ISSN: 0369-8963

Text to Image Translation using GAN with NLP and Computer Vision

R. PERUMALRAJA, A.S. ARJUNKUMAR AND S. KAMALESH

Velammal College of Engineering and Technology,
Madurai, Tamil Nadu, India.

ABSTRACT

Generating high-quality images from text queries is a challenging problem in computer vision and has many practical applications. This paper proposes Stacked Generative Adversarial Networks (StackGAN) to generate 256 x 256 photo-realistic images conditioned on text descriptions. We resolve the hard problem into more manageable sub-problems through a sketch-refinement process. The Stage-I GAN gives the primitive shape and colors of the object based on the given text description, yielding Stage-I low-resolution images. The Stage-II GAN uses Stage-I results and text descriptions as inputs and generates high-resolution images with photo-realistic details. It can correct defects in Stage-I results and add compelling details to the refinement process. To improve the generated images' variety and regulate the conditional-GAN training, we introduce a novel Conditioning Augmentation technique. Various experiments and comparisons with state-of-the-art benchmark datasets demonstrate that the proposed method achieves significant improvements in generating photo-realistic images conditioned on text queries.

Keywords: *Generative Adversarial Networks, Image, Generation, Text, Translation, Deep CNN layers.*

INTRODUCTION

Generative Adversarial Networks are a powerful class of neural networks that are used for unsupervised learning. Ian J. Goodfellow developed and introduced the GAN in the year of 2014. GANs are naturally made up of a system of two competing neural network models which compete with each other and can examine, capture and copy the variations within a particular dataset.It has been noticed most mainstream neural nets can be easily fooled into misclassifying things by adding only a small amount of noise into the original data.Unexpectedly, the model after adding noise has higher confidence in the wrong prediction than when it was predicted correctly. The reason for such a dispute is that most machine learning models learn from a limited amount of data, which is a huge drawback, as it is prone to overfitting. Also, the mapping between the input and the output is almost linear. Even though it may seem that the boundaries of separation between the various classes are linear, in reality, they are composed of linearities and even a small change in a point in the feature space might lead to misclassification of data.

GANs are a clever way of training a generative model by framing the problem as a supervised learning problem with two sub-models: the generator model that we train to generate new examples, and the discriminator model that tries to classify examples as either real (from the domain) or fake (generated). The two models are trained together in a zero-sum game, adversarial, until the discriminator model is fooled about half the time, meaning the generator model is generating plausible examples.GANs are an exciting and rapidly changing field, delivering on the promise of generative models in their ability to generate realistic examples across a range of problem domains, most notably in image-to-image translation

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

tasks such as translating photos of summer to winter or day to night, and in generating photorealistic photos of objects, scenes, and people that even humans cannot tell are fake.

The following sections show the Literature Review we have done for Text to Image Translation using GAN with NLP and Computer Vision, the Proposed System for Text to Image Translation using GAN with NLP and Computer Vision, then the Results and Discussion of Text to Image Translation using GAN with NLP and Computer Vision, and finally the Conclusion of the Text to Image Translation using GAN with NLP and Computer Vision.

LITERATURE REVIEW

The common approach to generating a fake image of people, pets, cartoons, or objects is with the Generative Adversarial Network (GAN) only. This GAN was first introduced in 2014 by Ian Goodfellow. GAN also works on image-to-image translation. GAN consists of two models that compete with each other. The two models are generative model G and a discriminative model D. The former entraps the data distribution, while the latter estimates the probability of sample data. The results obtained by the GAN in image generation, image transfer, and some other related tasks are quite impressive. Many types of GAN have emerged in recent years, such as DCGAN, StyleGAN, StyleGAN2, BigGAN, ProGAN, StarGAN, CycleGAN, GauGAN, etc.

Ming Tao et al. [1] proposed an approach for synthesizing high-quality realistic images from text descriptions. Existing text-to-image generative adversarial networks generally employ a stacked architecture as the backbone, yet there are still three flaws. The drawback of this approach is that it only introduces sentence-level text information, which limits the ability of fine-grained visual feature synthesis. Minfeng Zhu et al. [2] proposed the Dynamic Memory Generative Adversarial Network (DM-GAN) to generate high-quality images. The proposed method introduces a dynamic memory module to refine fuzzy image content when the initial images are not well generated. A memory writing gate is designed to select the important text information based on the initial image content, which enables our method to accurately generate images from the text description. The drawback of this approach is that the final results still rely heavily on the layout of multi-subjects in the initial images. Tingting Qiao et al. [3] proposed a novel global-local attentive and semantic-preserving text-to-image-to-text framework called MirrorGAN. MirrorGAN exploits the idea of learning text-to-image generation by redescription and consists of three modules: a semantic text embedding module (STEM), a global-local collaborative attentive module for cascaded image generation (GLAM), and a semantic text regeneration and alignment module (STREAM). The drawback of this approach is that STREAM and other MirrorGAN modules are not jointly optimized with complete end-to-end training due to limited computational resources.

Tao Xu et al. [5] proposed an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for a fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details in different subregions of the image by paying attention to the relevant words in the natural language description. The drawback is that the existing methods struggle to generate realistic high-resolution images of this dataset. Kai Hu et al. [6] proposed a novel Semantic-Spatial Aware Convolution Network that learns semantic-adaptive transformation conditioned on text to effectively fuse text features and image features and learns a mask map

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

in a weakly-supervised way that depends on the current text-image fusion process to guide the transformation spatially. The drawback is that the proposed method will generate complex images by modifying the text concerning objects. Han Zhang et al. [7] proposed a Stacked Generative Adversarial Network (StackGAN) to generate 256x256 photo-realistic images conditioned on text descriptions by decomposing the hard problem into more manageable sub-problems through a sketch-refinement process. The drawback is that conditioning augmentation is mainly needed for improving the diversity of the generated samples because of its ability to encourage robustness to small perturbations along the latent manifold.

Priyanka Mishra et al. [8] proposed an approach for synthesizing high-quality realistic images from text descriptions using Residual Generative Adversarial Networks. The drawback is that there is an absence of a proper standard evaluation metric for generative models. These models are judged by the quality of images generated by them, which is done under human supervision. Zixu Wang et al. [9] proposed a semantics-enhanced attention module and a semantics-enhanced batch normalization module. These modules improve the consistency of synthesized images by involving precise semantic features. The drawback is that sometimes balancing semantic consistency and individual diversity is a tedious process. Jezia Zakraoui et al. [10] proposed a new approach that decomposes the task of story visualization into three phases: semantic text understanding, object layout prediction, and image generation and refinement. The drawback is that, in some cases, the object shape, the location, and the size attributes sampled are still not adequate for a realistic image and need further improvement. Satya Krishna Gorti et al. [11] proposed a captioning network to caption generated images and exploit the distance between ground truth captions and generated captions to improve the network further. The drawback is that we need to manually find the ratio of the number of colors that are present in each image to all the colors that were mentioned in the caption.

PROPOSED SYSTEM

The framework we used in this paper is StackGAN. Stacked Generative Adversarial Networks to generate high-resolution images with photo-realistic details, we propose a simple yet effective Stacked Generative Adversarial Networks. It decomposes the text-to-image generative process into two stages. Stage-I GAN: it sketches the primitive shape and basic colors of the object conditioned on the given text description, and draws the background layout from a random noise vector, yielding a low-resolution image.Stage-II GAN: it corrects defects in the low-resolution image from Stage-I and completes details of the object by reading the text description again, producing a high-resolution photo-realistic image.

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

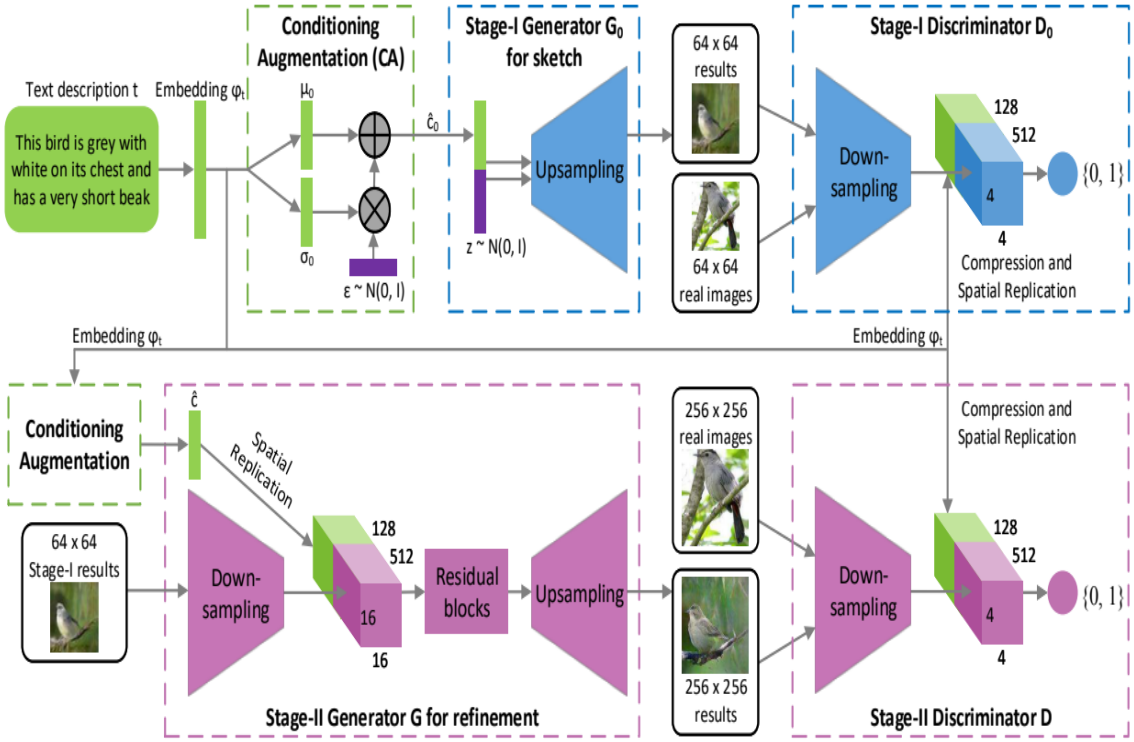

Figure 1. Stacked GAN

Generative Adversarial Networks (GAN) are composed of two models that are alternatively trained to compete with each other. The generator G is optimized to reproduce the true data distribution pdata by generating images that are difficult for the discriminator D to differentiate from real images. Meanwhile, D is optimized to distinguish real images and synthetic images generated by G. Overall, the training procedure is similar to a two-player min-max game with the following objective function

$$\min \max V(D,G) = Ex{\sim}pdata\ [logD(x)] + Ez{\sim}\text{pz}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

where x is a real image from the true data distribution pdata, and z is a noise vector sampled from distribution pz (e.g., uniform or Gaussian distribution). Conditional GAN is an extension of GAN where both the generator and discriminator receive additional conditioning variables c, yielding G(z, c) and D(x, c). This formulation allows G to generate images conditioned on variables c.

Stage-I GAN

Instead of directly generating a high-resolution image conditioned on the text description, we simplify the task to first generate a low-resolution image with our Stage-I GAN, which focuses on drawing the only rough shape and correct colors for the object. Let 't be the text embedding of the given description, which is generated by a pre-trained encoder. The Gaussian conditioning variables cˆ0 for text embedding are sampled from N (μ0('t),

https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

Σ0('t)) to capture the meaning of 't with variations. Conditioned on cˆ0 and random variable z, Stage-I GAN trains the discriminator D0 and the generator G0 by alternatively maximizing LD0 in Eq. and minimizing LG0.

$$L_{D_0} = E_{(I_U,t)\sim p_{data}}[logD_0(I_0,\varphi_t)] + E_{z\sim p_{data}}[\log(1 - D_0(G_0(z,\hat{c_0}),\varphi_t))]$$

$$L_{G_0} = E_{z\sim p_z,t\sim p_{data}}\left[\log\left(1 - D_0\left(G_0\left(z,\hat{c_0}\right),\varphi_t\right)\right)\right] + \lambda D_{KL}(N(\mu_0(\varphi_t),\Sigma_0(\varphi_t))||N(0,I))$$

where the real image I0 and the text description t are from the true data distribution pdata. z is a noise vector randomly sampled from a given distribution pz (Gaussian distribution in this paper). λ is a regularization parameter that balances the two terms in Eq. (4). We set λ = 1 for all our experiments. Using the reparameterization trick introduced in [13], both μ0('t) and Σ0('t) are learned jointly with the rest of the network.

Stage-II GAN

GAN Low-resolution images generated by Stage-I GAN usually lack vivid object parts and might contain shape distortions. Some details in the text might also be omitted in the first stage, which is vital for generating photo-realistic images. Our Stage-II GAN is built upon Stage-I GAN results to generate high-resolution images. It is conditioned on low-resolution images and also the text embedding again to correct defects in Stage-I results. The Stage-II GAN completes previously ignored text information to generate more photo-realistic details Conditioning on the low- resolution result s0 = G0(z, cˆ0) and Gaussian latent variables cˆ, the discriminator D and generator G in Stage-II GAN are trained by alternatively maximizing LD in Eq. (5) and minimizing LG in Eq.

$$L_D = E_{(I,t)\sim p_{data}}[logD(I,\varphi_t)] + E_{S_0\sim p_{G_0},t\sim p_{data}}[\log(1 - D(G(S_0,c\hat{}),\varphi_t))]$$

$$L_G = E_{S_0\sim p_0,t\sim p_{data}}\left[\log\left(1 - D(G(S_0,c\hat{}),\varphi_t)\right)\right] + \lambda D_{KL}(N(\mu(\varphi_t),\Sigma(\varphi_t))||N(0,I))$$

Different from the original GAN formulation, the random noise z is not used in this stage with the assumption that the randomness has already been preserved by s0. Gaussian conditioning variables cˆ used in this stage and cˆ0 used in Stage-I GAN share the same pre-trained text encoder, generating the same text embedding. However, Stage-I and Stage-II Conditioning Augmentation have different fully connected layers for generating different

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

means and standard deviations. In this way, Stage-II GAN learns to capture useful information in the text embedding that is omitted by Stage-I GAN.

We design the Stage-II generator as an encoder-decoder network with residual blocks. Similar to the previous stage, the text embedding 't is used to generate the Ng dimensional text conditioning vector cˆ, which is spatially replicated to form a Mg x Mg x Ng tensor. Meanwhile, the Stage-I result s0 generated by Stage-I GAN is fed into several down-sampling blocks (i.e., encoder) until it has a spatial size of Mg x Mg. The image features and the text features are concatenated along the channeldimension. The encoded image features coupled with text features are fed into several residual blocks, which are designed to learn multi-modal representations across image and text features. Finally, a series of up-sampling layers (i.e., decoder) is used to generate aW x H high-resolution image. Such a generator can help rectify defects in the input image while adding more details togenerate a realistic high-resolution image. For the discriminator, its structure is similar to that of the Stage-I discriminator with only extra down-sampling blocks since the image size is larger in this stage. To explicitly enforce GAN to learn better alignment between the image and the conditioning text, rather than using the vanilla discriminator, we adopt the matching-aware discriminator proposed by Reed et al. for both stages. During training, the discriminator takes real images and their corresponding text descriptions as positive sample pairs, whereas negative sample pairs consist of two groups. The first is real images with mismatched text embeddings, while the second is synthetic images with their corresponding text embeddings.

RESULTS AND DISCUSSION

By performing comparative analysis with Deep GAN considering various performance metrics such as Accuracy, Sensitivity, Specificity, Precision, Average Recall, and F-Measure Deep Stacked GAN is having better results than Deep GAN. So we conclude that Stacked GAN accuracy is better than Deep GAN.

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

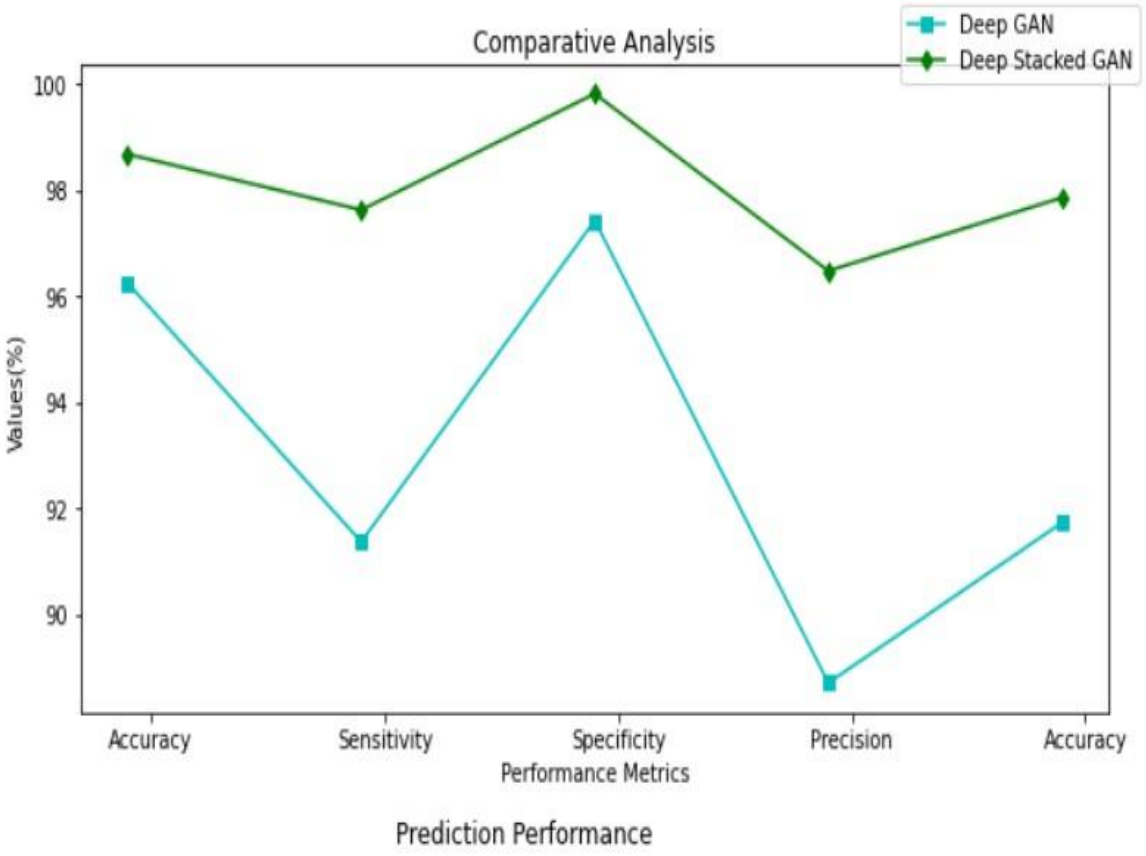

Figure 2. Prediction Performance 1

https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

	Average Precision	Average Recall	F-Measure
Deep GAN	0.93	0.92	0.91
Deep Stacked GAN	0.95	0.96	0.94

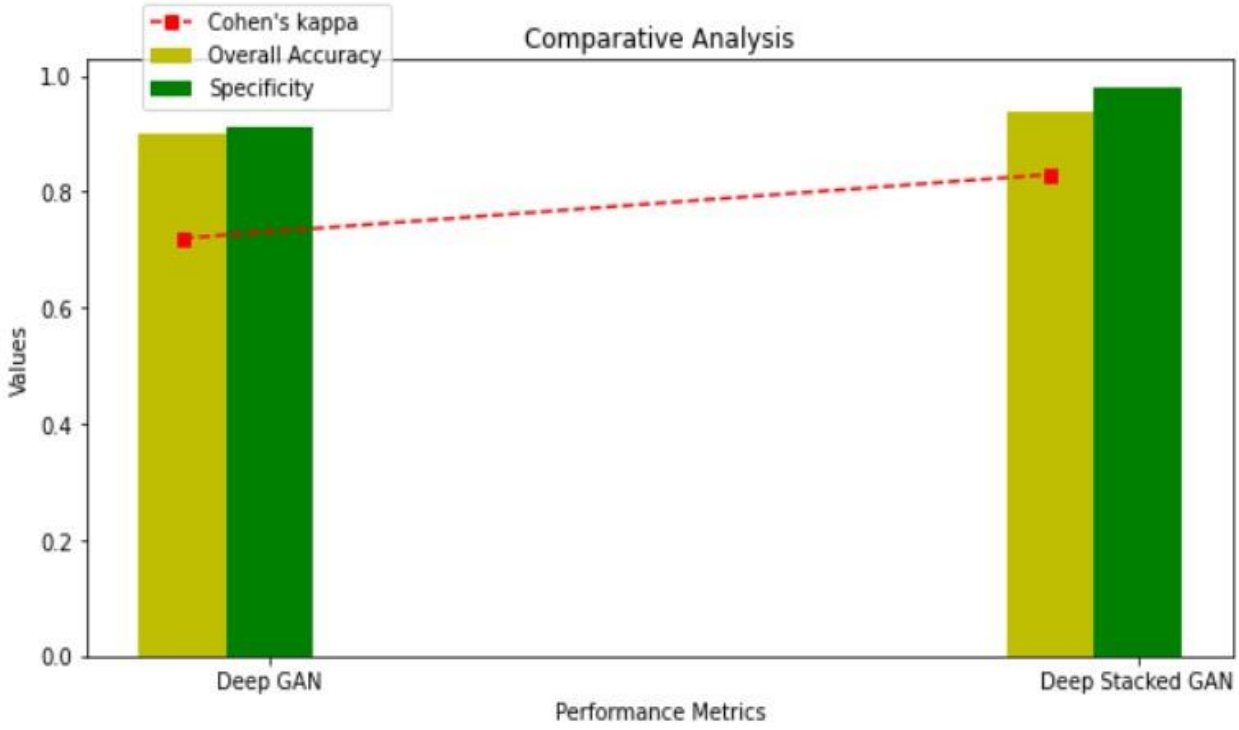

Figure 3. Prediction Performance 2

https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

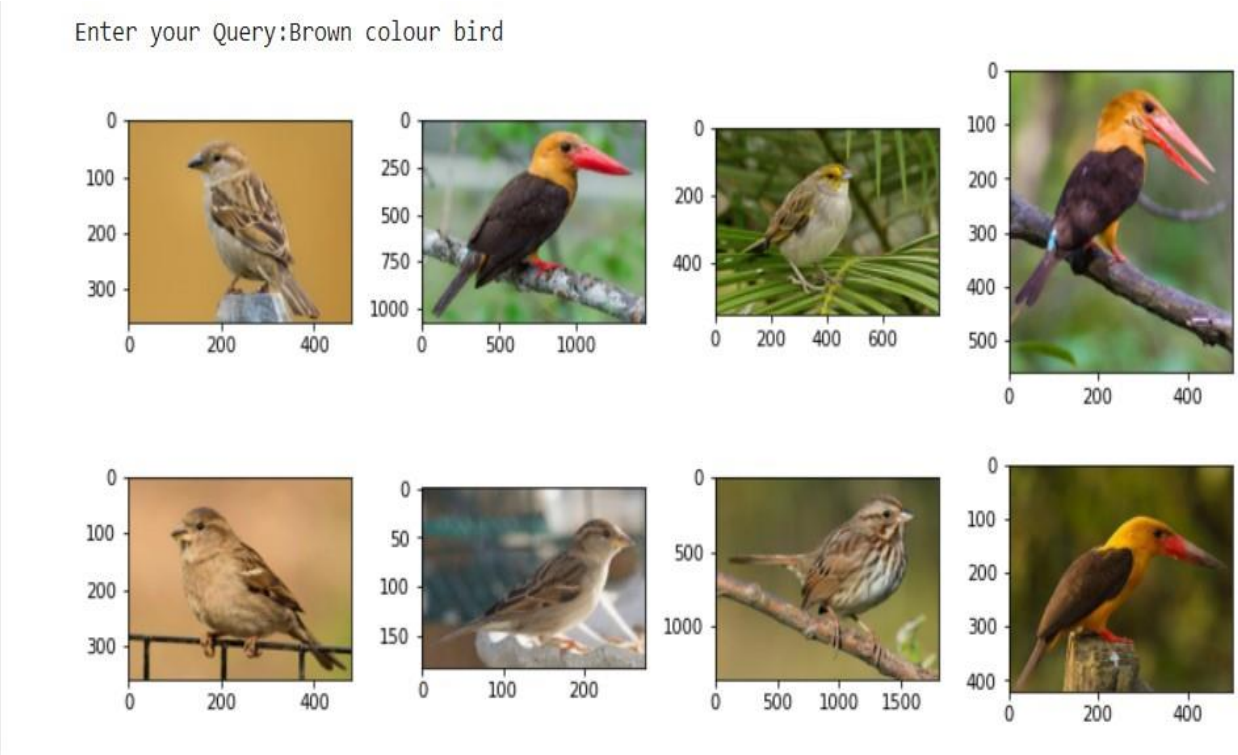

Figure 4. Brown Colour Bird

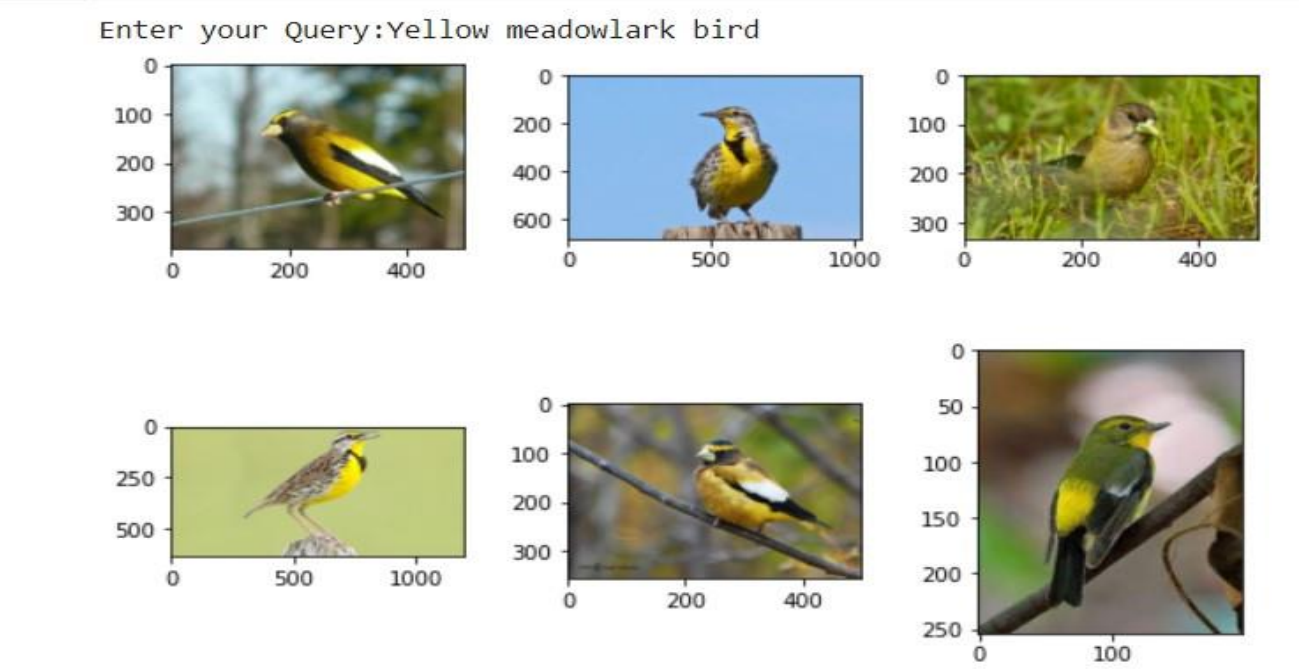

Figure 5. Yellow Meadowlark Bird

https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

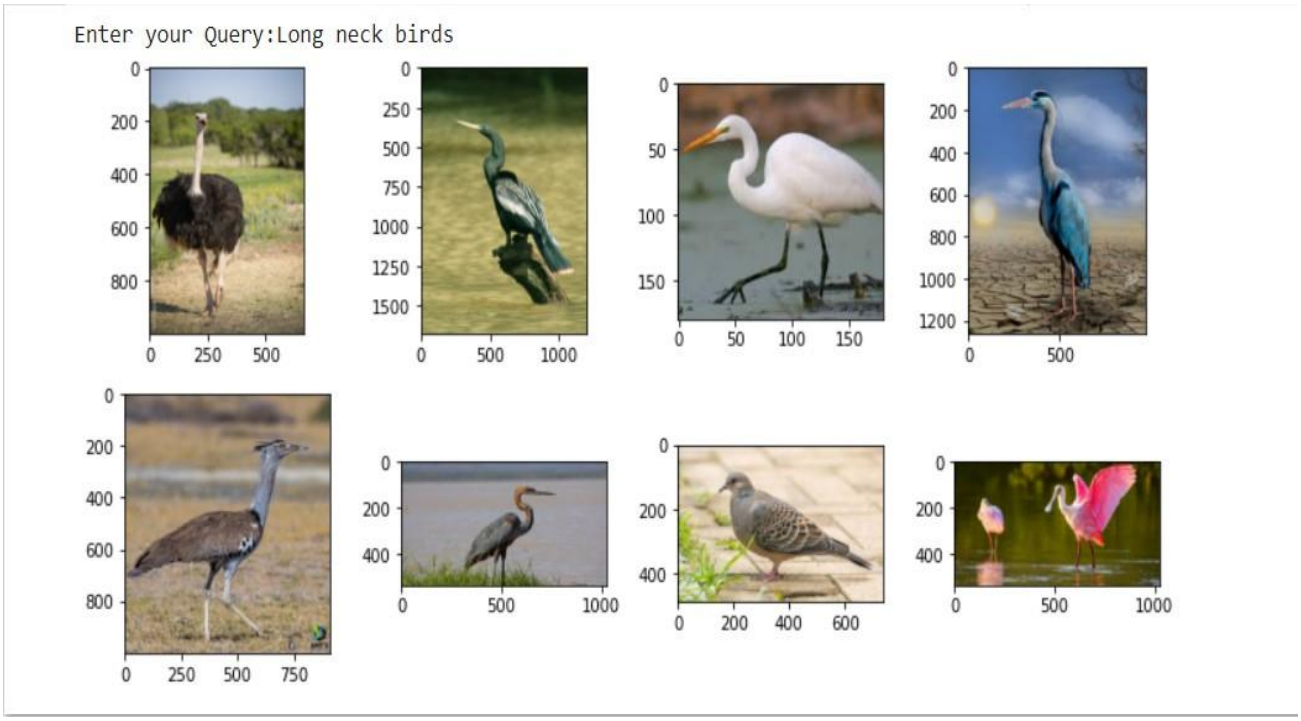

Figure 6. Long Neck Bird

CONCLUSION

This paper proposes Stacked Generative Adversarial Networks (StackGAN) with Conditioning Augmentation for synthesizing photo-realistic images. The proposed method decomposes the text-to-image synthesis into a novel sketch-refinement process. Stage-I GAN sketches the object following basic color and shape constraints from given text descriptions. Stage-II GAN corrects the defects in Stage-I results and adds more details, yielding higher resolution images with better image quality. Extensive quantitative and qualitative results demonstrate the effectiveness of our proposed method. Compared to existing text-to-image generative models, our method generates higher resolution images (e.g., 256 x 256) with more photo-realistic details and diversity. The limitation is that comparing performance metrics with other GANs is a time-consuming process. The future work is planning to do a comparative analysis with the rest of the GANs.

REFERENCES

1. Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiaoyuan Jing, Fei Wu and Bingkun Bao, "**Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis**",Cornell University (arVix:2008.05865), 24 Mar 2021.
2. Minfeng Zhu, Pingbo Pan, Wei Chen and Yi Yang, "**DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis**", Cornell University (arVix:1904.01310), 2 Apr 2019.

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

3. Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao, "**MirrorGAN: Learning Text-to-image Generation by Redescription**", Cornell University (arVix:1903.05854), 14 Mar 2019.
4. Shirin Nasr Esfahani and Shahram Latifi,"**Image generation with GANs-Based Techniques: A Survey**",International Journal of Computer Science & Information Technology (IJCSIT) Vol 11, No 5, October 2019.
5. Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang and Xiaodong He,"**AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks**", Cornell University (arVix:1711.10485), 28 Nov 2017.
6. Kai Hu, Wentong Liao, Michael Ying Yang and Bodo Rosenhahn,"**Text to Image Generation with Semantic-Spatial Aware GAN**", Cornell University (arVix:2104.00567), 24 Apr 2021.
7. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang and Dimitris Metaxas,"**StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks**", Cornell University (arVix:1612.03242), 5 Aug 2017.
8. Dr. Priyanka Mishra,Tribhuvan Singh Rathore,Shivani and Sachin Tendulkar,"**Text to Image Synthesis using Residual GAN**", 2020 3rdInternational Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), 8 Feb 2020.
9. Zixu Wang, Zhe Quan, Zhi-Jie Wang, Xinjian Hu and Yangyang Chen,"**Text to Image Synthesis with Bi Directional Generative Adversarial Networks**",2020 IEEE International Conference on Multimedia and Expo (ICME),10 July 2020.
10. Jezia Zakraoui , Moutaz Saleh, Somaya Al-Maadeed and Jihad Mohammed Jaam, "**Improving text-to-image generation with object layout guidance**", Multimedia Tools and Applications (2021) 80:27423–27443, 20 May 2021.
11. Satya Krishna Gorti and Jeremy Ma, "**Text-to-Image-to-Text Translation using Cycle Consistent Adversarial Networks**", Cornell University (arVix:1808.0538),14 Aug 2018.
12. Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba and Ruslan Salakhutdinov,"**Generating Images From Captions With Attention**",Cornell University (arVix:1511.02793), 29 Feb 2016.
13. Ming Ding, Zhuoyi Yang , Wenyi Hong , Wendi Zheng , Chang Zhou , Da Yin , Junyang Lin, Xu Zou , Zhou Shao, Hongxia Yang and Jie Tang,"**CogView: Mastering Text-to-Image Generation via Transformers**", Cornell University (arVix:2105.13290), 5 Nov 2021.
14. Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz and Philip H. S. Torr,"**Controllable Text-to-Image Generation**",Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 19 Dec 2019.
15. Tobias Hinz, Stefan Heinrich, and Stefan Wermter,"**Semantic Object Accuracy for Generative Text-to-Image Synthesis**",Cornell University (arVix:1910.13321),2 Jun 2020.
16. Weihao Xia, Yujiu Yang, Jing-Hao Xue and Baoyuan Wu,"**TediGAN: Text-Guided Diverse Face Image Generation and Manipulation**",Cornell University (arVix:2012.03308), 29 Mar 2021.
17. Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee and Yinfei Yang,"**Cross-Modal Contrastive Learning for Text-to-Image Generation**",Cornell University (arVix:2101.04702), 9 Jun 2021.

 https://doi.org/10.37896/pd91.4/91449

ISSN: 0369-8963

18. Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su and Qiang Liu,"**FuseDream: Training-Free Text-to-Image Generation with Improved CLIP+GAN Space Optimization**",Cornell University (arVix:2112.01573), 2 Dec 2021.
19. Jaemin Cho, Abhay Zala and Mohit Bansal,"**DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers**",Cornell University (arVix:2202.04053), 8 Feb 2022.
20. Md Aminul Haque, Palash Md Abdulla Al Nasim, Aditi Dhali and Faria Afrin,"**Fine-Grained Image Generation from Bangla Text Description using Attentional Generative Adversarial Network**",Cornell University (arVix:2109.11749), 24 Sep 2021.

 https://doi.org/10.37896/pd91.4/91449