

# Text-to-Image Fashion Item Generation: Be Your Own Designer (MLVU Final Report)

Junyeob Kim    Seokhyeon Park    Yeon Chae  
Seoul National University  
Seoul, Republic of Korea

{juny116, maheu, yeonchae62}@snu.ac.kr

## Abstract

*Generating text conditioned fashion image has promising impacts in real-world application, since it can assist people be their own designers for creating a range of fashion clothing for themselves. However, this is a challenging task as it requires rich understanding of diverse text inputs and high-quality image representation ability. Resulting models that requires complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks. Recent approach which autoregressively models text and image tokens as a single stream of data and applying self-attention showed great success. In this work, we apply Generative Adversarial Networks (GAN), highly compelling image generation model, to autoregressive transformer. In particular, we show (i) effectiveness of adopting GAN in transformer based fashion text-to-image generator, and (ii) how to properly train such model by conducting diverse experiments.*

## 1. Introduction

Especially in the fashion industry, trends change very fast with new designs or patterns come every day in the market. In recent years, advanced machine learning approaches have been successfully applied to various fashion-based problems such as attribute recognition[57, 17], attribute discovery[22, 62], recommendation[6, 23, 56], retrieval[21, 1, 2, 3]and human/fashion parsing[5].

In fact, a large body of literature exists that focuses on clothing segmentations[24, 31], recognition[?, 37], and fashion image retrieval[14, 19, 43]. Some other works have focused on identifying fashion ability[56, 64] or occupation[59] from the clothing in images. In addition, some researchers have explored methods for clothing retrieval, including within-scenario retrieval [41] and cross-scenario [42, 43].

According to rising demand in the market, recent re-

search is mostly on image recognition. Still, approach with text is important especially in the fashion industry in terms of the characteristics of fashion communication. Actually, in the age of digitalization, tons of data combined with text analyzing designs through images is provided. Especially at the stage of design, tons of detailed pattern and textiles are important factors separating clothes, which is still a demanding issue.

In the same context, recent researches with GAN-based methods have been developed to deal with text-to-image synthesis to deal with the generating high-quality issue[53, 5]. Still, such a problem with distinguishing clothes with detailed factor and generating high resolution images has left.

This study proposes a framework to enable high-quality image representation ability, while keeping the rich understanding of text inputs mainly in the fashion domain. In this study, SOTA T2I methods is applied with GAN based decoder, dealing with the problem with generation and understanding of context. As a result, combined with existing autoregressive transformer which shows great accuracy, highly compelling image has generated with far more detail.

## 2. Related Works

### 2.1. Text-to-Image Generation

Not only can humans easily describe images as text, but when reading text, humans can easily visually recall the corresponding image. With the rapid development of deep learning technology in recent years, it has become possible to mechanically implement the human visual cognitive structure. Computer vision research that can take an image as an input and explain the image in text form has been actively conducted, and has also demonstrated performance beyond humans. However, research on image generation including the field of compositing images from text-based description called *text-to-image(T2I)* has been relatively slow.

**GAN-based Text-to-Image Generation** Meanwhile, with the advent of Generative Adversarial Networks

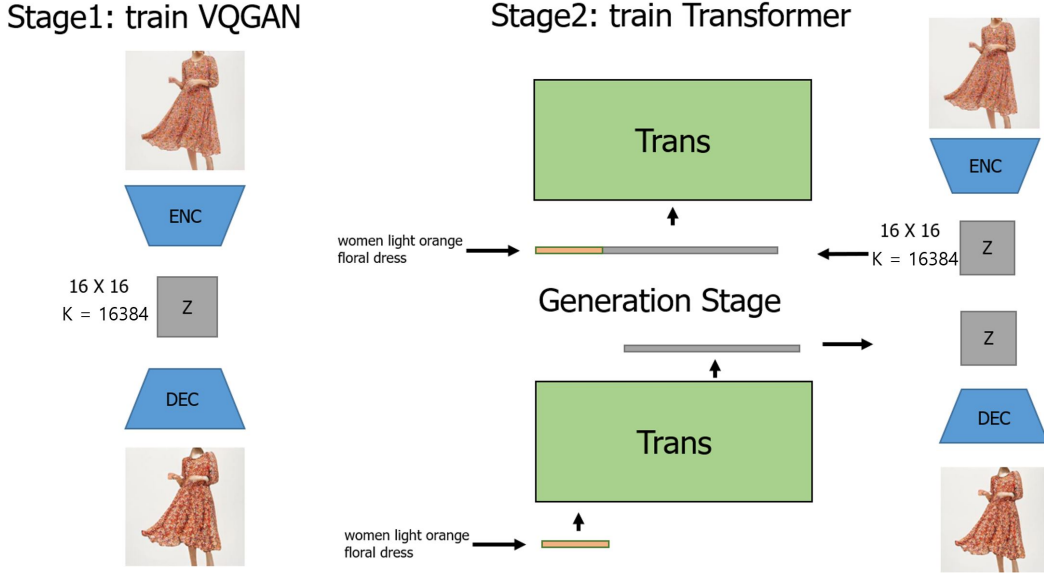


Figure 1. Overview of proposed method

(GANs)[20], research that can effectively generate images in an unsupervised manner has increased remarkably. It showed relatively good performance while competing and learning the two networks, and based on this, it was used in various fields such as human face synthesis, image translation and style transfer. Based on this technology, Reed et al. [52] paved the way for text-to-image research by presenting a model using conditional GAN using a pertained encoder. ver the next five years, with the comprehensive dataset such as COCO[40], this text-to-image field has grown rapidly. A text-to-image models were created through a variety of neural network architectures. In the early days, studies based on a stacked architecture using multiple stacked generators have been conducted to create a high-resolution image with better performance. Zhang et al.[66, 67] improved the image quality by changing the model using a multi-scale generator in StackGAN and StackGAN++, and similar studies such as FusedGAN[7] and HDGAN[68] were conducted. In addition, there was also a model using the attention mechanism. Representatively, AttnGAN[63] applied attention to text and images based on StackGAN++ to create a fine-grained model using words and global sentence vector. Huang et al[28] and ControlGAN[38] also tried to improve performance by using an attention mechanism. Methods such as SD-GAN[65], SEGAN[61], and Text-SeGAN[9] using Siamese Network, which two branches share and use model parameters, were also presented, and sentence and word embeddings were cascaded based on Cycle Consistency under the influence of CycleGAN[69]. Methods such as MirrorGAN[48] used in the generator structure

were conducted. Based on dynamic memory networks, DM-GAN[70] receives images and word features, and then generates high-quality images using memory writing gates. In addition, textStyleGAN[60], which enables semantic manipulation based on StyleGAN[33], was developed.

**Transformer-based image generation** Most text-to-image models using GAN, but text-to-image models using other architectures were also suggested. Models using Variational Autoencoders (VAEs) [36, 51] and Autoregressive models [27, 45, 44], flow-based models [15, 16, 35], score-matching networks [29, 32, 58], transformer models[46, 10, 18, 50]. Among these approaches, as transformers are widely used in the Natural Language Processing(NLP) field, there has been a growing trend to study the relationship between images and texts using transformers. OpenAI has developed CLIP[49] that can effectively learn visual concepts from natural language supervision. Based on this, a study was conducted to enable text-driven image manipulation, called SytleCLIP[47], using multi-modal embedding space and semantic similarity between text and image. In addition, OpenAI released a zero shot text-to-image model called DALL-E[50] based on GPT-3[8], an autoregressive language model . Also, by applying the GPT-based transformer model to the pixel sequence, OpenAI presented Image GPT[10] which is a high-performance model that can learn features without domain-specific model architecture.

## 2.2. AI in Fashion

**Image Generation in Fashion** Recent research proposes an approach that will accept text input from the user about

the fashion pattern and the model will generate images of fashion clothing based on the text input[30]. Still, generating images according to natural language descriptions is a challenging task[55],

In that text to image generation (T2I) model aims to generate photo-realistic images which are semantically consistent with the text descriptions[39]. The dataset used in this study is biased to birds and especially in the Fashion industry, such technology is still insufficient. There is a lack of research in the fashion industry simultaneously adopting the image with text understanding[12]. With the approach to image generation based on text, topics in fashion could be detailed.

**Text-to-Image task in Fashion** It is believed that images evoke deeper elements of human consciousness compared to text; this is partly due to the age of brain parts that process visual information[34]. Still in the fashion industry, visual information is given as combined with text messages.

Text-to-image synthesis is quite an important task which would allow an artist to design specific clothing products with text information[5]. Unlike conditioning on attributes[11, 4], the use of text offers more flexibility for specifying desired attributes for image synthesis[5].

Further, posts can be in various formats, including single photo, multiple photos, videos, text, text-embedded photos or text-embedded videos[5]. Fashion is the industry where value is shared through images and texts, especially from the user’s point of view, it is important to consider image and text at the same time. In other words, Synthesizing images based on text descriptions is an important task.

### 3. Approach

Our goal is to create fashion T2I model with rich understanding of diverse text inputs and high-quality image generation ability. Our work mostly follows the state of the art transformer based T2I generator DALL-E[50].

#### 3.1. Vector Quantised GAN

Esser et al.[18] recently proposed Vector Quantised GAN (VQGAN). VQGAN combines the effectiveness of the inductive bias of CNNs with the expressivity of transformers, enables them to model and thereby synthesize high-resolution images. VQGAN use CNNs to learn a context-rich vocabulary of image, and utilize transformers to efficiently model their composition within high-resolution images.

GAN based models are well known for higher image quality relative to VAE based models in terms of details in the image. Experiments in Esser et al.[18] shows that VQGAN with smaller latent size and smaller codebook size generated sharper image than Vector Quantised VAE (VQVAE). In addition, our experiment results in ?? also shows

that VQGAN has better reconstruction ability in fashion domain. Experiments details are in 4.2

Proposed model substitutes VQGAN for VQVAE in DALL-E fashioned training.

#### 3.2. Decoder-only multi-modal Transformer

In the field of natural language process, encoder only transformers (e.g. BERT[13] are known for better understanding, and decoder only transformers (e.g. GPT[8]) are known for better generation ability. We followed these findings in our work by applying decoder-only transformer. In detail, our final model used 8 layer transformer decoder with 8 attention heads with dimension 64.

#### 3.3. Training Method

Inspired by DALL-E, we applied same 2-stage method. Overview of proposed method is shown in 1.

In stage 1, train a VQGAN to compress each  $256 \times 256$  RGB image into a  $16 \times 16$  grid of image tokens, each element of which can assume 16384 possible values.

In stage 2, concatenate up to 80 BPE-encoded text tokens with the  $16 \times 16 = 256$  image tokens, and train an autoregressive transformer to model the joint distribution over the text and image tokens.

In generating stage, use 80 BPE-encoded text tokens as the input of the transformer to predict the  $16 \times 16 = 256$  image tokens. Decode  $16 \times 16 = 256$  image tokens to generate image.

### 4. Experiments

This section evaluates the ability of pre-trained reconstruction models (sec 4.2), comparison in diverse settings of training dataset (sec 4.3). Furthermore, in Sec. 4.4 and 4.4 we evaluate our model by providing a quantitative and qualitative comparison to our baseline model VQVAE DALL-E. Finally, in Sec. 4.6 we report other factors regarding the performance of the proposed model.

#### 4.1. Data preparation

Fashion dataset we used for training[53] contains various categories of apparel, including accessories, hat, etc; total 260,468 images with text-paired data for each image. On the stage of data pre-processing, extra categories like accessories and sunglasses, considered as a disturbance in training overall visual patterns, especially the form of the apparel, so these images are removed. Also for images, dataset of given images has a pose diversity, front, back, side view, etc. We pre-process the dataset to only consider about the front-view images to prevent the blurred results after training and to train well with the fixed pose.



Figure 2. Reconstruction Comparison. Original Images and 5 images generated by each VAEs which is used for evaluation

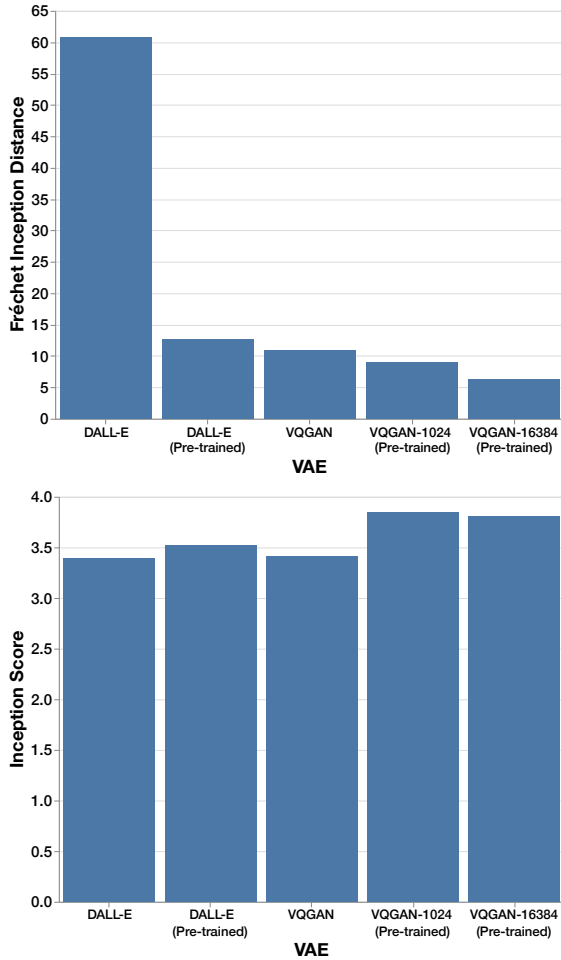


Figure 3. FID (Fréchet Inception Distance) and IS (Inception Score) of 5 VAEs

## 4.2. Reconstruction model

A comparative experiment was conducted to enable text-to-image generation using the VAE model with the best performance. Qualitative and quantitative analyses were con-

VAE	FID	IS
DALL-E	60.80	3.39
DALL-E (Pre-trained)	12.60	3.52
VQGAN	10.84	3.41
VQGAN 1024 (Pre-trained)	8.97	3.85
VQGAN 16384 (Pre-trained)	6.30	3.80

Table 1. Detailed FID (Fréchet Inception Distance) and IS (Inception Score) of 5 VAEs

ducted on the images generated by each model using the images of the dataset. There are a total of five VAE models used, and they can be broadly divided into two. The first is the VQVAE-based model suggested in the DALL-E[50], and the second is the VQGAN[18]-based model. The DALL-E VQVAE-based model was tested with a model trained with the prepared fashion dataset along with the existing pre-trained model. For the VQGAN-based model, a pre-trained model with codebook dimensionality of 1024 and 16384, respectively, and a model trained with a fashion dataset were used.

Qualitative analysis is performed based on reconstructed images by randomly selecting images from within the fashion dataset, and images that can represent each characteristic are shown in 2. First of all, the common point is that they cannot properly express the shape of the face. The image of the DALL-E series expressed the shape of the face well, but it was blurred, and the image of the VQGAN series showed a distorted shape although the sharpness was high. This point is well reflected in the overall image, including not only the face but also the fashion items. In the case of the dataset trained directly with the fashion dataset, the resolution and size of the dataset are insufficient, so the reconstruction quality is lower than that of the pre-trained model. It can be shown experimentally that the pre-trained DALL-E model shows the overall image characteristics well, and the detail and sharpness are well represented by the pre-trained VQGAN model with a codebook dimensionality of 16384.

Quantitative evaluation was performed using Fréchet Inception Distance (FID)[26] and Inception Score (IS)[54], which are mainly used as metrics to evaluate generative models. IS is a metric that evaluates performance based on the quality of the generated image and their diversity based on the entropy of the distribution of synthetic data, and means the higher the score, the better the performance. Unlike IS, which uses only generated images, FID evaluates the performance of the generator using the distribution of real images, and means the lower the value, the better the performance. The results of using this FID and IS as evaluation indicators for the five VAEs can be seen in 4.2 and 3. First, in FID score, it was shown that VQGAN-16384 showed the best performance, and in IS, VQGAN-1024 had the best score. A characteristic point is that the difference

in IS score, considering only the generated image, is not large, whereas FID showed low performance when manually trained with the fashion dataset. Based on the qualitative and quantitative analysis results, we selected the pre-trained VQGAN-16384 model which showed the best overall performance for text-to-image generation.

### 4.3. Data Comparison

First, in this section, we compare the result with varying the conditions of the data. We set two main variables and train with other conditions remaining the same. First variable is the categories used in training. Data is firstly divided into sub-categories: top-wear, bottom-wear, set. Here, due to the data imbalance matter, set-wear is trained together with the top-wear data as a unit class. Sub-categories are also divided into the type of apparel: jeans, skirts, dress, etc. 4 shows the detail composition of the data and text-caption information given in the data. With training whole ap-

	Top-wear	Bottom-wear	Set
Subcategory	T-SHIRTS (26%)	TROUSERS(24%)	DRESSES(91%)
	SHIRTS (13%)	JEANS(24%)	SUIT(1%)
	CREWNECKS (11%)	PANTS(5%)	
	SWEATSHIRTS (10%)		
	HOODIES&ZIPUPS (8%)		
	Input Name	First-Line	
Shirt	Slate Blue Long Sleeve Acid Print Top	Long sleeve crewneck printed 'Acid' sweatshirt in slate blue	
Jeans	Indigo & Turquoise Overdye Grupee Jeans	Skinnyfit jeans in indigo	
Dress	Black Tech Suiting Fitted Dress	Short sleeve waisted dress in black	

Figure 4. Data composition and example of caption set

parel, due to the massive amount of data which contains the various information of the apparels, rhetorical information, such as with pattern, styles is comparatively well-trained. Only training with one category did not work well when the number of data is not that large in one category to be well trained. As an example of result, here we shows the comparison among under three conditions: First train with the whole apparel data, second with bottom-wear sub-category apparel, third only with one type of apparel data. Result comparison shows in 5.

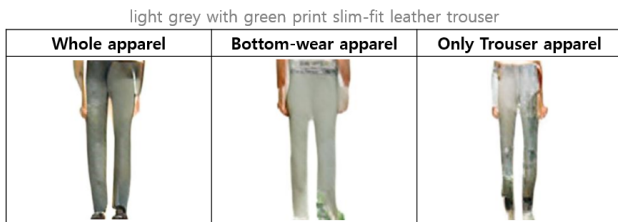


Figure 5. Generated images varying caption condition

Second variable is the text caption used in training. From the data[53] initially two types of text data is given. One type is a 7-line-length brief explanation of images(detail

caption), and the other is a text file with the name of image(name caption). Figure above 4 shows the example of comparison of two text files. For training, text caption we used is varied with below conditions: First, using whole detail caption not pre-processed, second, using only first-line information of detail caption including style, color, subcategory info. Third, using first line of detail caption with name caption. The result show in 6. With using whole captions, text information quite contains excessive information about the image, even not concerned with a visual form. Especially in terms of color and styles, third condition, using first line of detail caption with name caption quite well-trained on our data.



Figure 6. Generated images varying category condition

### 4.4. Quantitative results

We evaluate our model using three quantitative metrics. IS and FID scores are used to evaluate the quality of generated image. For evaluating the similarity between the input text and the generated image, we used CLIP score as the metric. Recent study[25] showed that CLIP, a cross-modal model pre-trained on web crawled image+caption pair data, can be used for robust automatic evaluation of image captioning without need for references. We believe that CLIP can also be a powerful tool in evaluating T2I models.

For in-dataset text experiments, we sample 5 random samples from training data. 512 images were generated for each samples. In total, 2560 images are evaluated. For random text experiment, we used 5 texts that are randomly created by human. Similar to in-dataset text experiments, 2560 images (512 for each text) are generated and evaluated. Details in sampled and created sentences are in table 3. As shown in table 2 VQGAN based transformer (ours) showed better results in all three metrics. As expected, in-dataset text showed better results than random created texts.



























Input	Long sleeve waffle-knit cotton in black			Skinnyfit jeans in deep indigo Dry Twill wash		
DALL-E						
OURS						
Input	Relaxed-fit nylon taffeta shorts in pink			Short sleeve t-shirt in blue		
DALL-E						
OURS						

Figure 7. Generated images, texts are sampled from training data

Input	light grey with green print slim-fit leather trouser			dark blue slim-fit narrow leg jean		
DALL-E						
OURS						
Input	baby-blue long sleeve wool crewnecks			short-sleeve crewneck dress in white		
DALL-E						
OURS						

Figure 8. Generated images, texts are randomly generated from user

#### 4.5. Qualitative results

CLIP score sorted top-3 generated images are compared. [7](#) shows generated images from training samples, [8](#) shows

generated images from random sample texts. Same texts in [3](#) are used to generate images.

Overall, generated images from our model showed much sharper images where images from DALL-E seem much

Model (data setting)	FID	IS	CLIP
DALL-E (in-dataset)	80.82	2.27	26.42
Ours (in-dataset)	70.5	2.85	27.45
DALL-E (random)	104.92	2.07	22.85
Ours (random)	90.41	2.56	26.26

Table 2. Quantitative results; DALL-E, OpenAI pre-trained model

Setting	Sampled Texts
In-dataset	Skinnyfit jeans in deep indigo Dry Twill wash Relaxed-fit nylon taffeta shorts in pink Long sleeve waffle-knit cotton pullover in black Short sleeve t-shirt in blue Sleeveless jersey knit dress in red Blue Faded Palm Leaf T-Shirt
Random	Dark blue slim-fit narrow leg jean Light grey with green print slim-fit leather trouser Baby-blue long sleeve wool crewnecks Long-sleeve green shirts with diagonal pattern Short-sleeve crewneck dress in white

Table 3. Texts used in evaluation

blurry. Unlike the results in 2 quality of images generated from random created text are not much worse than images generated from in-dataset texts. This shows that our model is well generalized.

#### 4.6. Other factors

In efficiency wise, since DALL-E model used 32 X 32 grid latent inference time of DALL-E model was about 8 times slower than our model (using single RTX Titan GPU).

Generated results are quite dependent on input text template. In our experiments, IS score varied in range of 0.2 points only by shuffling the order of the tokens. In addition, original text was not always optimal.

### 5. Conclusion

In this project, we trained T2I model by applying GAN to autoregressive transformer. By autoregressively modeling text and image tokens as a single stream of data, our proposed model was able to understand diverse text inputs and generate fine images. Due to the limited time and resources, our final results are far below our initial expectations. However, our work shows that applying GAN to transformer based T2I is effective in both quality and efficiency. With enough resources (data, hardware), we believe that the proposed model have potential to be used in real world fashion applications like fashion design.

For future work, applying encoder-decoder transformer to create multi-task multi-modal model may be a great

project. In addition, applying prompt optimization technics to boost the quality of the generated images will be a fun project.

### References

- [1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7708–7717, 2018. 1
- [2] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679. IEEE, 2018. 1
- [3] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Which shirt for my first date? towards a flexible attribute-based fashion query system. *Pattern Recognition Letters*, 112:212–218, 2018. 1
- [4] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10541–10550, 2019. 3
- [5] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network. *Pattern Recognition Letters*, 135:22–29, 2020. 1, 3
- [6] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE international conference on computer vision*, pages 388–397, 2017. 1
- [7] Navaneeth Bodla, Gang Hua, and Rama Chellappa. Semi-supervised fusedgan for conditional image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 669–683, 2018. 2
- [8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 2, 3
- [9] Miriam Cha, Youngjune L Gwon, and HT Kung. Adversarial learning of semantic relevance in text to image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3272–3279, 2019. 2
- [10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 2
- [11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image

- translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3
- [12] Alessandro Da Giau, Nicolai J Foss, Andrea Furlan, and Andrea Vinelli. Sustainable development and dynamic capabilities in the fashion industry: A multi-case study. *Corporate Social Responsibility and Environmental Management*, 27(3):1509–1520, 2020. 3
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3
- [14] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, pages 8–13, 2013. 1
- [15] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2
- [16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2
- [17] Qi Dong, Shaogang Gong, and Xiatian Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 520–529. IEEE, 2017. 1
- [18] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *arXiv preprint arXiv:2012.09841*, 2020. 2, 3, 4
- [19] Jianlong Fu, Jinqiao Wang, Zechao Li, Min Xu, and Hanqing Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *Asian conference on computer vision*, pages 420–431. Springer, 2012. 1
- [20] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2
- [21] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015. 1
- [22] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1463–1471, 2017. 1
- [23] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017. 1
- [24] Basela Hasan and David C Hogg. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, pages 1–11. Citeseer, 2010. 1
- [25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5
- [26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. 4
- [27] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018. 2
- [28] Wanming Huang, Richard Yi Da Xu, and Ian Oppermann. Realistic image generation using region-phrase attention. In *Asian Conference on Machine Learning*, pages 284–299. PMLR, 2019. 2
- [29] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 2
- [30] Anish Jain, Diti Modi, Rudra Jikadra, and Shweta Chachra. Text to image generation of fashion clothing. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 355–358. IEEE, 2019. 3
- [31] Nataraj Jammalamadaka, Ayush Minocha, Digvijay Singh, and CV Jawahar. Parsing clothes in unrestricted images. In *BMVC*, volume 1, page 2, 2013. 1
- [32] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Taquet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020. 2
- [33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2
- [34] Hyekyun Kim and Ella Kidd. A analysis of clothing perception through fashion image. *Psychology and Education Journal*, 58(2):2713–2718, 2021. 3
- [35] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018. 2
- [36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [37] Iljung S Kwak, Ana Cristina Murillo, Peter N Belhumeur, David J Kriegman, and Serge J Belongie. From bikers to surfers: Visual recognition of urban tribes. In *Bmvc*, volume 1, page 2. Citeseer, 2013. 1
- [38] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019. 2
- [39] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. *arXiv preprint arXiv:2104.00567*, 2021. 3
- [40] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2



- [41] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM international conference on Multimedia*, pages 619–628, 2012. **1**
- [42] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3330–3337. IEEE, 2012. **1**
- [43] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. **1**
- [44] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. *arXiv preprint arXiv:1812.01608*, 2018. **2**
- [45] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixellcn decoders. *arXiv preprint arXiv:1606.05328*, 2016. **2**
- [46] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. **2**
- [47] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. **2**
- [48] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrororgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019. **2**
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. **2**
- [50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. **2, 3, 4**
- [51] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019. **2**
- [52] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. **2**
- [53] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. **1, 3, 5**
- [54] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. **4**
- [55] Henning Schulze, Dogucan Yaman, and Alexander Waibel. Cagan: Text-to-image generation with combined attention gans. *arXiv preprint arXiv:2104.12663*, 2021. **3**
- [56] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2015. **1**
- [57] Edgar Simo-Serra and Hiroshi Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 298–307, 2016. **1**
- [58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. **2**
- [59] Zheng Song, Meng Wang, Xian-sheng Hua, and Shuicheng Yan. Predicting occupation via human clothing and contexts. In *2011 International Conference on Computer Vision*, pages 1084–1091. IEEE, 2011. **1**
- [60] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje ter Hoeve. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909*, 2020. **2**
- [61] Hongchen Tan, Xiuping Liu, Xin Li, Yi Zhang, and Baocai Yin. Semantics-enhanced adversarial nets for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10501–10510, 2019. **2**
- [62] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *European Conference on Computer Vision*, pages 252–268. Springer, 2016. **1**
- [63] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. **2**
- [64] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. Chic or social: Visual popularity analysis in online fashion networks. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 773–776, 2014. **1**
- [65] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2327–2336, 2019. **2**
- [66] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. **2**
- [67] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stack-

- gan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019. 2
- [68] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6199–6208, 2018. 2
- [69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2
- [70] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 2