PHOTOGRAPHIC TEXT-TO-IMAGE SYNTHESIS VIA MULTI-TURN DIALOGUE USING ATTENTIONAL GAN

Shiva Kumar Shrestha[1], Shashidhar Ram Joshi[2]

[1]*Department of Computer Engg., Khwopa College of Engineering, Bhaktapur, Nepal*
[2]*Department of Electronics and Computer Engg., Pulchwok Campus, Kathmandu, Nepal*

Abstract

The process of generating an image that depicts naturalness is not so easy. To address such problem this paper introduces a novel approach to synthesize a photo-realistic image from the caption. The user can adjust the image highlights turn-by-turn according to the caption. This leads to the integration of natural intelligence. For this, the input passed to dialogue state tracker to extract context feature. Then the generator produces an image. If image is not as per expectations then user gives another dialogue, but the system takes both recent input and previous image to generate a new one. In such a manner, user gets a chance to visualize as per the imagination. We performed extensive experiments on two datasets CUB and COCO to generate a realistic image each turn and obtained the results: Inception Score (IS) of 4.38 ± 0.05, R-precision of 67.96 ± 5.27 % on CUB dataset and IS of 26.12 ± 0.24, R-precision of 91.00 ± 2.31 % on COCO dataset. Further, the work could be enhance to synthesize HQ image, voice integration, and video generation from stories and so on. This research is limited to 256x256 image in each turn.

Keywords: *GAN, MultiTurnGAN, Text-to-image, Image generation, Realistic image synthesis*

1. Introduction

In short period of time since Generative Adversarial Network (GAN)'s introduction at 2014, many different enhancement methods (Xu *et al.*, 2018), (Zhang *et al.*, 2019) and training variants have been suggested to improve their performance regarding the synthesis of artificial data. The process of generating a photo-realistic image from the text, is one of the important problems(Xu *et al.*, 2018) in computer science. Moreover, GAN is the emerging technology and has tremendous applications, such as photo editing, CAD, interactive graphic, and so on. The basic structure of GAN shown in Fig. 1.

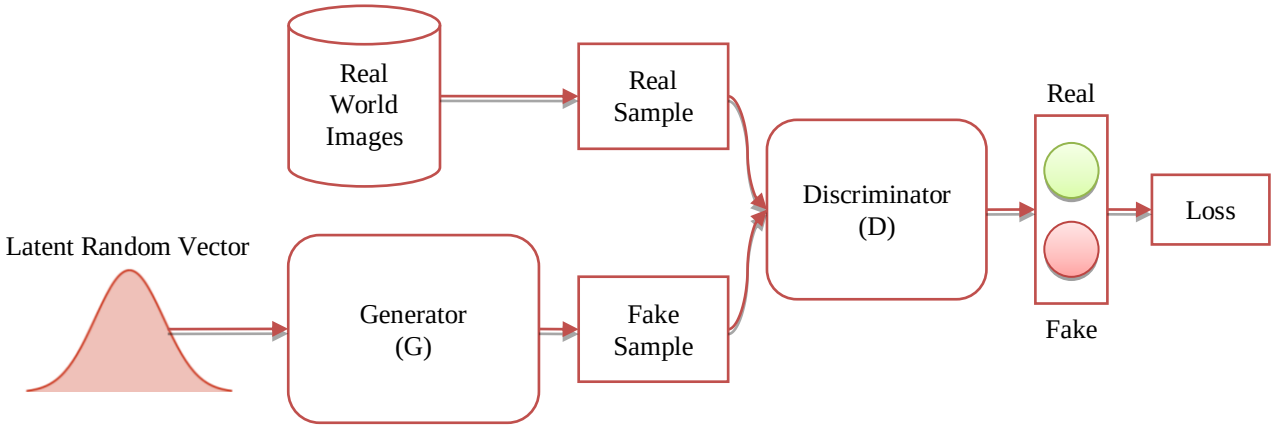

Fig. 1 Structure of GAN - Modified from McGuniess (2016)

There are impressive results from various researches works but that lacks to use conversational text/description, which can help to generate photo-realistic images with visual details.

*Corresponding author: Shiva Kumar Shrestha
Department of Computer Engineering, Khwopa College of Engineering, Libali – 8, Bhaktapur, Nepal
Email: ersks@khwopa.edu.np
(Received: December 08, 2019 Accepted: August 19, 2020)

2. Literature Review

The conditioning strategy is used to facilitate the smoothness of the latent conditioning(Chen Q.& Koltun, 2017). This approach increases the variety of synthesized images. The basic concept of the work is to convert the text definition into vectors and to generate the images from the given vectors using GAN(Goodfellow *et al.*, 2014).Some of these works generate 64x64 image where the proposed work generates 256x256 images.

Editing an image using Language-based captions (J. Chen *et al.*, 2018), (Manuvinakurike *et al.*, 2019) is a task designed for minimizing labor work while helping users create visual data. Some systems tried to figure-out the specific part that the user focusing. This is a very tedious and complex task, but it will be simpler if user and system have comprehensive understanding of both visual information and natural language. Some authors proposed a model for image editing via conversational language(Cheng *et al.*, 2018).

Many works used multiple GANs (Zhang *et al.*, 2019) to improve image quality. Some paper used GAN design and style to synthesize indoor scenes(Wang & Gupta, 2016).Some works added many GANs, but that could not generate high-resolution images with photo-realistic information(Huang *et al.*, 2018)where proposed single model can generate realistic image. The author(Goodfellow *et al.*, 2014) presented a method for producing photographic images based on semantic image descriptions by balancing the convergence between generators and discriminator sand stably modeling the huge pixel space in high-resolution images to ensure semantic consistency (Zhang *et al.*, 2019). Although some paper generated 1024x1024 image and more, we took idea of generating 256x256 image that is sufficient for visual aspects. Learning to generate meaningful and coherent sequences of images from a story(Li *et al.*, 2019) is a challenging task. However, the proposed work generates new image in every turn for the caption given rather than visualizing the whole story by several images. Translating a function from one image to another is a class of computer vision and graphics issues(Reed, Akata, Yan, *et al.*, 2016 a). Only the core concept of feature implementation for image synthesis taken from this work.

The research work that introduced an exploration work which incorporates content-to-picture and picture-to-content (picture inscribing) union to improve the exhibition of content-to-picture blend(Dong *et al.*, 2018).Another author proposed a novel worldwide neighborhood, mindful and semantic-safeguarding content-to-picture-to-content system, called MirrorGAN(Qiao *et al.*, 2019). Turbo learning deals with caption-to-image and vice versa for preparing a caption-to-picture generator (CaptionBot) and a caption-to-picture generator (DrawingBot) (Huang *et al.*, 2018).Various parallel works are coming weakly regarding the synthesis of text, image, sound, video, etc. Among them, only the concept of image generation is grabbed and attention driven dialogue mapping(Cheng *et al.*, 2018) is done. All the above papers have focused on the generation of artificial data/images. A large number of tasks have taken significant steps in several related areas, including image-to-text-to-image generation, drawing images, dialogue-based image synthesis, language-based image editing, generating images from text, image-to-image translation, image captioning, visual question-answering and so on.

3. Methodology

Multi-turn dialogue used in this research work for image generation, so the name for the network given as MultiTurnGAN. It is an advancement of the network, AttnGAN, in which the attention module focuses the attention of the network to generate fine-grained image from text. However, this GAN adds the attention of whole network in both text-to-image consistency and image quality to maintain naturalness in generated image. For the simplicity, only two steps shown in Fig.2. The global attention module uses dialog for automatic production of different sub-region objects. In addition, the DMS works to minimize the loss of matching fine-grained image-text. The model uses forward pass where user response $\{o_1, o_2, \dots, o_t\}$are passed to state tracker to generate the new image $(\widehat{x_t})$.

3.1. Global Attention Module

The user response o_t and previous image feature $h_{t1}$are passed to F_{attn} (). For the i^{th} sub-region of

the image (the i^{th} column of h_{t1}), a word-context vector c_i can be obtained by learning the attention weights of every word in o_t given the i^{th}sub-region of the image.

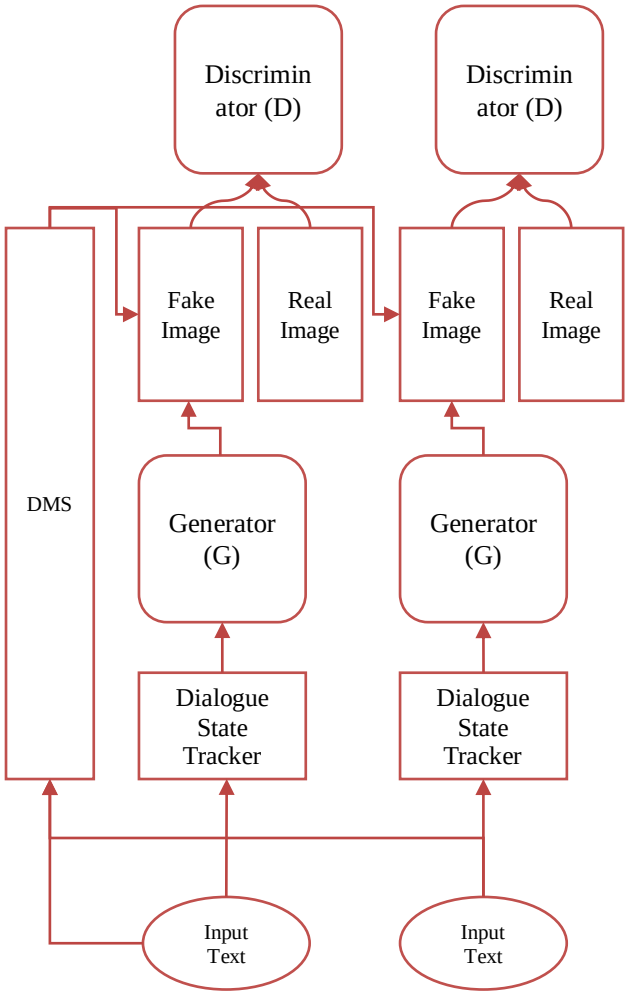

Fig. 2 Block Diagram of MultiTurnGAN

To generate an image in t^{th}step, $F_{attn}(o_t, h_{t-1})$ outputs a word-context matrix $(c_0, c_1, \dots, c_i, \dots)$ which is passed to the neural state tracker.The losses of G and D calculated as follows:

$$L_G = -\frac{1}{2}E_{x_t \sim P_G}[\log(D_t(x_t))] - \frac{1}{2}E_{x_t \sim P_G}[\log(D_t(x_t, h_t))] \quad (1)$$

$$L_D = -\frac{1}{2}E_{x_t \sim P_{data}}[\log(D_t(x_t))] - \frac{1}{2}E_{\widehat{x_t} \sim P_G}[\log(1 - D_t(\widehat{x_t}))] - \frac{1}{2}E_{x_t \sim P_{data}}[\log(D_t(x_t, h_t))] - \frac{1}{2}E_{\widehat{x_t} \sim P_G}[\log(1 - D_t(\widehat{x_t}, h_t))] \quad (2)$$

wherex_t is from the true data distribution P_{data} and $\widehat{x_t}$ is from the model distribution P_G.The AttnGAN(Xu *et al.*, 2018)uses different generator to generate 64x64, 128x128 and 256x256 images but the MultiTurnGAN use the same generator for different scales.

3.2. DMS Regularizer

Deep Multimodal Similarity Regularizer (DMS regularizer) developed to balance the semantics of the dialogues, and it stabilizes the image generator. Reference feedback and attributes are combined by the DMS as the text, which is different from those mentioned in Xu *et al.*(2018). The posterior probability of caption D_i matching the picture I_i is defined as:

$$P\left(\frac{D_i}{I_i}\right) = \frac{\exp(\gamma R(I_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma R(I_i, D_i))} \quad (3)$$

where γ is a smoothing factor. $R(I_i, D_i)$is the word-level matching score from attention-driven image-text pair. The loss function for matching photos with their corresponding captions for batch of M pairs is:

$$L_{DMS}^{i \to d} = -\sum_{i=1}^{M} \log P\left(\frac{D_i}{Ii}\right) \quad (4)$$

The loss function for matching captions with their corresponding pictures can be calculated by switching D_i and I_i. DMS function by combining both equations:

$$L_{DMS} = L_{DMS}^{i \to d} + L_{DMS}^{d \to i} \quad (5)$$

Overall objective function is:

$$L_{DMS} = -\frac{1}{T}\sum_{t=1}^{T} L_G + L_D + \lambda L_{DMS} \quad (6)$$

Where λ is the hyper-parameter to balance two losses.

3.3. Hyper Parameters

The hyper-parameters with corresponding values selected for the research works are termed as configurations. These were referenced from Xu *et al.*(2018)because this is the best method (IS: 4.36 ± 0.03, R-precision: 67.82 ± 4.43% on CUB and IS: 25.89 ± 0.47, R-Precision: 85.47 ± 3.69 on COCO) in state-of-the-art (SOTA) for CUB and COCO datasets. The fine-tuned values of base size, maximum no. of epochs, snapshot interval,l earning rate, smoothing factors, embedding dimensions and so on referenced for the model design, MultiTurnModel. The new configuration for our GAN are in each caption splitted into three/four parts for validation of model with previous SOTA.

3.4. Results and Discussions

The network is able to evaluate IS of 4.38 ± 0.05 and R-precision of 67.96 ± 5.27 on CUB dataset. In addition, IS of 26.12 ± 0.24and R-precision of 91.00 ± 2.31 was achieved on COCO dataset.

To obtain the result, the DMS Regularizer trained to balance the attention mechanism as per semantics of dialogues and then MultiTurnGAN trained. For the simplicity, Fig. 3 shows the turn-wise image synthesis by our system in four steps. Fig. 4 shows that the synthesized images of birds seem photo-realistic. Fig. 5 shows the outcomes comparison where our results look more attractive.

The inception score and R-precision calculated to validate the purposed model as shown in Fig. 6, Fig. 7, Fig. 8 and Fig. 9.This research work purposed the prototype for dialogue base text-to-image (T2I) synthesis. Based on two test sets, CUB and COCO, the MultiTurnGAN is compared to previous GAN models GAN-INT-CLS(Reed, Akata, Yan, *et al.*, 2016 c), GAWWN (Reed, Akata, Mohan, *et al.*, 2016 b), StackGAN v1 (Zhang *et al.*, 2017), StackGAN v2 (Zhang *et al.*, 2019) and AttnGAN(Xu *et al.*, 2018) for text-to-image (T2I) generation. The COCO dataset was more difficult than the CUB dataset as it contained object varieties with more complex scenarios. Examples in Fig.10 and Fig.11 illustrate that the caption used in previous research were broken down into three text-description just for model validation. Actually, the network designed to integrate human intelligence. In addition, the MultiTurnGAN is able to generate 256x256 resolution images step wise (in each turn) for different scenarios, although the COCO dataset images produced are not as photo-realistic as the CUB dataset images. The experimental results showed that, to create complex scenes, the MultiTurnGAN is more successful than other previous approaches.

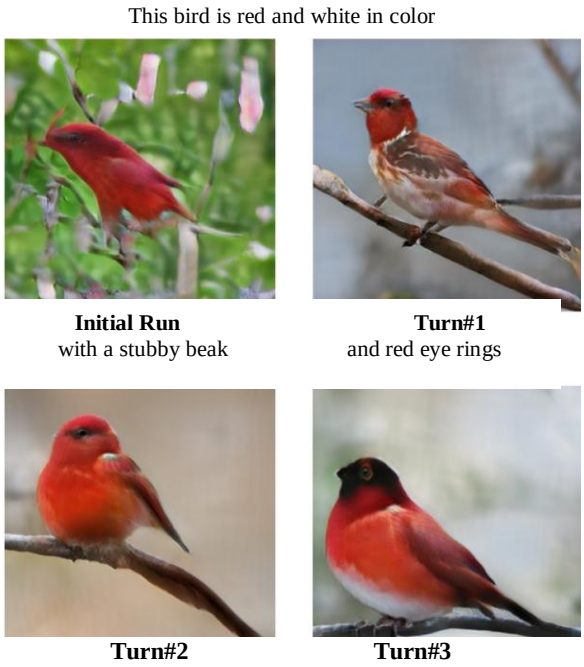

Fig. 3 Turn wise image synthesis by MultiTurnGAN

Table 1 shows the effectiveness of proposed model. Here the IS and R-precision was only compared with AttnGAN(Xu *et al.*, 2018) for being best method for selected datasets.

Fig. 4 Various synthesized images, of birds by our GAN (MultiTurnGAN), seems photo-realistic.

Fig.5 Comparison of outcomes of GAN-INT-CLS, GAWWN, Stack GAN v2, AttnGAN and MultiTurnGAN on CUB test sets

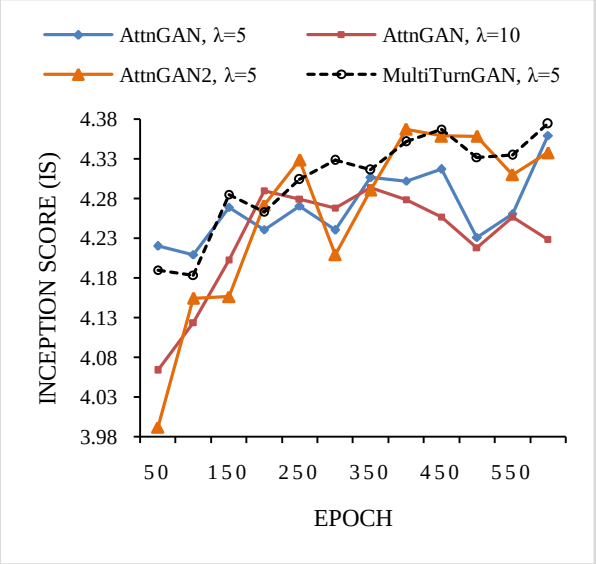

Fig. 6 Inception Score (IS) on CUB Test Sets

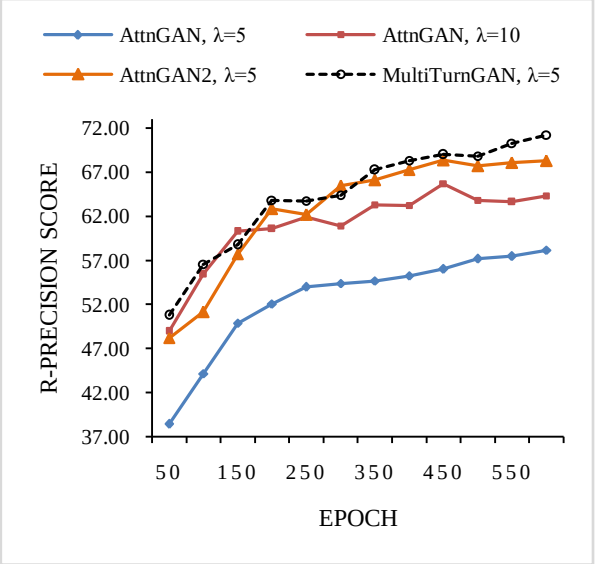

Fig. 7 R-precision Score on CUB Test Sets

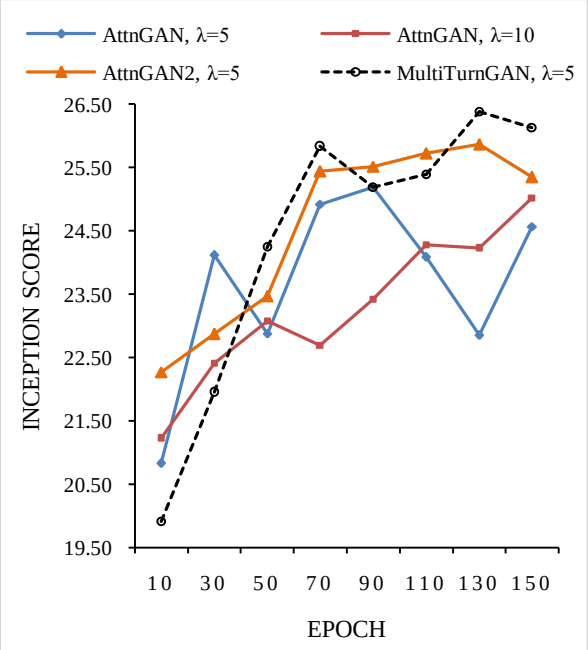

Fig.8 Inception Score on COCO Test Sets

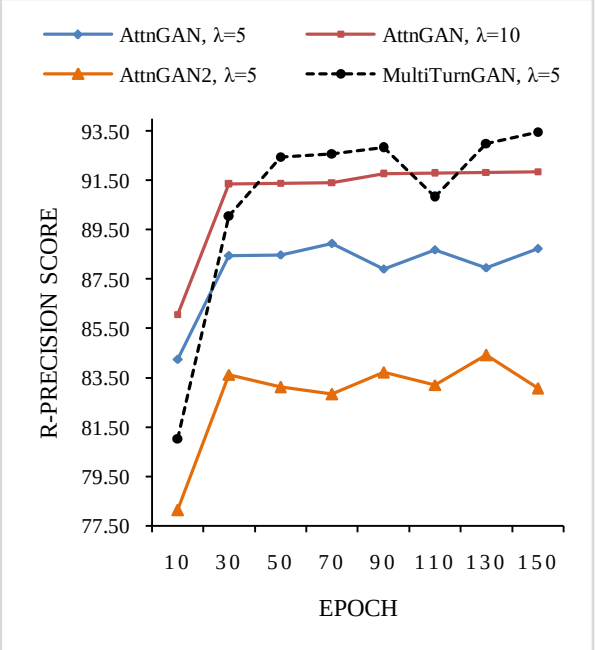

Fig.9 R-precision Score on COCO Test Sets

Inception score and R-precisions of our GAN just bid the AttnGAN2, latest experiments of AttnGAN. We can see the scores in Table 1 that contains only effective values. The scores of GAN-INT-CLS(Reed, Akata, Yan, *et al.*, 2016 c), GAWWN(Reed, Akata, Mohan, *et al.*, 2016 b) were low because these are the basic network for T2I generation. Inception scores of StackGAN v1 (Zhang *et al.*, 2017) were 3.70 ± 0.04 on CUB, 8.45 ± 0.03 on COCO which are lower than ours. The second version of StackGAN(Zhang *et al.*, 2019) gave IS: 4.04 ±0.05 on CUB and IS: 8.30 ± 0.10. These values are lower than about 0.34 and 59.66 on CUB and COCO datasets respectively. However, they computed other scores for the model validation and used other datasets.

Table 1: Results on CUB and COCO test sets

Method	**Inception Score**	**R-Precision (%)**
AttnGAN, λ = 5.0	4.35±0.04	58.65±5.41
AttnGAN, λ = 10.0	4.29±0.05	63.87±4.85
AttnGAN2, λ = 5.0	4.36±0.03	67.82±4.43
MultiTurnGAN, λ = 5.0	4.38±0.05	67.96±5.27
AttnGAN2 (COCO), λ = 50	25.89±0.47	85.47±3.69
MultiTurnGAN (COCO), λ = 50	26.12±0.24	91.00±2.31

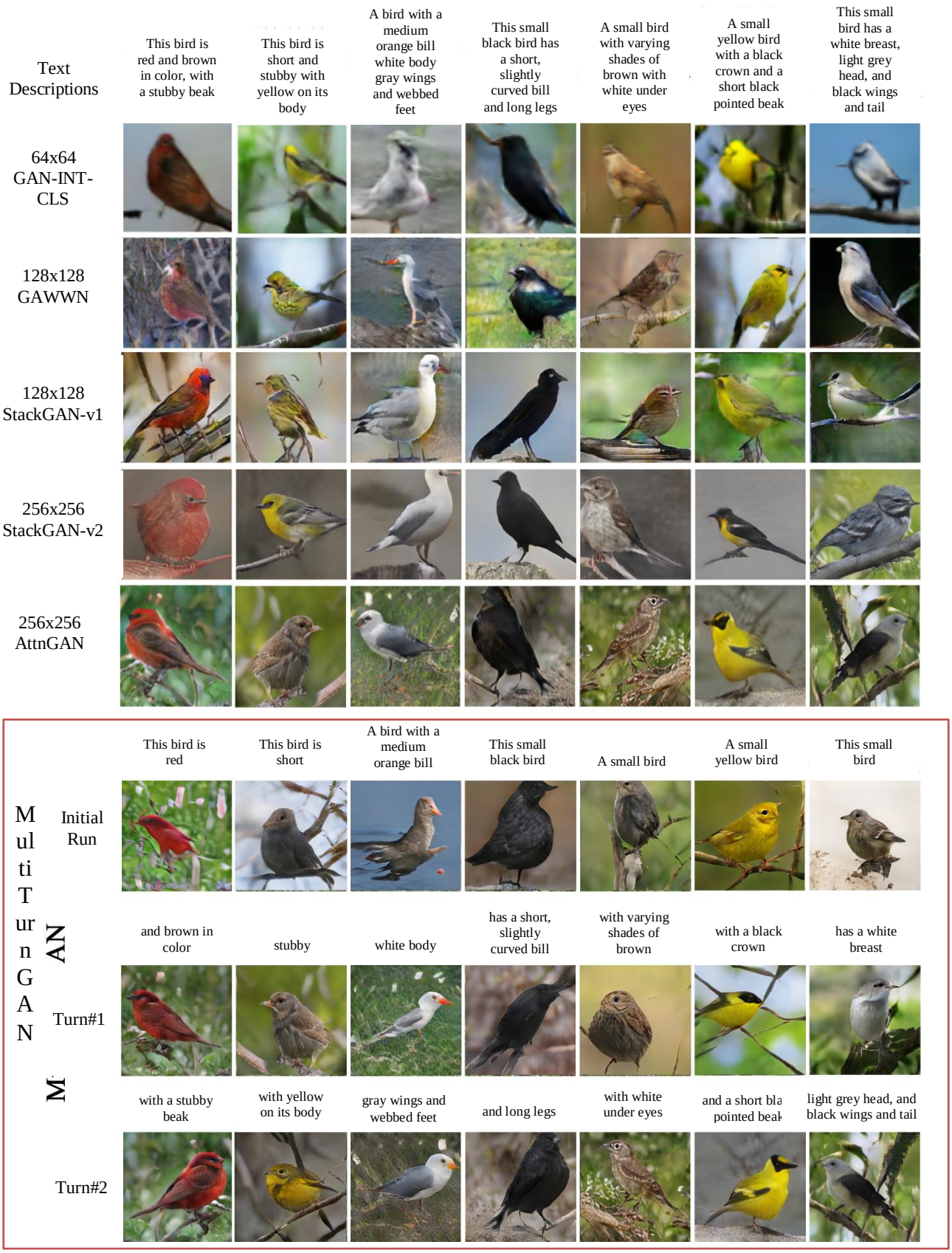

Fig. 10 Outcome comparison of GAN-INT-CLS, GAWWN, StackGAN v1, StackGAN v2, AttnGAN and MultiTurnGAN conditioned on text captions from CUB dataset (The colored picture is available on online version)

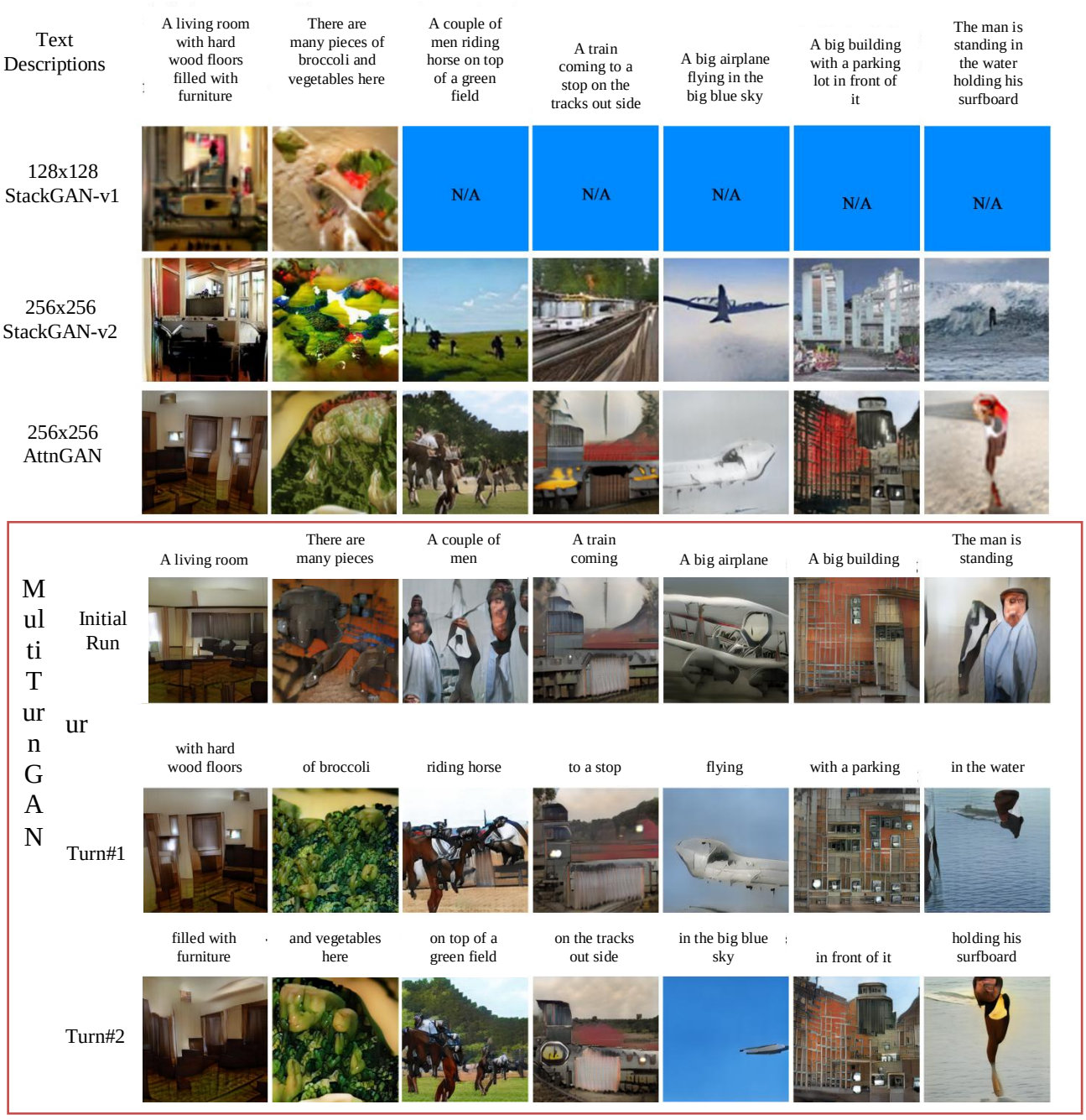

Fig. 11 Outcome of StackGANv1, StackGANv2, AttnGAN and MultiTurnGAN conditioned on text captions from COCO dataset (The colored picture is available on online version)

3.5. Conclusions and Recommendations

This research work proposes, a multi-turn network for the synthesis of fine-grained photo-realistic text-to-image, MultiTurnGAN designed to generate high-quality image in different turns with novel attention mechanism. In addition, a DMS regularizer developed to measure the loss of matching fine-grained image-text to train the MultiTurnGAN generator. This GAN outperforms GAN models substantially by improving the best recorded IS by 4.38 on the CUB dataset and 26.12 COCO data set as shown in the Table 1. Experiments have shown that the mechanisms proposed in the MultiTurnGAN are successful. The DMS regularizer was first trained, and the GAN then trained. This research proposed to aid natural intelligence from user to machine in turn-wise way. However, the whole caption was splitted into parts to compare with latest models.

There is a huge space where GANs can be applied, and there is still so much to research about them. In particular, GANs have made a lot of progress in Computer Vision recently: image impainting, style

transfer, image enhancement, etc. Nevertheless, new papers and improvements to existing models are being released almost every week. This work can be extended to the following fields:

- Voice to Text-to-Image Generation and Editing,
- Human Computer Interfacing,
- Brain Imagination to Text to Image Synthesis,
- Face Synthesis using CelebA Data-set,
- Video Generation from story,
- HQ Photo Generation from the caption,
- Generation and use of other datasets, etc.

References

[1] Chen, J., Shen, Y., Gao, J., Liu, J., & Liu, X. (2018). Language-Based Image Editing with Recurrent Attentive Models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8721–8729.

[2] Chen, Q., & Koltun, V. (2017). Photographic Image Synthesis with Cascaded Refinement Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October, 1520–1529.

[3] Cheng, Y., Gan, Z., Li, Y., Liu, J., & Gao, J. (2018). Sequential attention gan for interactive image editing via dialogue. *arXiv preprint* arXiv:1812.08352.

[4] Dong, H., Zhang, J., McIlwraith, D., & Guo, Y. (2018). I2T2I: Learning text to image synthesis with textual data augmentation. Proceedings - International Conference on Image Processing, ICIP, 2017-September, 2015–2019.

[5] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 3(January), 2672–2680.

[6] Huang, Q., Wu, D., Zhang, P., & Zhang, L. (2018). Turbo learning for captionbot and drawing bot. Advances in Neural Information Processing Systems, 2018-Decem (NeurIPS), 6455–6465.

[7] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., &Gao, J. (2019). Storygan: A sequential conditional gan for story visualization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 6322–6331.

[8] Manuvinakurike, R., Bui, T., Chang, W., & Georgila, K. (2018, July). Conversational image editing: Incremental intent identification in a new dialogue task. In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (pp. 284-295).

[9] McGuinness, Kevin (2016). Deep Learning for Computer Vision: Generative models and adversarial training, Slide Share, https://www.slideshare.net/xavigiro/deep-learning-for-computer-vision-generative-models-and-adversarial-training-upc-2016 Accessed on April 15, 2020.

[10] Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 1505–1514.

[11] Reed, S., Akata, Z., Lee, H., & Schiele, B. (2016) a. Learning deep representations of fine-grained visual descriptions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 49–58.

[12] Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016)b. Learning what and where to draw. Advances in Neural Information Processing Systems, Nips, 217–225.

[13] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016) c. Generative adversarial text to image synthesis. 33rd International Conference on Machine Learning, ICML 2016, 3, 1681–1690.

[14] Wang, X., & Gupta, A. (2016). Generative image modeling using style and structure adversarial networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9908 LNCS, 318–335.

[15] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1316–1324.

[16] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

[17] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2019). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 1947–1962.