Article

Generating Scenery Images with Larger Variety According to User Descriptions

Hsu-Yung Cheng ^{1,*} and Chih-Chang Yu ² ¹ Department of Computer Science and Information Engineering, National Central University, 300 Jong-Da Rd., Jong-Li, Taoyuan 32001, Taiwan² Department of Information and Computer Engineering, Chung Yuan Christian University, 200 Chung Pei Road, Chung-Li, Taoyuan 32023, Taiwan; ccyu@cycu.edu.tw

* Correspondence: chengsy@csie.ncu.edu.tw

Abstract: In this paper, a framework based on generative adversarial networks is proposed to perform nature-scenery generation according to descriptions from the users. The desired place, time and seasons of the generated scenes can be specified with the help of text-to-image generation techniques. The framework improves and modifies the architecture of a generative adversarial network with attention models by adding the imagination models. The proposed attentional and imaginative generative network uses the hidden layer information to initialize the memory cell of the recurrent neural network to produce the desired photos. A data set containing different categories of scenery images is established to train the proposed system. The experiments validate that the proposed method is able to increase the quality and diversity of the generated images compared to the existing method. A possible application of road image generation for data augmentation is also demonstrated in the experimental results.

Citation: Cheng, H.-Y.; Yu, C.-C. Generating Scenery Images with Larger Variety According to User Descriptions. *Appl. Sci.* **2021**, *11*, 10224. <https://doi.org/10.3390/app112110224>

Academic Editor: Yuan-Kai Wang

Received: 14 September 2021

Accepted: 25 October 2021

Published: 1 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: text to image; image generation; generative adversarial networks

1. Introduction

Generative adversarial networks (GAN) [1] have become an active research area since they were proposed. The principle of a generative adversarial networks is as follows. The generator generates fake data that serves as negative training samples for the discriminator. The discriminator learns to distinguish the real data from the fake data generated by the generator. Then, the discriminator penalizes the generator for producing implausible results. The generator learns to mimic the real data and generate better results hoping to fool the discriminator. Inspired by game theory, the generator and discriminator complete with each other to achieve the Nash equilibrium in the training process. As both the generator and the discriminator are well-trained, the GAN can produce plausible results. It usually requires a lot of training data and computation time to train a steady GAN model. Due to the weakness of the original GAN, various improved GANs models in terms of architecture optimization and objective function optimization have been proposed [2]. Radford et al. [3] proposed a Deep Convolutional Generative Adversarial Networks (DCGAN) which replaces the fully connected layer with the deconvolution layer in the generator to achieve better performance in image generation tasks. Mirza and Osindero [4] proposed Conditional Generative Adversarial Networks (CGANs), which introduce the conditional variable in both the generator and discriminator so that conditions such as labels, text or other data can be added to control to the model. Chen et al. [5] proposed another CGAN called InfoGAN by introducing mutual information and making the generation process more controllable and more interpretable. Based on CGAN, Odena et al. [6] proposed an Auxiliary Classifier GAN (ACGAN). The condition variable is not added in the discriminator, and another classifier is used to indicate the probability over the class labels.

The loss function is also modified to increase the probability of correct prediction for the discriminator.

In the training process of GAN, mode collapse has always been a problem. Sometimes it is feasible to fine tune the training parameters for better model learning. However, tuning parameters might not always work well. Therefore, Arjovsky et al. [7] used Wasserstein distance to improve the stability of learning and mode collapse. They showed that the Earth-Mover distance could produce better gradient behaviors when performing distribution learning compared with other distance metrics. Metz et al. [8] proposed an unrolled GAN that uses a gradient-based loss function to enhance the generator and applies the Jensen-Shannon (JS) divergence to minimize the loss function of the generator. Nguyen et al. [9] proposed a dual discriminator architecture. One of the discriminators rewards high scores for samples from true data distribution and the other discriminator rewards data from the generator, and thus produces data to fool both two discriminators.

The research area of GAN has many applications including art, design, data augmentation, and other applications [10–14]. For example, the research in [5] generates hand-written numbers. The DRAW system [15] constructs license plates. The work in [16] generates anime characters. The system in [3] demonstrates generating bedroom interiors using deep convolutional generative adversarial networks. PixelCNN [17] generates various different animals and human faces. Ledig et al. [18] proposed a super-resolution GAN to take a low-resolution image as input and generate a high-resolution image. Isola et al. [19] proposed an image-to-image translation framework called pix2pix based on CGAN to convert an image content from one domain to another. Iizuka et al. [20] proposed to perform image inpainting using two discriminators to monitor local features and global features for both local and global consistency. The development of the GAN techniques also leads to promising results in some novel applications such as 3D object generation, medicine, and autonomous driving [21]. Yu et al. [22] proposed a 3D object inpainting network that can process point cloud data directly without human labeling. Shin et al. [23] proposed a GAN-based method to generate synthetic, abnormal, MRI images with brain tumors. The work in [24] designed a systematic framework to automatically generate test cases to imitate real-world driving scenes.

With the progress of computer vision and natural language analysis, generating images according to natural language descriptions has become a popular research topic in recent years [25–32]. The work in [26] was one of the first works that attempted to generate images according to text descriptions. It translates single-sentence human-written descriptions directly into image pixels. Images of bird and flower images are synthesized based on human-written texts using a character-level text encoder and class-conditional GAN. It is the first end-to-end differentiable architecture from the character level to pixel level, which introduces a manifold interpolation regularizer for the GAN generator that significantly improves the quality of generated samples. In [27], a semantic compositional network is proposed for image captioning. The network uses a Long Short-Term Memory (LSTM) network and extends each weight matrix of the LSTM to an ensemble of tag-dependent weight matrices. In earlier research, the network architecture did not include training word embeddings. To address the research topic of the relationship between natural language and images, [28] proposed a network to extract visual descriptions that can represent images. To generate high-resolution photo-realistic images from text descriptions, it is not feasible to simply add more upsampling layers in the network. Therefore, StackGAN [29] used two stages of generative networks to first generate a low-resolution image conditioned on the text description, and then refine the results to produce a high-resolution image. Although some good results have been achieved, these studies are only conditioned on global sentence encoding when generating images. Without important fine-grained information at the word level, this degrades the quality of generated images when generating complex scenes. AttnGAN [30] improved the work in [29] by adding embeddings of both word-level and sentence-level descriptions and built an attention model to compute a region-context vector for each word. The work in [32] combined the method of reinforcement learning

in the framework to give higher rewards to novel generated images with unseen features and to give lower rewards to repetitive and similar images.

This work focuses on outdoor scenery image generation. To give the generated images more variety, the proposed system improves the architecture of AttnGAN [30] and adds an imagination mechanism in the framework to increase the diversity of the generated scenes. In this way, the variety of the generated images can be effectively increased. The rest of the paper is organized as follows. Section 2 elaborates the details of the proposed framework. Section 3 shows and discusses the experimental results. Finally, conclusions are made in Section 4.

2. Attentional and Imaginative Generative Networks for Scenery Image Generation

The proposed system framework is illustrated in Figure 1. The framework improves and modifies the architecture of AttnGAN [30] by adding the imagination models in addition to the original attention models. Therefore, the core is named as the attentional and imaginative generative networks. The architecture of the original AttnGAN is displayed in Figure A1 in Appendix A for comparison. We can observe that the main difference between the two architectures is the existence of the imagination model, S_i^{imag} . The details of the attentional and imaginative generative networks in this work are described below. There are m generators ($G_0, G_1, \dots, G_{m-1}$) in the generative model. Each of the m generators has its corresponding stage ($S_0, S_1, \dots, S_{m-1}$) and outputs a hidden layer ($h_0, h_1, \dots, h_{m-1}$) to generate images of lower to higher resolutions ($\hat{x}_0, \hat{x}_1, \dots, \hat{x}_{m-1}$) where $\hat{x}_i = G_i(h_i)$.

Figure 1. System Framework.

In Figure 1, the text encoder is a bidirectional LSTM [33], which extracts the global sentence vectors $\bar{e}$ from the input text descriptions. F_{CA} is the conditioning augmentation [29] that converts the sentence vector $\bar{e}$ to the conditioning vector c . The vector z is a noise vector sampled from a normal distribution. S_i is the upsampling network at each stage. All stages include attention models S_i^{attn} and imagination models S_i^{imag} , except for S_0 . Note that F_{CA} , S_i^{imag} , S_i^{attn} , S_i , and G_i are all modeled as neural networks. The relationship of hidden layers h_i and the upsampling networks S_i can be expressed as Equations (1) and (2).

$$h_0 = S_0(z; F_{CA}(\bar{e})) \quad (1)$$

$$h_i = S_i(h_{i-1}; S_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, \dots, m-1 \quad (2)$$

where e is the matrix of word vectors.

Attention models S_i^{attn} take the word features e and the image features h_{i-1} from the previous hidden layer as the inputs. A new perceptron layer is added to convert

the word features e into the common semantic space of the image features, denoted by $e_i^\sim$. Afterwards, a word-context vector C_j is computed for the j th subregion of the image according to its hidden features h . Each column of h is a feature vector of the j th subregion. The computation of the word-context vector C_j is expressed in Equations (3) and (4). In Equation (4), the softmax function converts $s_{j,i}^\sim$ into $\beta_{j,i}$, which is the weight the attention model learns from the i th word in the process of generating the j th subregion of the generated scene.

$$s_{j,i}^\sim = h_j^T e_i^\sim \quad (3)$$

$$C_j = \sum_{i=0}^{T-1} \beta_{j,i} e_i^\sim, \text{ where } \beta_{j,i} = \frac{\exp(s_{j,i}^\sim)}{\sum_{k=0}^{T-1} \exp(s_{j,k}^\sim)} \quad (4)$$

Imagination models S_i^{imag} are responsible for increasing the variety of the generated images. To elaborate the imagination models, the mechanism of LSTM cells is investigated first. The core concept of an LSTM cell is to add a cell state to enhance the decision in each step. It uses different gates to control the network. The first one is the forget gate, shown in Figure 2a. The inputs of the forget gate are the current input x_t and the previous output h_{t-1} . This gate uses a sigmoid function, denoted by σ , to control if it should forget or remember the information from the previous cell. The second gate is the input gate as shown in Figure 2b. The input gate determines how much new information should be updated to the new cell. The sigmoid function highlighted in Figure 2b determines the update value i_t . Additionally, the tanh function creates the new candidate value $\hat{C}_t$, which is to be added to the state value. When updating the old cell state C_{t-1} into the new cell state C_t , the old state is multiplied by the output of the forget gate f_t , and then added by $i_t \times \hat{C}_t$, as illustrated in Figure 2c. The output h_t is based on the cell state but being filtered. The old cell state C_{t-1} goes through a tanh function and is multiplied by the output of a sigmoid function. To summarize the mechanism, the forget gate decides what should be remembered. The input gate controls what information is relevant and should be added to the state in the current step. Additionally, the output gate decides what the next hidden state should be.

Figure 2. Cont.

Figure 2. LSTM Cells. (a) Forget Gate, (b) Input Gate, (c) Updating Cell State.

The imagination model utilizes the LSTM cell in the text encoder. In the upsampling processes of S_i , it utilizes the hidden features h_i to reinitialize the cell state in the text encoder. The main difference between the proposed framework and the original framework in [30] is that the imagination model extracts the features again from the text descriptions in every upsampling process. Such design enriches the text features given in the corpus description.

The loss function for the attentional and imaginative generative networks is defined in Equation (5).

$$\mathcal{L} = \mathcal{L}_G + \lambda_1 \mathcal{L}_C + \lambda_2 \mathcal{L}_{DAMS} \quad (5)$$

The first term $\mathcal{L}_G$ in Equation (5) is the summation of the loss at each stage i . We can express it as

$$\mathcal{L}_G = \sum_{i=0}^m \mathcal{L}_{G_i} \quad (6)$$

At the i th stage, the adversarial loss for the generator G_i is defined in Equation (7).

$$\mathcal{L}_{G_i} = -\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i))] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i, \bar{e}))] \quad (7)$$

The first part of Equation (7) is the unconditional loss, which discriminates if the image is real or fake. Additionally, the second part is the conditional loss, which determines whether the image and the text description match or not. In Equation (7), D_i denotes the corresponding discriminator for G_i . For each discriminator D_i , the cross-entropy loss defined in Equation (8) is minimized to classify the input as real or fake.

$$\begin{aligned} \mathcal{L}_{D_i} = & -\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log(D_i(x_i))] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{data_i}} [\log(1 - D_i(\hat{x}_i))] \\ & + -\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log(D_i(x_i, \bar{e}))] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{data_i}} [\log(1 - D_i(\hat{x}_i, \bar{e}))] \end{aligned} \quad (8)$$

Here, x_i is from the true image distribution p_{data_i} at the i th scale, and $\hat{x}_i$ is from the model distribution p_{G_i} at the i th scale.

The second term $\mathcal{L}_C$ in Equation (5) is designed to learn the relationship between the hidden state and image text in the imaginative model S_i^{imag} . The relationship r_h between each cell memory state c_h and word vector e_j is defined in the following equation.

$$r_h = c_h \times e_j^T \quad (9)$$

In this way, the relationship between the cell state initialized by the imaginative model S_i^{imag} and the j th word in the word vector can be learned. Then, the cosine similarity of r_h and the sentence vector s_h is calculated using Equation (10).

$$R(r_h, s_h) = r_h^T s_h / (\|r_h\| \|s_h\|) \quad (10)$$

For each minibatch, we calculate the matching score of the hidden state H and description D .

$$MS(H, D) = \log \left(\sum_{h=0}^{N_h} \exp(\gamma_{ms} R(r_h, s_h)) \right)^{\frac{1}{\gamma_{ms}}} \quad (11)$$

In (11), the term N_h is the number of cell memory states, and γ_{ms} is a parameter that determines the importance of the pair with the largest similarity $R(r_h, s_h)$ when calculating the matching score.

Suppose there are N_s samples in a minibatch. We calculate the matching score of the hidden state H_k and text description D_k of each sample k . Then, the probability of description D_k being matched with H_k is defined as Equation (12), where γ_p is an empirically determined parameter. Finally, the loss term $\mathcal{L}_c$ for the imaginative model is defined in Equation (13).

$$P(H_k|D_k) = \frac{\exp(\gamma_p R(H_k, D_k))}{\sum_{k=1}^{N_s} \exp(\gamma_p R(H_k, D_k))} \quad (12)$$

$$\mathcal{L}_c = - \sum_{i=1}^{N_s} \log P(H_i|D_i) \quad (13)$$

The third term $\mathcal{L}_{DAMSM}$ of the loss function in Equation (5) is a word-level image-to-text matching loss. The details of this loss can be found in [30]. Finally, λ_1 and λ_2 in Equation (5) are parameters to balance the three terms in the final loss function.

3. Experimental Results

The experimental results are presented in this section. To train the proposed framework, the training dataset is collected from Open Images Dataset V4 [34] and Places [35]. Open Images Dataset V4 contains 19,995 machine-labeled classes. Each label is provided with a confidence score. We use these labels to select the desired training images for the proposed system. Additionally, images with desired labels are selected from the Places dataset automatically. We select 26 different training descriptions for the Open Images Dataset V4, and 37 different training texts for the Places dataset in the experiments. We refer to the collected dataset set as the Scene dataset. In the Scene dataset, there are total 63 different descriptions. Each description contains 685 images. The total image count is 43,155. To test the proposed system, 1200 images are generated for each training description. Additionally, additional testing descriptions are established by combining different place descriptions randomly with some season or time descriptions in the training texts. There are a total of 76 additional testing description combinations. We also generate 1200 images for each additional testing description. Consequently, 166,800 images are generated to evaluate the image generation system in total. The experiment environment and settings are listed in Table 1.

Table 1. Experiment Environment and Settings.

Hardware	
CPU	i7-8700k
RAM	64GB
GPU	nvidia 1080ti
Environment	
OS	Ubuntu 16.04.5 LTS
Docker	18.06.1-ce
Python	python3.5.2
IDE	Vscode1.34.0

Table 1. Cont.

Tools	
pytorch	1.0.1post2
easydict	1.9
python-dateutil	2.7.2
pandas	0.22.0
torchfile	0.1.0
nltk	3.4
scikit-image	0.14.2
piexif	1.1.2

Two metrics, Inception Score (IS) [36] and Fréchet Inception Distance (FID) [37], are used to evaluate the generated images in the experiments. The Inception Score uses an Inception v3 network pretrained on ImageNet. It calculates the statistics of the outputs of the Inception v3 network when applied to generated images. The definition of IS is expressed in Equation (14), where $x \sim p_g$ denotes that x is an image sampled from the trained generative model.

$$IS(G) = \exp\left(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x) \| p(y))\right) \quad (14)$$

$$D_{KL}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (15)$$

$D_{KL}(P \| Q)$ is the KL divergence between the distributions p and q . In Equation (14), $p(y|x)$ is the conditional class distribution, and $p(y)$ is the marginal class distribution. Higher inception scores indicate that the generated images have larger variety, and each image can be distinctly classified as some category. Therefore, the higher the inception score, the better the generative model is.

As indicated in [36], there are some drawbacks to use inception score alone. Fréchet Inception Distance is proposed to remedy the issues of IS. To calculate FID, the inception network is used to extract features from an intermediate layer. The definition of FID between real data $\mathbb{P}_r$ and generated data $\mathbb{P}_g$ is expressed in Equation (16).

$$FID(\mathbb{P}_r, \mathbb{P}_g) = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(C_r + C_g - 2(C_r C_g)^{\frac{1}{2}}\right) \quad (16)$$

In Equation (16), μ_r and μ_g denote the mean of the real data and generated data, respectively. Additionally, C_r and C_g denote the covariance matrices of the real data and generated data, respectively. Finally, Tr sums up all the diagonal elements. Lower FID values indicate better quality and diversity for generated images. FID is more robust to noise than IS, and it is also a better measurement for image diversity.

Figure 3 shows the inception scores of the proposed method and AttnGAN [30] using a different number of training epochs. The number of epochs ranges from 10 to 65 in the training process. Figure 4 compares the FID values of the proposed method and AttnGAN using a different number of training epochs. We can observe that IS and FID do not improve much after number of epochs is larger than 50. Therefore, we select 50 as the number of epochs in the rest of the experiments to save training time. From Figures 3 and 4, we can see that the proposed method can achieve lower FID while maintaining similar or slightly higher IS compared to AttnGAN. Figure 5a displays scenery images generated using training descriptions. Images in Figure 5a are generated using the description “forest in the morning”. Figure 5b–d display selected images generated using testing description combinations. Images in Figure 5b are generated using the description “hayfield in autumn morning.” Images in Figure 5c are generated using the description “beach in winter morning,” and those in Figure 5d are generated using the description “sand in autumn.” We can observe that, although images generated under the training

descriptions look more realistic, images generated under testing description combinations still exhibit decent quality. This indicates that the model has learned from the word vectors to generate season, time and place combinations that it has never seen in the training images. The limitation of this work is similar to existing GAN-based methods. The model can generate good results if the input is mapped into the learned subspace. However, data that is unseen would be mapped out of the subspace. For example, since a baseball field is not in the training dataset, text descriptions containing baseball fields could not be generated properly.

Figure 3. Comparisons on Inception Score.

Figure 4. Comparisons on Fréchet Inception Distance.

Figure 5. Images Generated by Different Description Combinations Using the Proposed Method. (a) “forest in the morning”, (b) “hayfield in autumn morning”, (c) “beach in winter morning”, (d) “sand in autumn”.

Figure 6 compares some images generated by the proposed method and AttnGAN. We use the same description to generate 12 images at a time using both methods to observe the differences. Figure 6a,b shows images generated under the description “canal at night.” Figure 6c,d displays images generated under the description “canal in autumn.” Images in Figure 6a,c are generated by AttnGAN. Additionally, images in Figure 6b,d are generated by the proposed method. We can see that similar images appear repeatedly in Figure 6a,c, and the images in Figure 6b,d exhibit higher diversity. From Figure 6a, we can see that the results have similar colors. Additionally, the buildings along the canal are very much alike. We can observe that the images in Figure 6b contain different types of buildings, terrains, and reflection patterns. It is clear that the proposed imagination model is able to generate images with larger variety. Although the generated image in the top right corner of Figure 6b is an unsuccessful case, such failed examples also happen in Figure 6c (image on row 3, column 2 and column 3) generated by AttnGAN. Both methods inevitably generate unsatisfying results occasionally.

Figure 6. Images Generated by Different Methods. (a) “canal at night” by AttnGAN, (b) “canal at night” by the proposed method, (c) “canal in autumn” by AttnGAN, (d) “canal in autumn” by the proposed method.

For subjective qualitative evaluation, twenty users are invited to evaluate the results generated by the proposed method and AttnGAN. Each user is asked to evaluate the generated images using 30 different descriptions. For each description, twelve images are generated at a time so that the user can evaluate the variety of the generated images under the same description. Each user performs evaluation by choosing which result he or she prefers in terms of quality and variety. Among all user scores, 53% of them prefer the proposed method in terms of image quality. Additionally, 79% of the user scores prefer the proposed method in terms of variety. This subjective evaluation shows that the proposed method has both superior quality and more variety.

One possible application for the proposed framework is data augmentation for autonomous vehicles. Realistic and high-quality road scenes can be generated through the proposed system. For example, phrases such as “forest road in the morning”, “field road at night”, or “mountain road in winter” can be used to specify the kind of road scenes to be generated. With the help of additional image to video generation techniques proposed in [38–40], training and testing videos can be generated. This could save a lot of time and human power spent on driving and recording datasets under different weather conditions, in different seasons, and on different types of roads. Figure 7 displays the generated road images using the proposed framework. The descriptions used to generate the images from the first row to the last row in Figure 7 are “forest road in the morning”, “forest road in autumn”, and “forest road in winter.” The proposed system successfully gives the generated images desired characteristics described by the input texts. Additionally, the generated images exhibit sufficient variety to provide useful data augmentation for

training so that the autonomous vehicles can adapt to different road conditions. From Figures 6 and 7, we can see that the proposed framework is not limited to generating road images. It is also feasible to extend the framework to generate different types of scenes for miscellaneous robots or unmanned vehicles that can travel in different terrains or surfaces.

Figure 7. Different Types of Road Images Generated by the Proposed Method. (a) “forest road in the morning”, (b) “forest road in autumn”, (c) “forest road in winter”.

4. Conclusions

This paper proposes a framework to generate natural scenery images according to user descriptions. With text-to-image generation techniques, the proposed system can generate images under specified places, times and seasons. The framework improves and modifies the architecture of a generative adversarial network with attention models by adding the imagination models. The proposed imagination model utilizes the LSTM cell in the text encoder. It extracts the features again from the text descriptions in every up-sampling process. Such design enriches the text features given in the corpus description. We construct a Scene dataset which includes different types of scenery images for experimental purposes. The experiments show that the proposed framework can generate realistic and high-quality image scenes with sufficient diversity. The IS and FID scores are superior compared with the original AttnGAN. For future works, the work can be improved to generate more detailed descriptions specifying objects in the scenes. The work can also be extended to generate more general scene images in addition to natural sceneries. Moreover, the work can be combined with motion-generation techniques to generate videos. We can expect to find possible applications in data augmentation for autonomous vehicles by generating driving-training videos on desired type of roads and weather conditions with the help of the proposed framework.

Author Contributions: Conceptualization, H.-Y.C.; methodology, H.-Y.C.; validation, C.-C.Y.; formal analysis, H.-Y.C.; investigation, H.-Y.C.; resources, C.-C.Y.; data curation, H.-Y.C.; writing—original draft preparation, H.-Y.C.; writing, C.-C.Y.; supervision, H.-Y.C.; project administration, H.-Y.C.; funding acquisition, H.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology, grant number 110-2221-E-008-074-MY2.

Institutional Review Board Statement: Not applicable.

Acknowledgments: This research was funded by Ministry of Science and Technology, Taiwan.

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

Appendix A

Figure A1. The original architecture of the AttnGAN. Source of the figure: [30].

References

1. Ian, J.G.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*; MIT Press: Cambridge, MA, USA, 2014; pp. 2672–2680.
2. Pan, Z.; Yu, W.; Yi, X.; Khan, A.; Yuan, F.; Zheng, Y. Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access* **2019**, *7*, 36322–36333. [CrossRef]
3. Radford, A.; Metz, L. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico, 2–4 May 2016; Volume 1511, p. 06434.
4. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784. Available online: <https://arxiv.org/abs/1411.1784> (accessed on 2 March 2021).
5. Chen, X.; Duan, Y.; Houthoofd, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv* **2016**, arXiv:1606.03657.
6. Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 2642–2651.
7. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 214–223.
8. Metz, L.; Poole, B.; Pfau, D.; Sohl-Dickstein, J. Unrolled generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, Toulon, France, 24–26 April 2017; pp. 1–25.
9. Nguyen, T.D.; Le, T.; Vu, H.; Phung, D. Dual Discriminator Generative Adversarial Nets. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 4–9 December 2017; pp. 2672–2680.
10. Chen, Q.; Koltun, V. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22–29 October 2017; pp. 1520–1529.
11. Ozkan, S.; Ozkan, A. Kinshipgan: Synthesizing of Kinship Faces from Family Photos by Regularizing a Deep Face Network. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 7–10 October 2018; pp. 2142–2146.
12. Chen, W.; Hays, J. SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9416–9425.

13. Chen, Y.S.; Wang, Y.C.; Kao, M.H.; Chuang, Y.Y. Deep Photo Enhancer: Unpaired Learning for Image Enhancement from Photographs with GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6306–6314.
14. Zhang, Z.; Xie, Y.; Yang, L. Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6199–6208.
15. Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.J.; Wierstra, D. DRAW: A Recurrent Neural Network for Image Generation. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1462–1471.
16. Jin, Y.; Zhang, J.; Li, M.; Tian, Y.; Zhu, H. Towards the high-quality anime characters generation with generative adversarial network. In Proceedings of the Machine Learning for Creativity and Design, NIPS Workshop, Long Beach, CA, USA, 8 December 2017; pp. 1–13.
17. van den Oord, A.; Kalchbrenner, N.; Vinyals, O.; Espeholt, L.; Graves, A.; Kavukcuoglu, K. Conditional Image Generation with PixelCNN Decoders. In Proceedings of the Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4797–4805.
18. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
19. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
20. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 107. [\[CrossRef\]](#)
21. Aggarwal, A.; Mittal, M.; Battineni, G. Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 10004.
22. Yu, Y.; Huang, Z.; Li, F.; Zhang, H.; Le, X. Point Encoder GAN: A deep learning model for 3D point cloud inpainting. *Neurocomputing* **2020**, *384*, 192–199. [\[CrossRef\]](#)
23. Shin, H.; Tenenholtz, N.A.; Tenenholtz, A.; Rogers, J.; Schwarz, C.; Senjem, M.; Gunter, J.; Andriole, K.; Michalski, M. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging, Granada, Spain, 16 September 2018; pp. 1–11.
24. Tian, Y.; Pei, K.; Jana, S.; Ray, B. DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars. In Proceedings of the International Conference on Software Engineering 2018, Gothenburg, Sweden, 27 May–3 June 2018; pp. 303–314.
25. Fang, H.; Gupta, S.; Iandola, F.N.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
26. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative Adversarial Text to Image Synthesis. In Proceedings of the International Conference on Machine Learning, New York City, NY, USA, 19–24 June 2016; pp. 1060–1069.
27. Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L. Semantic compositional networks for visual captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 1141–1150.
28. Reed, S.; Akata, Z.; Schiele, B.; Lee, H. Learning Deep Representations of Fine-grained Visual Descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 49–58.
29. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5908–5916.
30. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1316–1324.
31. Zhu, Y.; Elhoseiny, M.; Liu, B.; Peng, X.; Elgammal, A. A generative adversarial approach for zero-shot learning from noisy texts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1004–1013.
32. Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Montréal, QC, Canada, 15–18 May 2017; pp. 2852–2858.
33. Schuster, M.; Paliwal, K.K. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [\[CrossRef\]](#)
34. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Mallocci, M.; Duerig, T.; et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [\[CrossRef\]](#)
35. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. A 10 million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Salimans, T.; Goodfellow, I.J.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. *Neural Inf. Process. Syst.* **2016**, *29*, 2234–2242.
37. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Proceedings of the International Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 6629–6640.

-
38. Ying, G.; Zou, Y.; Wan, L.; Hu, Y.; Feng, J. Better guider predicts future better: Difference guided generative adversarial networks. In *Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
 39. Endo, Y.; Kanamori, Y.; Kuriyama, S. Animating landscape: Self-supervised learning of decoupled motion and appearance for single-image video synthesis. *ACM Trans. Graph.* **2019**, *38*, 1–19. [[CrossRef](#)]
 40. Kwon, Y.; Park, M.G. Predicting future frames using retrospective cycle GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; pp. 1811–1820.