International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 02, ISSN No. 2455-2143, Pages 01-08
Published Online June 2023 in IJEAST (http://www.ijeast.com)

CLOTHING FASHION IMAGE GENERATION FROM TEXT USING ARTIFICIAL INTELLIGENCE

Aqsa Shaheen
Department of Computer Science
University of Engineering and Technology Taxila, Pakistan

Dr. Javed Iqbal
Assistant Professor
Department of Computer Science, University of Engineering and
Technology, Taxila, Pakistan

Abstract. **Development of dynamic, intensely engaging, and fascinating images has greatly benefited from the recent exponential advancements in image synthesis techniques. The architecture proposed in this research allows users to enter text regarding a particular dress, and the model then create images of fashionable apparel based on that content. The model suggested can let people become their own fashion designers by utilizing the strength of Deep Learning and Artificial intelligence to create a variety of fashionable outfits for themselves. DALL-E model is utilized to engender realistic images based on text description. DALL-E is an artificial intelligence model that generates realistic images from a description in natural language. While there are alternative text-to-image systems, the DALL-E produces far more coherent visuals. The world and the relationships between objects appear to be well understood by this technology. DALL-E uses GPT-3 model and dataset of text-image pairs for image synthesis. Image is encoded into size of 32×32 grid using VQ-VAE. Then image and text are combined together in the form of single stream for training of DALL-E. Deep Fashion dataset is used for training of DALL-E, which is simply more realistic dataset and contains High definition images that further enable accurate generation. After training DALL-E produce more accurate results and provides higher inception score than preceding models.**

Keywords: **DALL-E, VQ-VAE, AI, Synthesis, Encoding, GAN, Fashion, Machine Learning**

I. INTRODUCTION

Nowadays, fashion is all around us from products like footwear, clothing, and makeup, accessories to hair styles, interior and lifestyle. Fashion industry is growing rapidly and large part of the population is directly or indirectly associated with fashion. Machine learning and artificial intelligence is now implemented in several domains and while promoting numerous applications. In fashion it has become very appealing and influential while providing solutions in every part. Penetration of AI and machine learning in fashion industry overcomes the limitation of designers' inventive abilities to create novel and fresh designs. Machine learning is widely used in multiple fields of fashion such as fashion design classification, style prediction, style recommendation, and sales forecasting. However, image generation of fresh and novel fashion items is not much explored such as generating novel images of shirts, T-shirts, trousers, jeans, footwear and bags. When we read or hear a story, we visualize the content in our head in the form of pictures. There is a natural relationship between human language and visual world that we rarely think about it. Building a system which can understand the relationship between language and vision, and can create images based on text description, is a major advancement in the history of machine learning and artificial intelligence. Image synthesis is a field of deep learning which is a process of generating new images or manipulating existing. Image synthesis is being used in multiple applications such as image editing, art generation, video games, visual reality and computer aided designs [1]. Let's imagine a designer is designing a new outfit. She wants to design a white frock with floral patterns and short sleeves to wear in summer. She is in need of reference images based on her imagination or thinking. AI and machine learning can help in this regard providing solution in the form of generative models GAN [1] and Auto encoders [2]. Generative Adversarial Networks GAN [3]is seeking much attention due to its adaption in multiple applications such as human face synthesis [4], style transfer [5], image in painting [6] [7], image to image translation [8] [9],and data augmentation [10]. Early efforts on Auto encoder were used for dimensionality reduction or feature learning and were first proposed by LeCun in 1987. Auto encoder has recently gained prominence in generative modelling due to the popularity of deep learning research. Auto encoders come in a variety of forms that have been effectively used in a wide range of applications, including computer vision, speech recognition, and natural language processing. The input layer, the hidden layer, and the output layer are the three successive layers that make up a common neural network called an Auto encoder (AE). An unsupervised neural network called an Auto encoder compresses data from a multidimensional format to a chosen dimensionality. Utilizing 2 | P a g e the hidden layer weights generated by encoding, it reconstructs the input data [2]. Generative Adversarial Networks had the aptitude to synthesize real world images such as birds, flowers, faces, photo albums, dogs, home interior etc. All proposed GAN models had the capability to generate real images based on captions or class labels, but there was no control over image or object location and pose. Scott Reed et al. provided the solution in Generative Adversarial what where network (GAWWN) [11]. CUB dataset was used and images was conditioned based on object location and informal text description. Location was controlled accurately by bounding box and 128×128 resolution bird images were generated [11]. Fashion image synthesis is a difficult task as it involves multiple fashion items to generate a compatible final look. Diverse set of fashion image synthesis is needed for satisfying Virtual try-on experience and fashion items compatibility [13]. To overcome the problem FiNet (Fashion inpainting Networks) [12] provided fashion transfer and clothing reconstruction which could be useful for new fashion recommendation and compatibility-aware fashion design.

1.1 Uses

Uses of Text to image synthesis using AI are

- The model has been designed to accept ideas from the user's mind as a written description and convert the text into an image by creating realistic images of clothing.
- The system is able to help users create their own fashion clothes without the need of a professional designer.
- Everyone can become their own fashion designer using the provided template that turns their thoughts and visualizations into images and displays them in real time.

II. LITERATURE REVIEW

Text to image synthesis has emerged as one of the prominent research domains. The application of image synthesis in computer aided design (CAD) has made it one of prominent research domains. GANs have been applied in various fields such as: image synthesis, image translation, image-attribute editing, domain Adaptec's, etc. [14] [15] [16] [17] Particularly, GANs have achieved state-of-art results in the image synthesis. Other than GANs another neural network auto-encoders has also been used in many image generation tasks. Generating images based on text description is an active research problem and has gained significance in computer vision. Text to image synthesis is a task of generating photo-realistic images from user's text description.

2.1 Auto encoders based image synthesis

We found that conventional Auto encoders are capable of learning to compress and reconstruct data but are not particularly useful for creating new data. Variational Encoder (VAE) proved useful in this situation. Instead of only learning the compressed image, VAE learns the distribution of the data, and by exploiting the distribution, we can decode and produce new data. VAEs (Variational Auto-encoders) have also been highly successful, to the point where they are frequently mathematically more accurate at producing images that closely resemble their original dataset [9].

2.2 GAN Based Image Synthesis

GAN comprises two different neural networks: a generator and a discriminator. The purpose of the generator is to add random noise and generate the synthetic data G (z). The synthetic data is then passed to the discriminator as shown in Figure 1. The discriminator takes either true data or G (z). Its purpose is to give the probability whether the data is synthetic or true data [18].

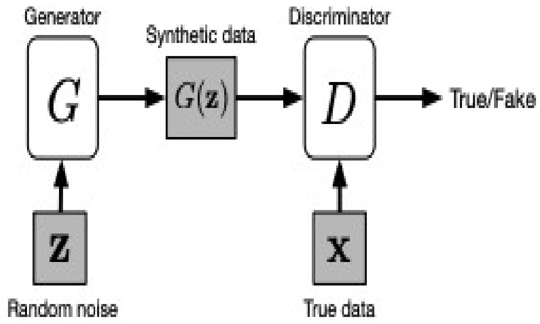


Fig. 1 General Structure of GAN [19]

Image generative models have gained significant attention over the period of time. Image synthesis from text to image has a significant amount of application in CAD and computer vision. Though a lot of work has been done in the image domain, text to image-based synthesis was not solved. Since the evolution of GAN, a lot of work has been done for text to image-based generation. Attention GAN generated a very detailed region over a multi-stage refinement process. In contrast to the previous techniques, it resulted in better quality and performance. Attention GAN consists of consecutive attentional GAN's, this setting aids the generation of multi-scale images [20]. Despite the improvement Attention GAN brought in text to image synthesis domain, it was not tested on Fashion dataset [21].

GAN was firstly used by Reed et. al [22] to generate photorealistic images from detailed visual concepts in text. They were able to generate low resolution images (64^2) using CUB [23] dataset. Character to pixel synthesis was limited to birds and flowers images. In [24] authors proposed Generative Adversarial What-Where Network (GAWWN) on Caltech-UCSD Birds dataset with resolution of (128^2). Bounding box around the bird and its respected parts were also controlled. Hong et. al in [24] generated photo-realistic images from text using semantic layout. Proposed algorithm decomposed generation process into multiple steps and instead of direct mapping from text, a layout generator was utilized which generates semantic layout from text. Semantic layout was then converted to image by image generator. Model is demonstrated on MS-COCO [23] dataset and complicated images were generated. Spatially adaptive Normalization [24] performed on semantic layout generated diverse realistic images comprising of indoor, landscape, outdoor, and street scenes.

Jain et al. proposed a model to take textual information regarding user's query about fashion and to synthesize images from given textual input. The proposed framework used StackGAN [25] with two stages and has been tested on Fashion dataset [17]. The aim of the first stage is to synthesize low resolution images. In the second stage, by high resolution images with more realistic and fine-grained information in accordance with the user input. The proposed model has currently been exploited over the cloud. Particularly with the focus on data set enhanced Attention GAN (e-AttnGAN) has been proposed in [29]. e-AttnGAN uses Feature-wise Linear Modulation (FiLM) which utilizes sentences and words. FiLm adds the manipulative ability for visuals with no additional support. It has been tested on Deep Fashion [30] and FashionGen [31]. Both these datasets have text description corresponding to the images.

2.3 DALL-E

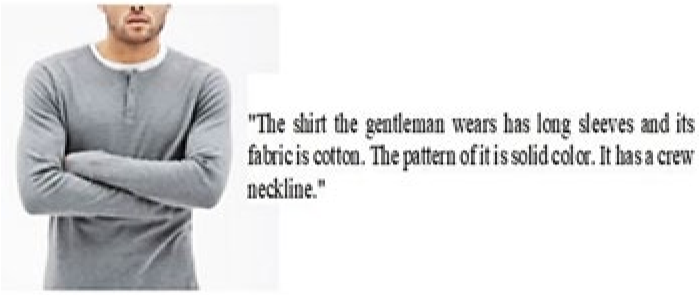
DALL E [35] has been trained to generate text from images using text-image pair. It has diverse set of capabilities and could generate images of objects, landscape, animals and others. The model takes image and text as describing that image as one stream to synthesize new related images based on given text. It consists of two stages in stage one dVAE discrete vibrational auto encoder to compress 256×256 RGB image into 32×32 RGB image and size of transformer is reduced to 192 where image quality is not degraded. Then in second stage text is concatenated to image and trained over text image tokens. Zero-shot text to image is trained on JFT-300M 250 million text-image pairs dataset collected from internet. Model provide better result on MS-COCO and cub dataset without use of training labels. Text to image generation and image to image translation is provided in one module [34].

2.4 Dataset

The dataset for text to picture synthesis utilizing generative models and convolutional neural nets must be consistent with the GAN model. To create images from text input, a model must learn about various keywords and link them to the visual semantics they denote. As a result, the necessary dataset must include photographs of various types of fashionable apparel as well as the attributes of each item of clothing that is associated with the image. Only a small number of datasets are available for text to image synthesis. The dataset used is DeepFashion [30] Multimodal, consisting of 78,979 images and their respective human annotated captions.

TABLE 1 DEEP FASHION DATASET STATISTICS [30]

Dataset	Deep Fashion
Total Images for synthesis	78,979
Categories	50
Attributes	1000



"The shirt the gentleman wears has long sleeves and its fabric is cotton. The pattern of it is solid color. It has a crew neckline."

Fig. 2 Example Images from dataset [36]

"The gentleman wears a sleeveless tank shirt with solid color patterns. The tank shirt is with cotton fabric."

Fig. 3 Image from dataset [36]

III. PROPOSED METHODOLOGY

DALL-E [35] module is used for training on fashion dataset and generating images from text. DALL-E module is explained below.

3.1 DALL-E Working:

Using a collection of text-image pairs, the DALLE model creates images from text descriptions. DALL-E is a transformer language model, just like GPT-3. GPT-3 demonstrated that a large neural network can be trained to carry out a range of text generation tasks using language. The Image GPT demonstrated that the same kind of neural network may also be used to generate the high-quality images.

DALL-E is a transformer language model. It accepts the text and the image as a single stream of data with up to 1280 tokens, and it is trained to produce each token one at a time using maximum likelihood. Any symbol from discrete vocabulary is referred to as a token, and each letter of the 26-letter English alphabet is a token for humans. Both text and image concept tokens are included in the DALL-E vocabulary. A maximum of 256 BPE encoded tokens with a vocabulary size of 16384 are used to represent each image caption, while 1024 tokens with a vocabulary size of 8192 are used to represent the actual image. The methodology shown in Figure 4 illustrates various phases that are used to carry out the research work. The proposed system is implemented to generate real images using textual input.

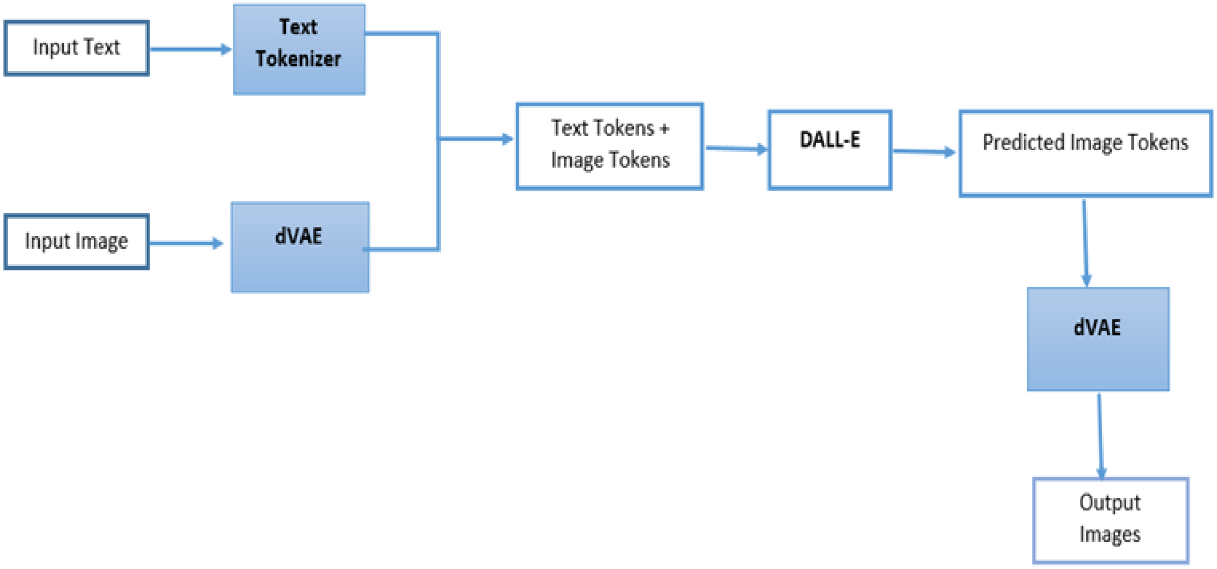


Fig. 4 DALL-E Flow Diagram

3.2 Encoding and decoding:

Dimensionality reduction in machine learning is the process of lowering the number of characteristics used to represent a given set of data. Many applications that call for low dimensional data (data visualization, data storage, intensive computation, etc.) can benefit from this reduction, which is carried out either by selection (only some existing features are conserved) or by extraction (a smaller number of new features are created based on the old features). Encoders create "new features" representations from "old features" representations (either by selection or extraction), and decoders do the opposite. Therefore, dimension reduction can be understood as data compression, with the encoder compressing the data (from the initial space to the encoded space, also known as latent space), and the decoder uncompressing it.

3.3 VAE:

An architecture made up of both an encoder and a decoder that is trained to minimize the reconstruction error between the encoded-decoded data and the starting data is known as a Variational Auto encoder (VAE) [9]. Instead of encoding an input as a single point, we encode it as a distribution over the latent space in order to introduce some regularization of the latent space. This modifies the encoding-decoding process slightly. After that, the model is trained as follows.

- The input is first encoded as a distribution across the latent space
- Then a point from that latent space is sampled from the distribution
- Which is then decoded to allow the computation of the reconstruction error
- Finally, the reconstruction error is back propagated through the network.

3.4 VQ-VAE:

It is easier to comprehend the VQ-VAE [34] model as a communication system. It is made up of a decoder, which reconstructs the observations from these discrete variables, and an encoder, which maps observations into a series of discrete latent variables. A shared codebook is used by the encoder and decoder.

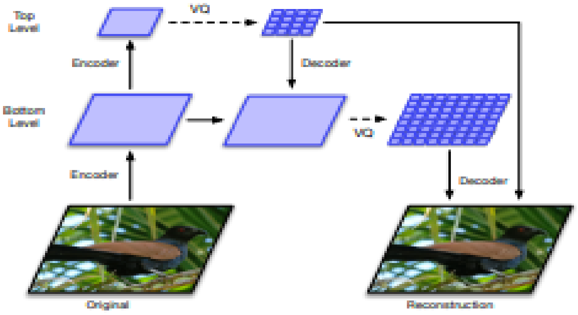


Fig. 5 VQ-VAE [34]

Deep neural networks are used in both the encoders and decoders. A 256 x 256 image is compressed to quantized latent maps with sizes of 64 x 64 for the bottom level and 32 x 32 for the top level serves as the model's input. From the two latent maps, the decoder creates the image.

3.5 Tokenization:

Any symbol from a limited lexicon is referred to as a token, and each letter of the 26-letter English alphabet is a token for humans. Both text and image concept tokens are included in the DALLE vocabulary. The image itself is represented using 1024 tokens with an 8192-word vocabulary, and each image caption is expressed using a maximum of 256-byte pair-encoded tokens. During training, the photos are preprocessed to 256x256 resolution. Similar to VQ-VAE, we pre-train a discrete VAE utilizing a continuous relaxation to compress each image to a 32x32 grid of discrete latent codes. DALL-E [35] is a straightforward decoder-only transformer that models each of the 1280 tokens it gets 256 for the text and 1024 for the picture auto regressively. It receives both the text and the image as a single stream. Each picture token is able to attend to every text token thanks to the attention mask at each of its 64 self-attention layers. For the text tokens, DALL-E use the conventional causal mask, and for the image tokens, it employs sparse attention with either a row, column, or convolutional attention pattern, depending on the layer.

3.6 DALL-E Training:

1. Training for encoder and decoder image into 32x32 grid of 8k potential codes word tokens (d-VAE)
2. Combine picture and text tokens into a single array
3. Anticipate the next image token based on the previous tokens (autoregressive transformer).
4. Only the image decoder and the next predicator is kept; image encoder is discarded.

3.7 G. Comparison with Existing Models: Inception Score is used to evaluate the performance of the model. A popular statistic for measuring synthesis model performance is Inception Score (IS). The evaluation is based on a pre-trained Inception V3 model.

Model	**Inception Score**
StackGAN++ [26]	1.74±0.02
AttnGAN [28]	4.12±0.06
e-AttnGAN [29]	4.77±0.10
DALL-E (Proposed Work)	5.6±0.40

Table 2 Comparison with Existing Models

Below are examples of output generated for given text. Model is not learned to create faces so faces are not generated accurately.

The shirt the female wears has long sleeves and its fabric is knitting. The pattern of it is pure color.
This person is wearing a ring on her finger.

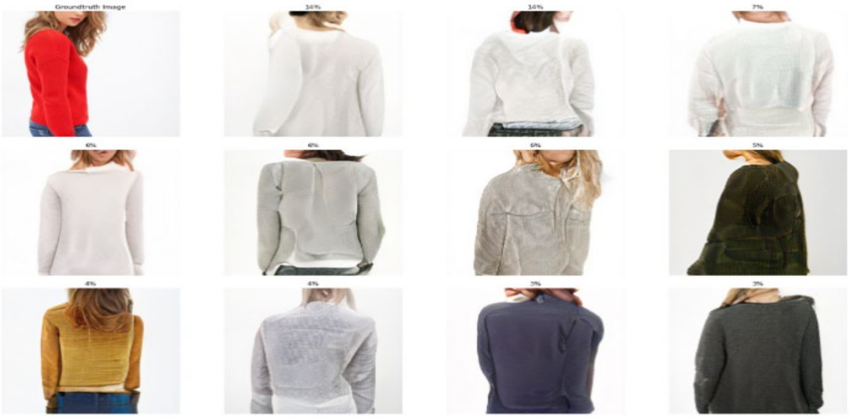


Fig. 6 Output Images with given text description

The person is wearing a medium-sleeve shirt with lattice patterns and a long pants.
The shirt is with cotton fabric and its leckline is lapel. The pants are with denim fabric and solid color patterns.
This female is wearing a ring on her finger

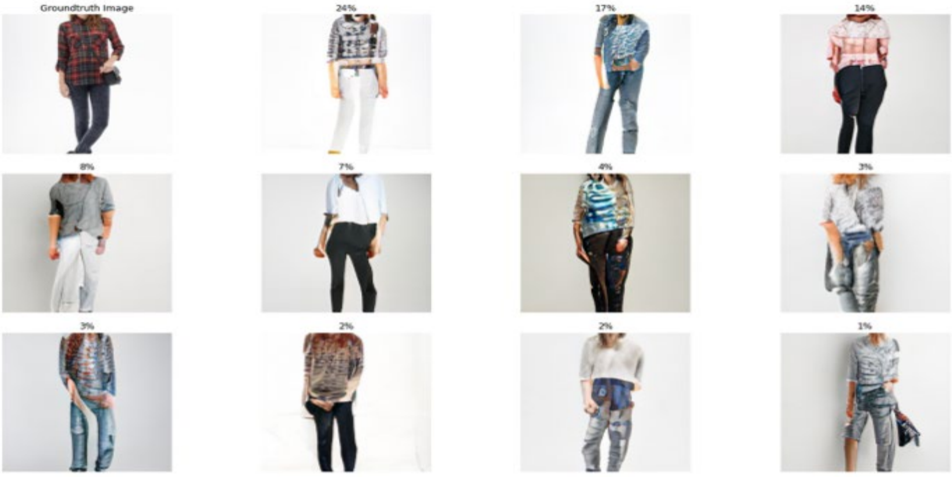


Fig. 7 Output Images

IV. CONCLUSION

The model is intended to take user input in the form of written descriptions and translate that information into visual representations by creating photos of apparel that are as accurate as possible. Without the aid of a professional designer, the system helps users create their own fashionable apparel. By utilizing the suggested approach, individuals can become their own fashion designers by having their ideas and visualizations translated into an image in real time. DALL-E is a new AI model used to for image synthesis. DALL-E model is trained on Deep Fashion dataset. The acquired results are promising as inception score calculated is more than existing clothing images synthesis models. Text-to-image generating technology converts ideas into on-screen visual feedback in real time. Everyone can become their own designer through the transformation of thoughts into images. People would be able to visualize the outfit in reality by simply describing it in their minds.

V. FUTURE WORK AND LIMITATIONS:

The suggested strategy is successfully put into practice, and the results of the text input appear promising. The suggested methodology has limitations in that the results are restricted to the dataset that was used; for example, because color training was not included in the dataset, the model was not trained on colors. The dataset needs to be updated frequently in order to increase the variety of results and the searches the user can enter. Model usability can be increased by creating web interface or application for model deployment. Additionally, the output of the suggested model could not accurately reflect pictures of clothes in the real world. The approach can be expanded to create wearable clothes given the user's body dimensions when this product is combined with the garment production business. Thus, a method for producing individualized and tailored apparel on demand can likely be created in the future. Work is still being done on text to image synthesis to give real world images.

VI. REFERENCES

[1]. T. H. F. R. H. D. Stanislav Frolov, "Adversarial text-to-image synthesis: A review," Neural Networks, vol. 144, pp. 187-209, December 2021.

[2]. J. Zhai, S. Zhang, J. Chen and Q. He, "Autoencoder and Its Various Variants," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 2018, pp. 415-419, doi: 10.1109/SMC.2018.00080

[3]. J. P.-A. M. M. B. X. D. W.-F. A. C. Y. B. Ian J. Goodfellow, "Generative adversarial nets," in NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems Volume 2, 2014.

[4]. T. A. T. L. S. &. L. J. Karras, "Progressive growing of GANs for improved quality, stability, and variation," International conference on learning representations. 2018.

[5]. L. A. E. A. S. &. B. M. Gatys, "Image style transfer using convolutional neural networks," Proceedings of the IEEE computer vision and pattern recognition, p. 2414–2423, 2016.

[6]. J. L. Z. Y. J. S. X. L. X. &. H. T. S. Yu, "Free-form image inpainting with gated convolution," Proceedings of the IEEE international conference on computer vision, p. 4471–4480, 2019.

[7]. R. A. C. C. L. T.-Y. S. A. G. H.-J. M. &. D. M. N. Yeh, "Semantic image inpainting with deep generative models," Proceedings of the IEEE computer vision and pattern recognition, p. 5485–5493, 2016.

[8]. P. Z. J.-Y. Z. T. &. E. A. A. Isola, "Image-to-image translation with conditional adversarial networks," Proceedings of the IEEE computer vision and pattern recognition, p. 1125–1134, 2016.

[9]. M. Sami and I. Mobin, "A Comparative Study on Variational Autoencoders and Generative Adversarial Networks," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 1-5, doi: 10.1109/ICAIIT.2019.8834544.

[10]. D. I. K. E. A. M. G. J. G. H. Frid-Adar M., "GAN-Based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, pp. pp. 321-331, 2018.

[11]. Z. A. S. M. S. T. S. H. L. ScottReed, "Learning what and Where to Draw," Computer Vision and Pattern Recognition (cs.CV); Neural and Evolutionary Computing (cs.NE) arXiv: 1610.02454 [cs.CV], 2016.

[12]. T. W. V. D. K. A. B. S. A. B. A. Creswell, "Generative Adversarial Networks: An Overview." in Institute of Electrical and Electronics Engineers (IEEE) 53–65, 2018.

[13]. M. M. M. Z. F. Ben Aissa, "A survey on generative adversarial networks and their variants methods," in International Conference on Machine Vision (ICMV) 114333N, 2020.

[14]. J. Z. D. L. Y. Z. J. C. S. Zhang, "Recent Advance on Generative Adversarial Networks," in International Conference on Machine Learning and Cybernetics (ICMLC) 69–74, 2018.

[15]. U. H. J. Y. S. Y. Y. Hong, " How generative adversarial networks and their variants work," in ACM Computing Surveys 1–43., 2019.

[16]. H. Y. P. S. &. W. C. Huang, "An introduction to image synthesis with generative adversarial nets arXiv preprint arXiv: 1803.044 69," 2018.

[17]. M. D. J. R. C. S. Jain A, "Text to Image Generation of Fashion Clothing," in 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 355-358). IEEE. 2019 Mar 13 .

[18]. J. P.-A. M. M. B. X. D. W.-F. A. C. Y. B. Ian J. Goodfellow, "Generative adversarial nets," in NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems Volume 2, 2014.

[19]. H. L. B. S. L. L. X. Y. Z. A. Scott Reed, "Generative Adversarial Text to Image Synthesis," in Neural and Evolutionary Computing (cs.NE); Computer Vision and Pattern Recognition (cs.CV), arXiv:1605.05396, 2016.

[20]. C. W. e. al., "The caltech-ucsd birds-200-2011 dataset," 2011.

[21]. S. M. S. T. B. S. H. L. S. R. Z. Akata, " "Learning What and Where to Draw,"," in Computer Vision and Pattern Recognition (cs.CV) ,Neural and Evolutionary Computing (cs.NE), rXiv:1610.02454 , 2016.

[22]. S. Hong, D. Yang, J. Choi and H. Lee, "Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[23]. M. M. S. B. J. H. P. P. D. R. P. D. a. C. L. Z. T.-Y. Lin, "Microsoft coco: Common objects in context," in ECCV, 2014.

[24]. T. Park, M.-Y. Liu, T.-C. Wang and J.-Y. Zhu, "Semantic Image Synthesis With Spatially-Adaptive Normalization," in IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019

[25]. T. X. H. L. S. Z. X. W. X. H. a. D. M. H. Zhang, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in IEEE International Conference on Computer Vision (ICCV, Venice. Italy, 2017.

[26]. T. X. H. L. S. Z. X. W. X. H. D. M. Han Zhang, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," in Pattern Analysis and Machine Intelligence (TPAMI), IEEE, 2018.

[27]. P. Z. L. Z. Q. H. X. H. S. L. a. J. G. W. Li, "Object-Driven Text-To-Image Synthesis via Adversarial Training," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, USA, 2019.

[28]. T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang and X. He, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[29]. H. L. J. Y. T. J. K. A. Emir AK K, "Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network." in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[30]. Z. a. L. P. a. Q. S. a. W. X. a. T. X. Liu, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), june 2016.

[31]. S. H. T. B. W. S. Y. Z. C. J. a. C. P. N. Rostamzadeh, " Fashion-Gen: The Generative Fashion Dataset and Challenge. ArXiv e-prints," June 2018.

[32]. J. Z. D. L. Y. Z. J. C. S. Zhang, "Recent Advance on Generative Adversarial Networks," in International Conference on Machine Learning and Cybernetics (ICMLC) 69–74., 2018.

[33]. B. C. a. A. Kea, "Toward Realistic Image Compositing With Adversarial Learning," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8407-8416, 2019.

[34]. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever, "Zero-Shot Text-ToImage Generation", In Computer Vision And Pattern Recognition (Cs.Cv); Machine Learning (Cs.Lg),2021

[35]. DALL·E: Creating Images from Text (openai.com)

[36]. https://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html