StyleSphere: Conversational Fashion Outfit Generator powered by Generative AI

Sujit Akulwar[1] Q Rushikesh Bakshi[1] Q Vedant Dahale[1] Q Tanay Dubey[1] Q
Mr.Anandrao Deshmukh[2] Q

1 Student, Department of Computer Engineering, PES Modern College of Engineering
2 Assistant Professor, Department of Computer Engineering, PES Modern College of Engineering

1. Abstract

Artificial Intelligence (AI) has gained traction in the fashion domain, particularly with Generative AI technologies. Our study presents a unique text-to-image generator tailored for fashion, leveraging Generative Adversarial Networks (GANs), notably the StackGAN variant. This model translates textual fashion descriptions into visually appealing images, offering a stream-lined approach to design conception. Validation demonstrates its ability to produce photo-realistic fashion visuals, bridging the gap between concept and visualization. Our research con-tributes to advancing AI-driven solutions in fashion, facilitating streamlined design workflows and encouraging creativity. This integration of AI with fashion design practices has the potential to revolutionize the industry, offering new avenues for creative expression and addressing contemporary design challenges.

Keywords

Fashion Design, Generative Adversarial Networks (GANs), StackGAN, Image Synthesis

2. Introduction

The landscape of visual media content creation has undergone a rapid transformation with the ~~integration of artificial intelligence. When it~~ comes to fashion design, traditional methods such as computer-aided designing or hand-drawn sketches demand considerable time and numerous iterative refinements. Crafting fashionable attire entails a meticulous process of precision design and intricate detailing, requiring not only diligence but also creativity and attention to de-tail. However, such tasks can now be automated with the assistance of artificial intelligence. In to-day's context, Generative Adversarial Networks (GANs) have emerged as a powerful technique for image generation [1, 4, 19]. While training GANs on specific datasets poses challenges, numerous research studies have demonstrated their efficacy in image synthesis. Nonetheless, prior research primarily focused on generalized data rather than clothing imagery. In contrast, our proposal involves a version of StackGAN specifically trained on images of fashion garments.

StackGAN were originally proposed by Han Zhang et al. [2]. The name is so, because the network consists of Sequentially stacked layers of generator and discriminator. The process consists of two stages: In Stage-I, primitive shapes and background colors are generated based on the provided text descriptions. The stage-I generates low resolution images. Subsequently, in Stage-II, additional details are incorporated by conditioning on the text descriptions & Stage-I results providing detailed synthesized image through the network. Previously, when Ian Goodfellow first introduced GAN [3], the key barrier to generating high-resolution im-ages was the training instability of GAN [5], which can probably result in the non-overlapping of image distributions with model distributions. However, the low-resolution image produced in Stage-I of StackGAN has the better probability to align with the support of model distribution. As a result, StackGAN Stage-II can produce high-resolution photo-realistic images. In addition, the novel Conditioning Augmentation technique, introduced by Han Zhang et al. [2], has proven to be highly effective. This innovative approach has substantially improved the quality & diversity of images produced.

Accordingly, with reference to [2], we aim to build StackGAN model with two stages, Stage-I & Stage-II. Stage-I produces a low-level image of 64 x 64 pixels, whereas, Stage-II fine-tunes the image with multiple sequentially stacked upsam-

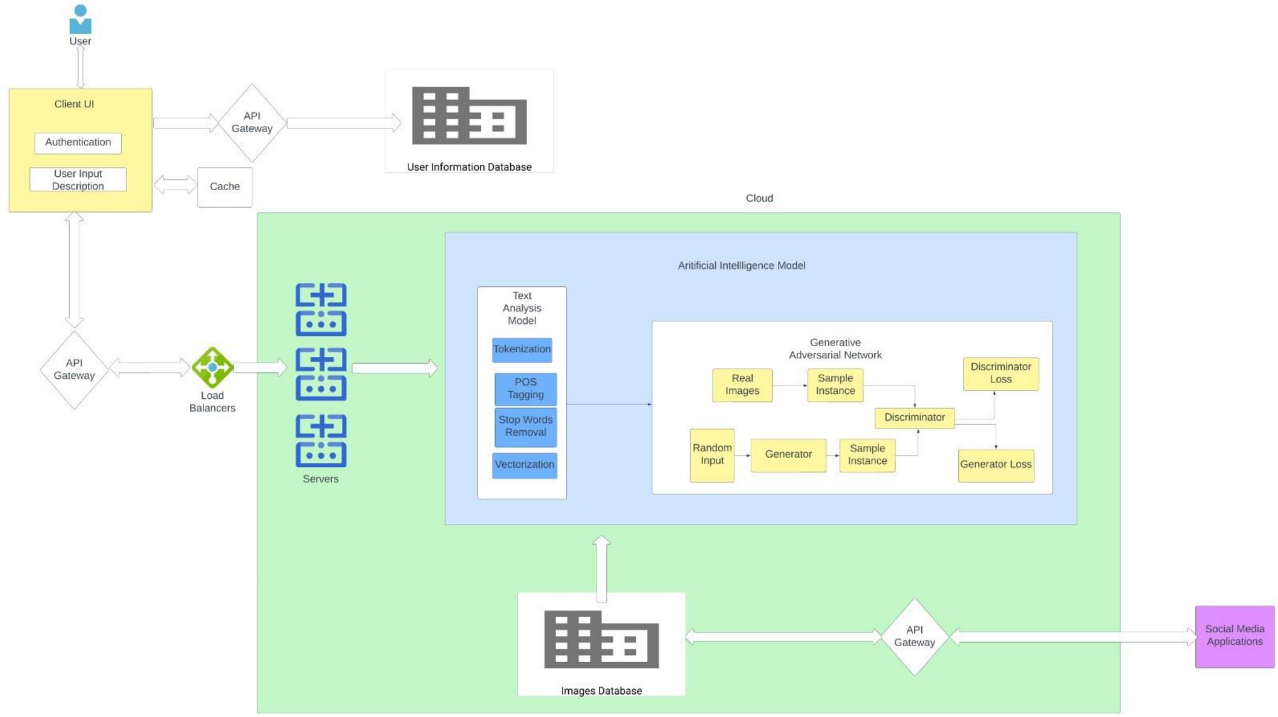


Figure 1: System architecture diagram

pling & down sampling layers.

In synopsis, we table our work with three important advancements 1) An innovative approach, utilizing Stacked Generative Adversarial Networks (GANs) to generate photorealistic im-ages from text descriptions, thus breaking down the challenge of high-resolution image synthesis and achieving significant progress. Utilizing StackGAN, images with 256x256 resolution and lifelike details are produced directly from text prompts. 2) Additionally, a novel Conditioning Augmentation [15] technique is introduced to stabilize conditional GAN training and enhance sample diversity. 3) Comprehensive analyses confirm the effectiveness of the model design and components, offering insights for future conditional GAN development.

3. Related Work

In recent years, the fashion industry has embraced generative methods for image creation [6, 7, 8], driven by advances in computer vision and machine learning. Leveraging techniques such as generative adversarial networks (GANs) and variational autoencoders (VAEs), these methods are revolutionizing fashion design and retail, enabling personalized clothing design, virtual try-on experiences, and innovative approaches to advertising and marketing. The earlier mentioned works on diffusion methods, like Isolated Diffusion by Jingyuan Zhu et al. [9], focus on addressing the issue of Concept Bleeding to improve the quality of image generation. Meanwhile, EDIT-VAL proposed by Samyadeep Basu et al. [10]. offers an image editing technique with conversational features. GAN-based methods such as AttnGAN by Tao Xu et al. [11], introduce attention mechanisms to generate images from text descriptions, enhancing the model's ability to focus on specific details and improve realism. Scott Reed et al. [12, 13], explore image generation based on the semantic context of text descriptions, resulting in detailed images.

Furthermore, within the fashion community, Fashion-AttGAN by Qing Ping et al. [14] intro-duces a novel approach for fashion image editing using Multi-Objective Generative Adversarial Networks (GANs). This model enables precise manipulation of fashion attributes like color, pat-tern, and style within images. Leveraging attention mechanisms, Fashion-AttGAN achieves realistic results by focusing on relevant areas during attribute editing. Proposed by the same authors, InGAN [15] facilitates virtual try-on experiences and fashion editing by seamlessly integrating clothing items onto individuals within existing images. The model employs advanced techniques in GANs and conditional image synthesis to deliver realistic and visually appealing outcomes.

While our work shares similarities with previously published methods like StackGAN, we aim

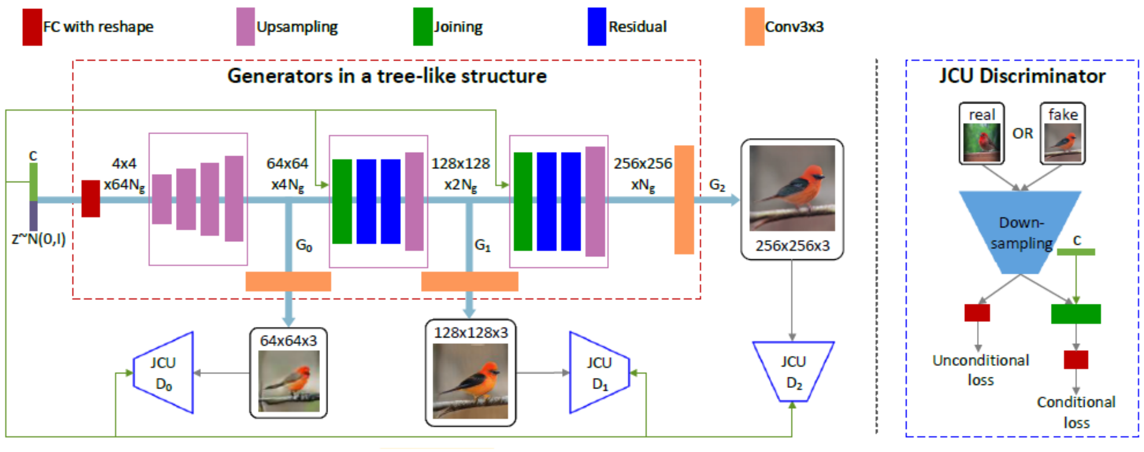


Figure 2: StackGAN-v2 architecture

to address domain-specific problem to the fashion industry.

4. Stacked Generative Adversarial Networks

StackGAN-v2 has a series of multi-scale image distributions. It consists of multiple generators (Gs) and discriminators (Ds) in a tree-like structure, where the image quality gradually increases from low-resolution to high-resolution from different branches of tree. At each branch, generator tries to capture the image distribution whereas the discriminator estimates whether the sample image came from training dataset or generated by the generator by calculating corresponding probabilities of those steps.All the generators are trained jointly so that they can approximate the multiple distributions, while the generators & the discriminators are trained in an alternative manner.

4.1. Multi-scale image distributions approximation

StackGAN-v2 can adapt the noise vector $z \sim p_{noise}$ because of its tree structure. The p_{noise} is taken as the standard normal distribution. The latent variable z are reshaped into hidden features layer by layer. The computation of hidden features hi for each Generator G_i is done by a non-linear transformation,

$$h_0 = F_0(z); \quad h_i = F_i(h_{i-1}, z), \quad i = 1, 2, \ldots, m - 1$$

where h_i represents hidden features for the i^{th} branch, m is the total number of branches, and F_i are modelled as neural networks. To capture the omitted information in previous branches, the noise vector z is concatenated to the hidden features h_{i-1} as the inputs of F_i for calculating h_i. Based on hidden features at different layers ($h_0, h_1, \ldots, h_{m-1}$), generators produce samples of small-to-large scales ($s_0, s_1, \ldots, s_{m-1}$),

$$s_i = G_i(h_i), \quad i = 0, 1, \ldots, m - 1$$

where G_i is the generator for the i^{th} branch. After each generator G_i, a discriminator D_i, takes real image x_i or a fake sample s_i as input and classify inputs into real or fake by minimizing the cross-entropy loss,

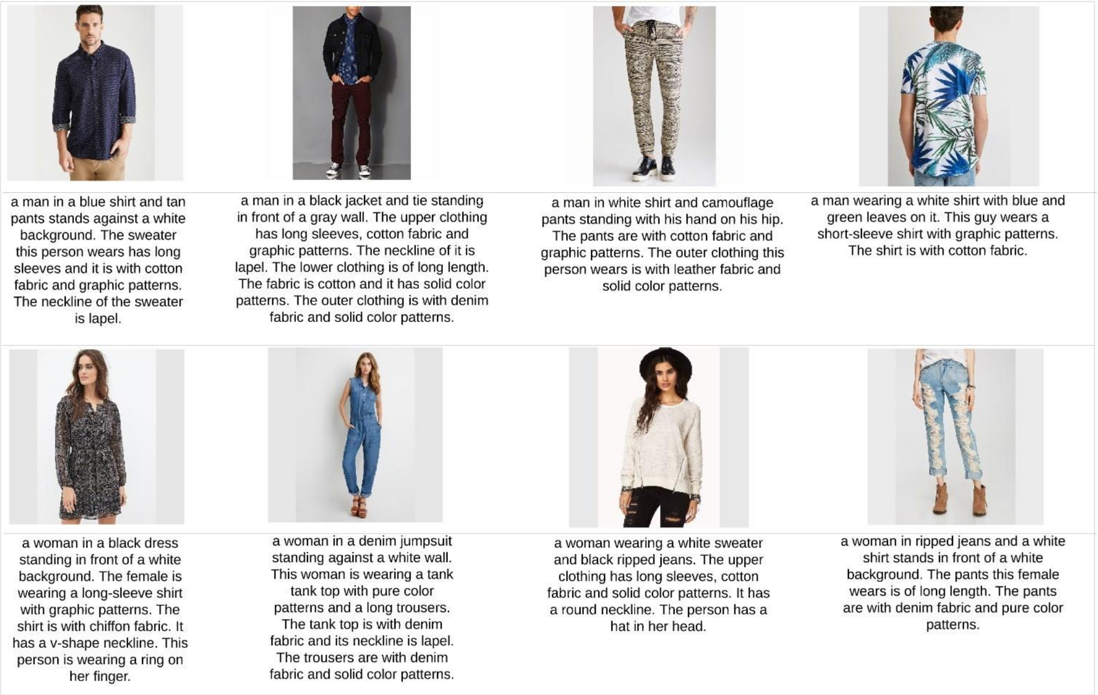
$$L_{D_i} = -E_{x_i \sim p_{data_i}} [\log D_i(x_i)] - E_{s_i \sim p_{G_i}} [\log(1 - D_i(s_i))]$$

where x_i is from the true image distribution p_{data_i} at the i^{th} scale, and s_i is from the model distribution p_{G_i} at the same scale. These discriminators are trained in parallel and focused on a single image scale.

Guided by the trained discriminators, the generators are optimized to jointly approximate multi-scale image distributions (p_{data_0}, p_{data_1}, ..., $p_{data_{m-1}}$) by minimizing the following loss function,

$$L_G = -\sum_{i=1}^{m} L_{G_i}, \quad L_{G_i} = -\mathbb{E}_{s_i \sim p_{G_i}} [\log D_i(s_i)]$$

where p_{data_i}, or the loss function for estimating the image distribution at the i^{th} scale, is represented by L_G. The generators G_i and discriminators D_i are alternately optimized until convergence throughout the training process.



a man in a blue shirt and tan pants stands against a white background. The sweater this person wears has long sleeves and it is with cotton fabric and graphic patterns. The neckline of the sweater is lapel.

a man in a black jacket and tie standing in front of a gray wall. The upper clothing has long sleeves, cotton fabric and graphic patterns. The neckline of it is lapel. The lower clothing is of long length. The fabric is cotton and it has solid color patterns. The outer clothing is with denim fabric and solid color patterns.

a man in white shirt and camouflage pants standing with his hand on his hip. The pants are with cotton fabric and graphic patterns. The outer clothing this person wears is with leather fabric and solid color patterns.

a man wearing a white shirt with blue and green leaves on it. This guy wears a short-sleeve shirt with graphic patterns. The shirt is with cotton fabric.

a woman in a black dress standing in front of a white background. The female is wearing a long-sleeve shirt with graphic patterns. The shirt is with chiffon fabric. It has a v-shape neckline. This person is wearing a ring on her finger.

a woman in a denim jumpsuit standing against a white wall. This woman is wearing a tank tank top with pure color patterns and a long trousers. The tank top is with denim fabric and its neckline is lapel. The trousers are with denim fabric and solid color patterns.

a woman wearing a white sweater and black ripped jeans. The upper clothing has long sleeves, cotton fabric and solid color patterns. It has a round neckline. The person has a hat in her head.

a woman in ripped jeans and a white shirt stands in front of a white background. The pants this female wears is of long length. The pants are with denim fabric and pure color patterns.

Figure 3: Dataset Archives

5. Implementation Details

To generate text embeddings for the captions in the training data, we have used the OpenAI CLIP (Contrastive Language-Image Pre-training), a pre-trained model which can associate an image with a text caption which can contain any words from the English language. It does this by jointly mapping their embedded representations in a shared semantic space. CLIP consists of two main components: an image encoder and a text encoder. The image encoder is generally based on convolutional neural net-works (CNNs) such as ResNet, which processes input images to extract visual features. Similarly, the text encoder is often based on transformer architectures like BERT, which encodes textual descriptions into embedded representation. CLIP is pretrained on a large dataset containing diverse image-text pairs. During pretraining, the model learns to associate semantically similar image-text pairs while contrasting them with dissimilar pairs. After pretraining, CLIP achieves a common embedding space where both images and text are represented as vectors. In this embed-ding space, images and text that have similar semantics are mapped close to each other in this space. This model was fine-tuned on DeepFashion dataset to associate the images with their captions. During the inference, textual embed-ding vectors were extracted for each caption. As a result, pairs of images were obtained along with their corresponding textual descriptions.

StackGAN-v2 model is designed to generate images with resolution of 256x256. The input text embedding is concatenated with noise and is denoted by c (Conditioned variable) and is first transformed to a 4×4×64 N_g feature tensor. Here, N_g denotes the number of channels in the tensor. Then, this 4×4×64N_g tensor is gradually transformed to 64×64×4N_g, 128×128×2N_g, and eventually 256×256×1N_g tensors at different layers of the network by six up-sampling blocks. The intermediate 64×64×4N_g, 128×128×2N_g, and 256×256×1N_g features are used to generate im-ages of corresponding scales with 3×3 convolutions. During this process, conditioning variable c is also directly fed into intermediate layers of the network to ensure that encoded information within it is not omitted.

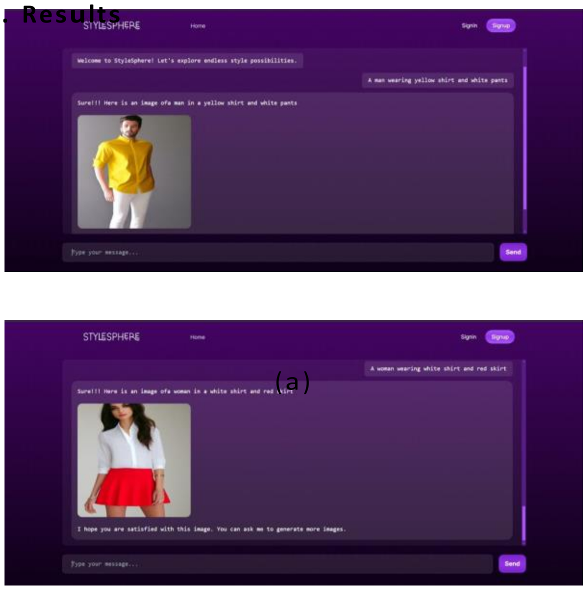
All the discriminators D_i have down-sampling blocks and 3×3 convolutions to transform the in-put image to a 4×4×8N_d tensor. The sigmoid function at the end of the discriminator is used to generate probabilities. We set N_g = 32, N_d = 64 and used two residual blocks between every two generators. ADAM optimizer with beta1 = 0.5 and a learning rate of 0.0002 is used for all models.

6. Dataset Details

The DeepFashion, curated by the Hong Kong University by Ziwei Liu et al.[24] consists over 800000 images with their respective images description in the range of 256 x 256 pixels.

Fashion-Gen consists of 293,008 high defination (1360 x 1360 pixels) fashion images paired with item descriptions provided by professional stylists.

7. Results



Welcome to StyleSphere! Let's explore endless style possibilities.

A man wearing yellow shirt and white pants

Sure!!! Here is an image ofa man in a yellow shirt and white pants

Type your message...

(a)

A woman wearing white shirt and red skirt

Sure!!! Here is an image ofa woman in a white shirt and red skirt

I hope you are satisfied with this image. You can ask me to generate more images.

Type your message...

(b)

Figure 4: Images generated corresponding to the textual input

8. Evaluation Metrics

Method	Inception Score
64×64 StackGAN-v2	3.26 ± .01
256×256 StackGAN-v2	4.04 ± .05

Table 1: Inception score of StackGAN-v2

Method	Inception Score
GAN-INT-CLS	2.88 ± .04
GAWWN	3.62 ± .07

Table 2: Inception score of models previous to StackGAN-v2

9. Conclusion

In this paper, we have proposed StyleSphere, a conversational text-to-image generator for fashion apparels with the help of Stacked Generative Adversarial Networks (StackGAN). The images generated are conditioned on text descriptions to provide fine-tuned image.

References

[1]Huang, H., Yu, P.S., & Wang, C. (2018). An Introduction to Image Synthesis with Gener-ative Adversarial Nets. ArXiv, abs/1803.04469.

[2]Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D.N. (2016). StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 5908-5916.

[3]Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.

[4]Reed, Scott & Akata, Zeynep & Yan, Xinchen & Logeswaran, Lajanugen & Schiele, Bernt & Lee, Honglak. (2016). Generative Adversarial Text to Image Synthesis. arXiv:1605.05396

[5]Thanh-Tung, Hoang & Tran, Truyen & Venkatesh, Svetha. (2019). Improving generalization and stability of generative adversarial networks. arXiv:1902.03984

[6]Lomov, Ildar & Makarov, Ilya. (2019). Generative Models for Fashion Industry using Deep Neural Networks. 1-6. 10.1109/CAIS.2019.8769486.

[7]Sun, Wei & Bappy, Md Jawadul & Yang, Shanglin & Xu, Yi & Wu, Tianfu & Zhou, Hui. (2019). Pose Guided Fashion Image Synthesis Using Deep Generative Model. arXiv:1906.07251

[8]Patricia Cortajarena, University of Amsterdam, Amsterdam, The Netherlands. A novel generative model for fashion clothing based on sketch + user control image transformations.

[9]Jingyuan Zhu, Huimin Ma, Jiansheng Chen, Jian Yuan. (2024). Isolated Diffusion: Optimizing Multi-Concept Text-to-Image Generation Training-Freely with Isolated Diffusion Guidance. https://doi.org/10.48550/arXiv.2403.16954

[10]Basu, S., Saberi, M., Bhardwaj, S., Chegini, A.M., Masciceti, D., Sanjabi, M., Hu,

S.X., & Feizi, S. (2023). EditVal: Benchmarking Diffusion Based Text-Guided Image Editing Methods. ArXiv, abs/2310.02426.

[11]Xu, Tao & Zhang, Pengchuan & Huang, Qiuyuan & Zhang, Han & Gan, Zhe & Huang, Xiaolei & He, Xiaodong. (2017). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. arXiv:1711.10485

[12]Ak, Kenan & Lim, Hwee & Tham, Jo & Kassim, Ashraf. (2019). Semantically Consistent Hierarchical Text to Fashion Image Synthesis with an Enhanced-Attentional Generative Adversarial Network. 10.1109/ICCVW.2019.00379.

[13]Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., & Shao, J. (2019). Semantics Disentangling for Text-To-Image Generation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2322-2331. arXiv:1904.01480

[14]Qing Ping, Bing Wu, Wanying Ding, Jiangbo Yuan. Fashion-AttGAN: Attribute-Aware Fashion Editing With Multi-Objective GAN DOI:10.1109/CVPRW.2019.00044. Published in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

[15]Heewoo Jun, Rewon Child, Mark Chen, John Schulman, Aditya Ramesh, Alec Radford, Ilya Sutskever. Conditional Augmentation for Generative Modeling Published in ICML 2020·

[16]Mirza, Mehdi & Osindero, Simon. (2014). Conditional Generative Adversarial Nets. arXiv:1411.1784

[17]Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 649–657. arXiv:1509.01626

[18]Reed, Scott & Akata, Zeynep & Lee, Honglak & Schiele, Bernt. (2016). Learning Deep Representations of Fine-Grained Visual Descriptions. 49-58. 10.1109/CVPR.2016.13.

[19]Reed, Scott & Akata, Zeynep & Yan, Xinchen & Logeswaran, Lajanugen & Schiele, Bernt & Lee, Honglak. (2016). Generative Adversarial Text to Image Synthesis. arXiv:1605.05396

[20]X. Fu, E. Ch'ng, U. Aickelin and S. See, ”CRNN: A Joint Neural Network for Redundancy Detection,” 2017 IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China, 2017, pp. 1-8, doi: 10.1109/SMARTCOMP.2017.7946996.

[21]J. Redmon, S. Divvala, R. Girshick and A. Farhadi, ”You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[22]C. Szegedy et al., ”Going deeper with convolutions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[23]Liu, Ziwei & Luo, Ping & Qiu, Shi & Wang, Xiaogang & Tang, Xiaoou. (2016). DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. 1096-1104. 10.1109/CVPR.2016.124.

[24]International Conference on Computer Vi-sion (ICCV), Seoul, Korea (South), 2019 pp. 4169-4178. arXiv.1905.12384