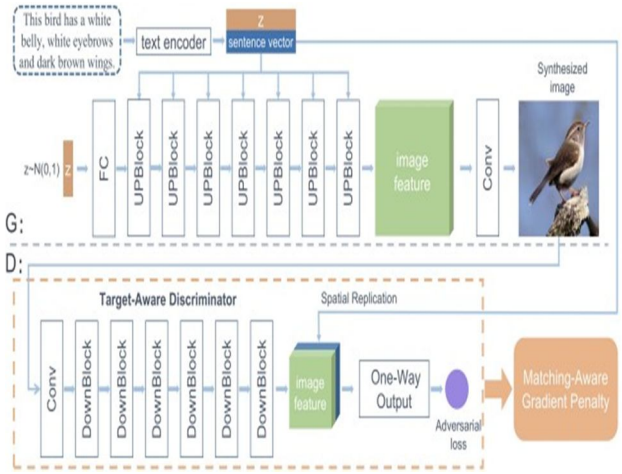
3) We propose a novel Deep text-image Fusion Block (DF Block) to more fullyfuse text and visual features.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

We evaluate our proposed DF-GAN model on the CUB dataset and compare it with several state-of- the-art text-to-image synthesis models. The experimental results show that our model can synthesize more realistic and text-matching images than the compared models.

A. Text-to-Image

Our proposed one-stage text-to-image backbone for generating high-resolution images directly from text descriptions. Previous text-to-image GANs usually employ stacked architecture to generate high-resolution images from low-resolution ones. However, the stacked architecture introduces entanglements between different generators, resulting in final refined images that look like a simple combination of fuzzy shapes and some details.



To address this issue, we propose a novel one- stage text-to-image backbone that can synthesize high-resolution images directly by a single pair of generator and discriminator.

We adopt hinge loss to stabilize the adversarial training process. Since there is only one generator in the one-stage backbone, it avoids the entanglements between different generators. As the single generator in our one-stage framework needs to synthesize high-resolution images from noise vectors directly, it must contain more layers than previous generators in stacked architecture. To train these layers effectively, we introduce residual networks to stabilize the training of deeper networks. The formulation of our one- stage method with hinge loss is as follows:

$$L\_D = - E\{x \sim P\ r\} [\min(0, -1 + D(x, e))] - (1/2)E\_\{G(z)P\ g\} [\min(0, -1 - D(G(z), e))] -(1/2)E\_\{ xP\ mis\} [\min(0, -1 - D(x, e))]$$

$$L\_G = - E\_\{G(z)\sim P\_g\} [D(G(z), e)]$$

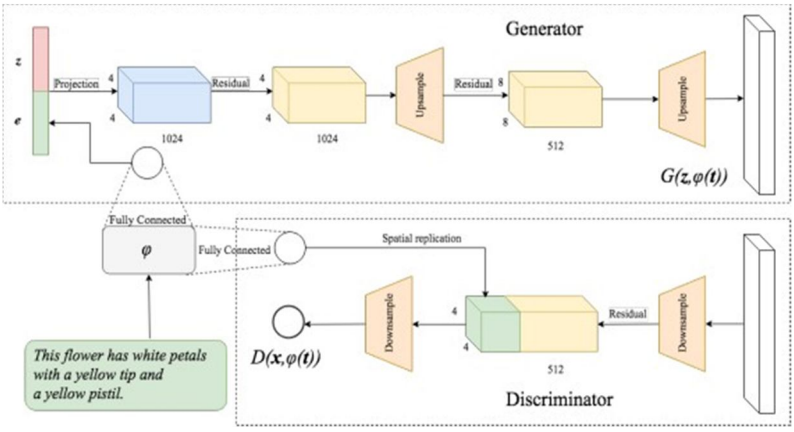
where z is the noise vector sampled fromGaussian distribution; e is the sentence vector;P_g, P_r, P_{mis} denote the synthetic datadistribution, real data distribution, and mismatching data distribution, respectively.

By using a one-stage backbone, we can avoid the entanglements between different generators andgenerate high-resolution images directly from textdescriptions. However, this approach introduces a new challenge: training a single generator to synthesize high-resolution images from noisevectors directly is more difficult than training multiple generators in a stacked architecture. To address this challenge, we introduce residual networks to stabilize the training of deepernetworks. We evaluate our proposed one-stage text-to-image backbone on the CUB dataset and compare it with several state-of-the-art text-to- image synthesis models. Our experimental results show that our model can synthesize more realistic and text-matching images than the compared models.

In summary, we propose a novel one-stage text- to-image backbone for generating high-resolutionimages directly from text descriptions. By using hinge loss and residual networks, we can train the generator effectively and avoid the entanglements between different generators. Our experimental results demonstrate the effectiveness of our proposed method.

III. RELAVENT WORK

1) Generative Adversarial Networks (GANs) [8]: GANs are a class of generative models that can learn complex data distributionsby training a generator and a discriminatorin a two-player min-max game.



2) Conditional GANs [37, 38]: Conditional GANs are a variant of GANs that can generate images based on input conditions,such as text descriptions.
3) Stack GAN [56, 57]: Stack GAN is a multi-stage generative model that can generate high-resolution images by stacking multiple generators and discriminators.
4) AttnGAN [50]: AttnGAN is a text-to- image generation model that introduces a cross-modal attention mechanism to help the generator synthesize images with moredetails.
5) Mirror GAN [33]: Mirror GAN is a text- to-image generation model that regenerates text descriptions from generated images for text-image semantic consistency.
6) SD-GAN [51]: SD-GAN is a text-to-imagegeneration model that employs a Siamese structure to distill the semantic commons from texts for image generation consistency.
7) DM-GAN [60]: DM-GAN is a text-to- image generation model that introduces a Memory Network to refine fuzzy image contents when the initial images are not well generated in stacked architecture.
8) Transformer-based text-to-image methods [7, 24, 35]: These methods tokenize theimages and take the image tokens and word tokens to make auto-regressive training by a unidirectional Transformer.

Our DF-GAN is different from these previous methods as it generates high-resolution images directly by a one-stage backbone, adopts a Target- Aware Discriminator to enhance text-image semantic consistency without introducing extra networks, and fuses text and image features moredeeply and effectively through a sequence of DF Blocks. Compared with previous models, our DF-GAN is much simpler but more effective in synthesizing realistic and text-matching images.

IV. EXPERIMENTS

DF-GAN has a one-stage text-to-image backbone, which can synthesize high-resolution images directly without entanglements between different generators. It also has a Target-AwareDiscriminator, which concatenates the image feature and sentence vector to output only oneadversarial loss through two convolution layers, making the gradient pointed to the target data points directly. This optimizes and accelerates the convergence of the generator.

DF-GAN also has a Deep text-image Fusion Block (DF Block), which stacks multiple Affine Transformations and ReLU layers in Fusion Blockto deepen the text-image fusion process. This makes the generator more fully exploit the text information when fusing text and image features and enlarges the representation space of the fusion module, which is beneficial to generate semantic consistent images from different text descriptions. The paper evaluates DF-GAN and its variants quantitatively and qualitatively on two challenging datasets, CUB bird and COCO. The evaluation metrics used are Inception Score (IS), Frechet Inception Distance (FID), and the number of parameters (NoP). The paper compares DF- GAN with several state-of-the-art methods, including StackGAN, StackGAN++, AttnGAN, MirrorGAN, SD-GAN, and DM-GAN, and recent models such as CPGAN, XMC-GAN, DAE- GAN, and TIME.

The results show that DF-GAN has a significant smaller Number of Parameters (NoP) but still achieves a competitive performance comparedwith other leading models. It improves the ISmetric

and decreases the FID metric on both CUB and COCO datasets. The visualization results also show that images synthesized by DF-GAN have better object shapes and realistic fine-grained details than AttnGAN and DM-GAN. The ablation studies conducted on the testing set of theCUB dataset verify the effectiveness of each component in the proposed DF-GAN. The results demonstrate that the proposed One-Stage text-to- image Backbone, Matching-Aware Gradient Penalty, One-Way Output, and Deep text-image Fusion Block improve the performance of the model**.** image Backbone, Matching-Aware Gradient Penalty, One-Way Output, and Deep text-image Fusion Block improve the performanceof the model. AttnGAN and DM-GAN The ablation studies conducted on the testing set of theCUB dataset verify the effectiveness of each component in the proposed DF-GAN. The results demonstrate that the proposed One-Stage text-to- image Backbone, Matching-Aware Gradient Penalty, One-Way Output, and Deep text-image Fusion Block improve the performance of the model.

V. CONCLUSION AND FUTURE WORK

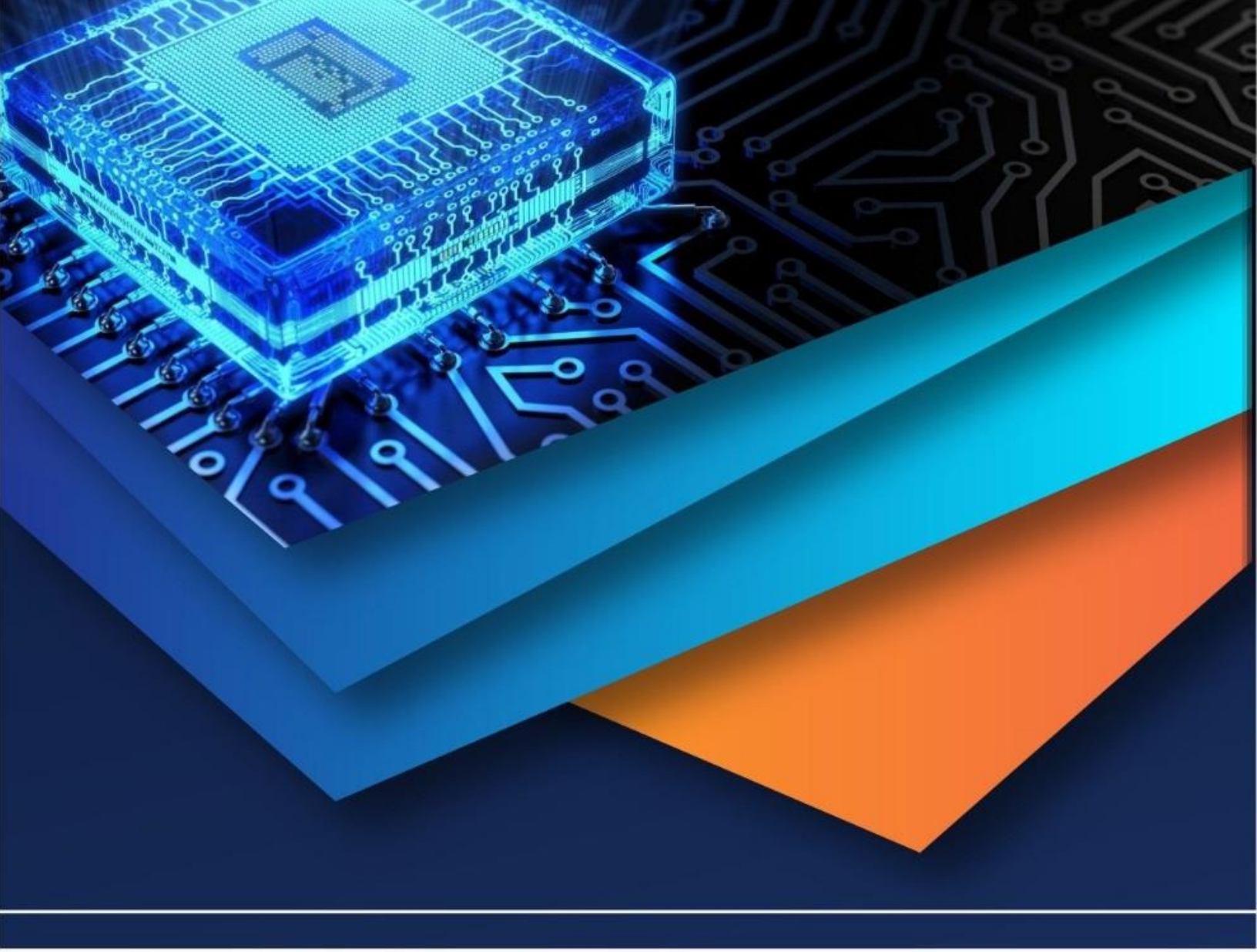
In this paper, we proposed a novel DF-GAN for text-to-image generation tasks, which includes a one-stage text-to-image backbone, a Target- Aware Discriminator, and a Deep text-image Fusion Block (DF Block). Our model can synthesize high-resolution images directly with enhanced text-image semantic consistency. Extensive experiments on the CUB and COCO datasets demonstrate that our proposed DF-GAN outperforms current state-of-the-art models.

A. Future Work

While our proposed DF-GAN has shown superior performance in text-to-image generation tasks, there are still some limitations that can be addressed in future work. First, our model only introduces sentence-level textinformation, which limits the ability of fine-grained visual feature synthesis. Second, incorporating pre-trained large language models to provide additional knowledge may further improve the performance. In addition, exploring the application of our proposed DF-GAN in other related tasks, such as video generation from text,is also an interesting direction for future research.

REFERENCES

[1] Reed, S., Akata, Z., Yan, X., Lo, J., Schiele, B.,& Lee, H. (2016). Generative adversarial textto image synthesis. In Proceedings of the International Conference on Machine Learning (ICML).
[2] Zhang, H., Xu, T., Li, H., Wang, X., Huang, X., & He, X. (2017). StackGAN: Text to photo- realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on ComputerVision (ICCV).
[3] Xu, T., Zhang, P., Huang, 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 PatternRecognition (CVPR).
[4] Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[5] Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., & Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. In Proceedings ofthe IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[6] Zhang, H., Li, H., Wang, X., Wang, X., Huang, X., & He, X. (2017). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 1947-1962.
[7] Liang, J., Pei, W., & Lu, F. (2020). Cpgan:Content-parsing generative adversarial networksfor text-to-image synthesis. In Proceedings of theEuropean Conference on Computer Vision (ECCV).
[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... &Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing System (NIPS).
[9] Devlin, J., Chang, M. -W., Lee, K., &Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).
[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Askell, A. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
[11] Cheng, W. -H., Song, S., Chen, C. -Y., Hidayati, S. C., & Liu, J. (2021). Fashionmeets computer vision: A survey. ACM Computing Surveys(CSUR), 54(4), 1-41.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY