Let AI Clothe You: Diversified Fashion Generation

Rajdeep H. Banerjee(✉), Anoop Rajagopal, Nilpa Jha, Arun Patro, and Aruna Rajan

Myntra Designs Pvt. Ltd., Bengaluru, India
{rajdeep.banerjee,anoop.kr,nilpa.jha,arun.patro,aruna.rajan}@myntra.com

Abstract. In this paper, we demonstrate automation of fashion assortment generation that appeals widely to consumer tastes given context in terms of attributes. We show how we trained generative adversarial networks to automatically generate an assortment given a fashion category (such as dresses and tops etc.) and its context (neck type, shape, color etc.), and describe the practical challenges we faced in terms of increasing assortment diversity. We explore different GAN architectures in context based fashion generation. We show that by providing context better quality images can be generated. Examples of taxonomy of design given a fashion article and finally automate generation of new designs that span the created taxonomy is shown. We also show a designer-in-loop process of taking a generated image to production level design templates (tech-packs). Here the designers bring their own creativity by adding elements, suggestive from the generated image, to accentuate the overall aesthetics of the final design.

Keywords: Fashion · GAN · Context

1 Introduction

Fashion is an interesting interplay of art and science - where science (data) informs us as to the customer's choice, art aims to create aesthetics with wide appeal that has relevance to the times we live in. Building fashion merchandise involves stocking customer needs and keeping a healthy long tail of merchandise that appeals to diverse aesthetics. Therefore, it is not enough to merely study what sells more to exploit the design process into making more of it, but also explore and create a diverse range of styles.

To appeal to a wide range of tastes in fashion - of the masses, as well as of fashionistas, fashion catalogs need week on week revisions, given the ephemeral nature of trends, and shifting consumer habits. How can we achieve this at scale, in a completely data driven way? Can we automate the process of design selection, and allow our human designers to focus on creative exploration over sorting through what has been done before to borrow elements? How can we inspire our designers by creating a base template of diverse designs, and allow them to innovate on top? We show how automated clothes design using generative adversarial

G. Carneiro and S. You (Eds.): ACCV 2018 Workshops, LNCS 11367, pp. 75–87, 2019.
https://doi.org/10.1007/978-3-030-21074-8_7

networks (GANs) [3] - the promise of AI [6,12] - can be realised to the fullest to solve these problems.

Fig. 1. Fashion generation with DC-GAN (Color figure online)

Tops:
{"type":["regular","bardot","tank",...."a-line"] , "length":["cropped","regular","longline"] ,
"neck":["round","v-neck","boat","polo"] , "pattern":["checked","solid","printed","stripes"], "sleeve length":
["short", "long", "sleeveless" , "three-quarter"]}

Dresses:
{"shape":["a-line","fit and flare","bodycon",....] , "length":["mini","maxi","midi"] ,
"neck":["round","v-neck","boat","halter"] , "pattern":["checked","solid","printed","stripes"], "sleeve length":
["short", "long", "sleeveless" , "three-quarter"]}

Fig. 2. Tops & Dresses taxonomy

To begin with, we use DC-GAN framework [6,7] and train with our catalogue images to produce images of fashion articles shown in Fig. 1. Many duplicates were observed, and also GAN generated images resembled our existing catalog too closely. For example (see Fig. 1), we observe many red/pink outfits in similar neck type and sleeve lengths. This resulted in producing more of what we already have, and added less diversity to our mix of fashion assortment. Also most of the images generated were hard to interpret by designers. For instance in second row third image from left of Fig. 1 shows a style with half sleeves on right hand and short/cap sleeves on the left. This leads to an ambiguity in designers mind on what sleeves should be used. Hence, can we make the design process efficient in capturing both the popular elements of fashion through some context and the longer tail of diverse tastes?

In this work, we explore fashion generation with context (in terms of class attributes/text) to address this problem of generating a wider taxonomy of designs. We come out with an exhaustive list of taxonomy of design attributes curated by fashion designers. We use this taxonomy as context and explore two approaches: (1) AC-GAN [5] modified for multi-label classification and (2) Attention-GAN [14] where we use these attributes as text input to the network.

Approach (1) provided good results for coarse attributes present in Fig. 2. However, it is observed that fine grained attributes in a fashion image is most often present in the accompanying text description. Hence we use approach (2) to generate fashion images with fine grained text appended with coarse attributes present in the catalog.

We propose to use a conditional and Attention GAN [14] framework using the fashion attributes/text for every fashion article type, so that we maintain the diversity of the catalogued assortment by providing context in the generation process. We exploit descriptive text to generate a diverse assortment, by focussing the generative network onto important (attention) regions in the input images. This is a powerful technique as we demonstrate, because together with our work described in the later sections, it can be used to create a rich corpus of text to image mappings, and learn regions of interest that descriptive text of images refer to, eventually enabling text based generation of images. Visualize a scenario where a consumer of fashion can specify their wish as: "I want to wear a classy dress to attend a business dinner, the neckline should be conservative, yet not boring - and the prints and pattern should represent *bhil art* [1] - from

Fig. 3. Tops types examples.

Fig. 4. Dress shapes examples.

(a) Tops (b) Dresses

Fig. 5. Sample text description for tops and dresses (Color figure online)

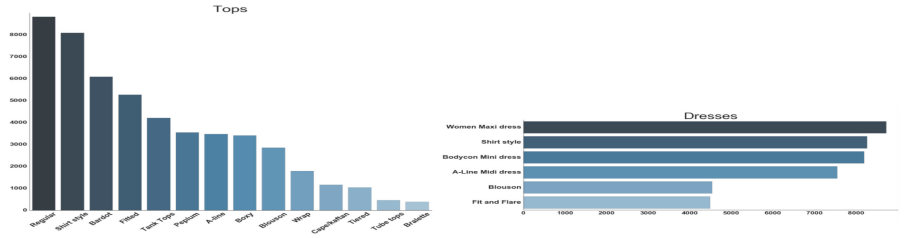

Fig. 6. Data distribution by taxonomy.

the region I belong to," and our fashion generation model can generate images of possible interest to the consumer given her description. This would not only solve the discovery problem in fashion retail for consumers, but also greatly simplify supply chain processes, that need a 3–6 month long predictive planning process and rely on a customer getting interested in a garment over mapping what the customer has in mind and manufacturing exactly that.

2 Related Work

Generation of fashion articles using generative models has generated a lot of interest [6,12]. In [12] a VAE/GAN is proposed to generate article types wherein the fashion image is first encoded using Variational Auto Encoders (VAE) and encoded data is then passed as input to GAN to learn its distribution.

In [6], DCGAN [7] was used to generate clothes and a practical system of taking it to production was evaluated with a human in-loop. [15] explores generation by separating components of a fashion article into texture, colour, and shape. In [2] a neural-style transfer approach is proposed to personalize and generate new fashion clothes. A system used to recommend styles to users and also aid the design of new products that matches user preferences is proposed in [4]. In [18] a two stage approach is proposed to generate new fashion based on design coding (using text) and a segmentation map. GANs have shown great results in generating sharper context through text. New image generation conditioned on text descriptions is explored in [10]. In [8,9,16,17] have encoded the whole text description into a global sentence vector as the condition for GAN-based image

generation. This lacks in encoding fine grained information at the word level, which AttnGAN [14] addresses.

In our work, we propose to use a conditional and Attention GAN [14] framework using the fashion attributes/text for every fashion article type so that we maintain the practicality of the catalog assortment by providing context in the generation process.

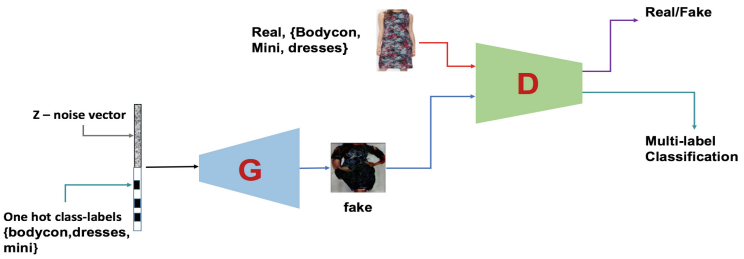

Fig. 7. Multi-label AC-GAN network.

3 Fashion Generation Using Taxonomy

Our dataset consists of images taken from a fashion e-commerce catalog for 2 major categories, women's tops and dresses. There are 50,630 images from Tops category and 62,236 images from Dresses category, all are front poses. The images are high resolution and are photographed under studio conditions. Additionally we have product descriptions and attribute provided for each image by an internal design team. We come up with a modified AC-GAN [5] which uses multi labeleled attribute tags for each image. For getting richer context through taxonomy we use product description along with the tags in the AttnGAN [14] framework. We talk more about the tags and description related taxonomy in the next section.

The term *fashion* encompasses substantial diversity in clothing apparel and accessories limited only by the stretch of imagination. Expressing this enormous complexity of visual experience in terms of well articulated distinctive classes is a challenging problem. In Fig. 2 we present the taxonomies of dresses and tops respectively. The tagged attributes represent high level features for the apparels as shown in Figs. 3 and 4, whereas the description has more fine grained details about the apparel. For example in Fig. 5, the description for the Tops image additionally tells it has *cold shouldered sleeves with a floral print*, whereas the tagged attribute *pattern* just tells us it is printed. For the red dress, we know from the description that the print is *ethnic motifs* and it has a *flared hemline*. Attribute tags usually contain a long tail, Fig. 6 shows the distribution of tops type and dress shape.

In our work, we explore attribute classes and text description as context using generative adversarial network (GAN) for fashion generation. For fashion generation with attribute classes we use the popular AC-GAN [5]. However, as a fashion image always come with multiple attributes (like round neck, printed, bardot Tops) we modify the AC-GAN softmax loss (for multi-class classification) with a multi-label soft margin loss for training that optimizes a multi-label one-versus-all loss based on max-entropy. Let x be an image and y_i^C be the one-hot encoding of multi-label representation where C is the total number of classes. Then loss function is given by

$$loss(x, y) = -\sum_i y_i \log((1 + exp(-x_i))^{-1}) + (1 - y_i) \log\left(\frac{exp(-x_i)}{1 + exp(-x_i)}\right) \quad (1)$$

We then use this Multi-label AC-GAN (network architecture in Fig. 7) for fashion generation using attributes. However, as mentioned earlier, attribute tagging is a cumbersome process and most e-commerce sites only give coarse attributes. Hence we use the fine grained descriptions often provided in the text additionally with attributes (used as text instead of classes) to generate fashion images. To this end, we use Attn-GAN [14]. Attentional Generative Adversarial Network (AttnGAN) allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. We describe how we create dataset to run AttnGAN in the next section.

4 Experiments

For a robust learning of the generative network we do augmentation of the training images, artificially creating more samples through scale change, adding noise etc. In all our experiments, we generate images of size 256×256.

In the multi-label AC-GAN network we use DC-GAN network architecture [7] for the generator and discriminator adding more up-sampling and down-sampling blocks to generate images at 256×256 pixels. We trained our network with an ADAM optimizer with learning rate = 0.0008, $\beta_1 = 0.9$ and one sided label smoothing [11] in the discriminator.

For fine grained image to text synthesis we use product descriptions in addition to coarse attribute tags in the AttnGAN [14] framework. As we had only one caption (description) per image, we augmented the descriptions per image. Within a category (for eg., dresses) for each image we take its coarse attribute key-value (see Fig. 2) combinations to form *attribute-classes*. Among all possible such classes we consider those with at least 3 images. Then for an image we randomly sample descriptions from the *attribute class* corresponding to that image, thus ensuring we augment descriptions including color from related images as in Fig. 8.

Original Description: (1 caption)
off-white and black striped top has a v-neck three-quarter sleeves with lace detail

Post augmentation: (3 captions)
1. off- white v-neck regular horizontal stripe regular regular sleeve three-quarter sleeve off-white black stripe top v-neck three-quarter sleeve lace detail
2. white v-neck regular stripe regular regular sleeve three-quarter sleeve white navy knit stripe top v-neck three-quarter sleeve
3. white v-neck regular horizontal stripe regular regular sleeve three-quarter sleeve white navy knit top stylised stripe v-neck three-quarter sleeve

Fig. 8. Augmented image captions

For the attention model we use a 3 stage Generator-Discriminator architecture as in [14]. For the DAMSM loss [14] we use $\lambda = 10$, image and text embedding dimension as 256 and 3 captions per image.

For a quantitative evaluation of our models we use Inception Score [11]. The intuition behind this metric is that good models should generate diverse but meaningful images. We compare outputs of the two generative models (sample images in Figs. 9, 10 and 11) using Inception Score [11], see Table 1. We see that multi-label AC-GAN even with coarse attributes has scores similar to AttnGAN, indicating a multi-label classifier in the discriminator is a good approach for context based generation.

To quantitatively evaluate image similarity we use multiscale structural similarity (MS-SSIM) [13]. MS-SSIM is a multi-scale variant of a well-characterized perceptual similarity metric that attempts to discount aspects of an image that are not important for human perception. MS-SSIM values range between 0.0 and 1.0; higher MS-SSIM values correspond to 'perceptually more similar images. We measure the MS-SSIM scores between 50 randomly chosen pairs of images within a given class for Multi-Label ACGAN model and same text attributes for the AttnGAN model, see Table 1. Figure 11 shows samples of image variants generated using the same fine grained text attribute. We also experiment to see how interpolating, as in [7], between 2 generated images helps in synthesizing new designs as shown in Fig. 12. The results in general are aesthetically well formed and additionally add to the variability of the generated outputs.

The above evaluations though talk about diversity and quality of data generated, they are not of use much in practice. In order to manufacture these type of newly generated styles, they need to be converted into what is called a "Tech pack". A tech pack is vital to production and a fashion designer spends good amount of time generating them. Tech pack contains sketches of the actual design and also some design elements added. This is where designer brings his inspiration and creativity to convert a generated image into an aesthetic sketch that can be taken to production.

Fig. 9. MultiLabel-ACGAN Generation. Labels used for generation shown below images

Fig. 10. Attention GAN Generation. Captions used for generation shown for above images

Figures 13 and 14 shows instances wherein the generated reference image is taken to a sketch detail and physical aesthetics are added in accordance with generated image by designers thought and inspiration. In Fig. 13 we a contrast center patch is added in accordance with reference image and in Fig. 14 a front

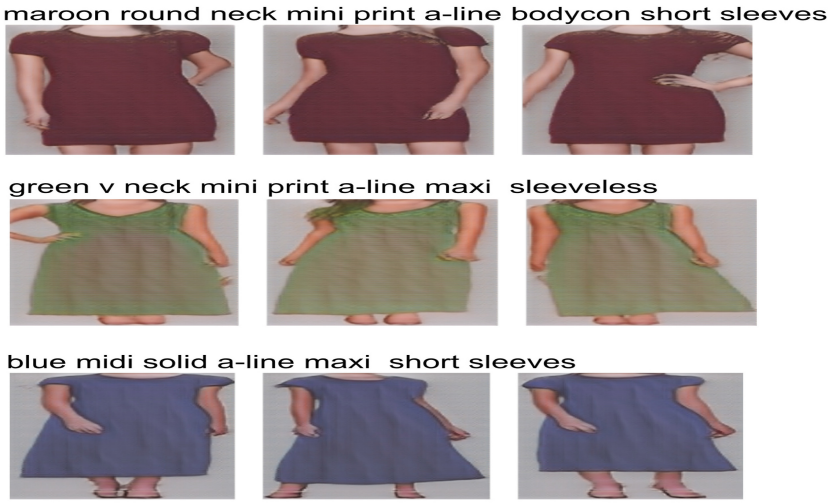

Fig. 11. Attention GAN Generation. Image variants generated with the same set of captions

Table 1. Inception and MSSIM score

Model	Inception score	MS-SSIM score
Multi-label AC-GAN	2.86 ± 0.133	0.15 ± 0.004
AttnGAN	2.96 ± 0.306	0.37 ± 0.007
Training dataset	3.46 ± 0.134	0.22 ± 0.006

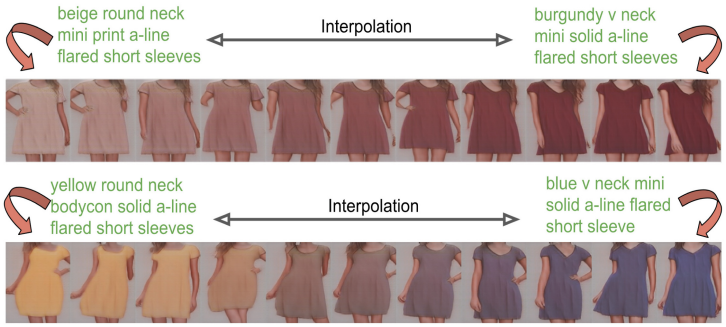

Fig. 12. Interpolation results between first and last images of each row

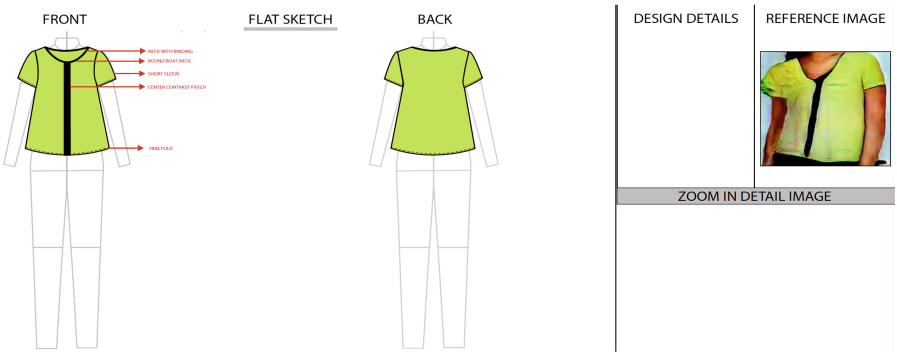

Fig. 13. Tech pack for Tops. The reference image refers to generated images from our approach. Flat sketch is the actual image which goes for manufacturing. We observe the interpretation of center patch adding to the aesthetics

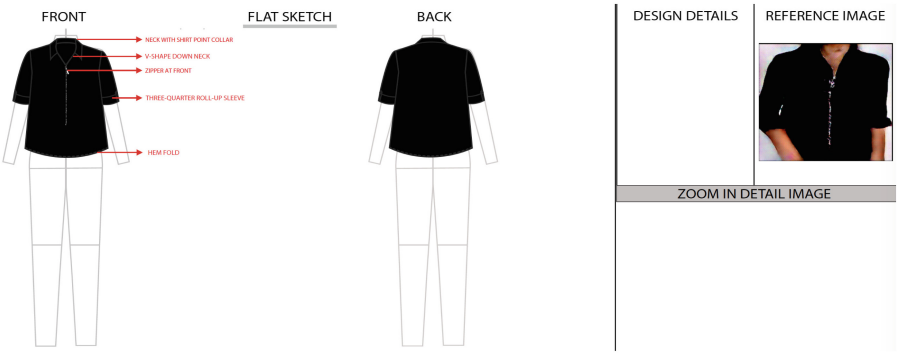

Fig. 14. Tech pack for Tops. The reference image refers to generated images from our approach. Flat sketch is the actual image which goes for manufacturing. The aesthetics is improved by the adding of zipper at the front which is indicative from the reference images.

zipper is added as a design element. Similarly in Figs. 15 and 16 we see tech packs for dresses. Notice the elasticated belt in Fig. 15 added by the designer to accentuate the looks which is vaguely indicated in the reference generated image. On the other hand in Fig. 16 there is a concealed zip added which is not visible. These calls on emphasizing the looks of the reference image is purely a designer call based on their experience but suggestive from the generated reference images. Thus it can be easily emphasized that Fashion generation with generative models enhances and eases the inspiration that a designer can obtain in creating aesthetic tech packs for manufacturing in fashion domain.

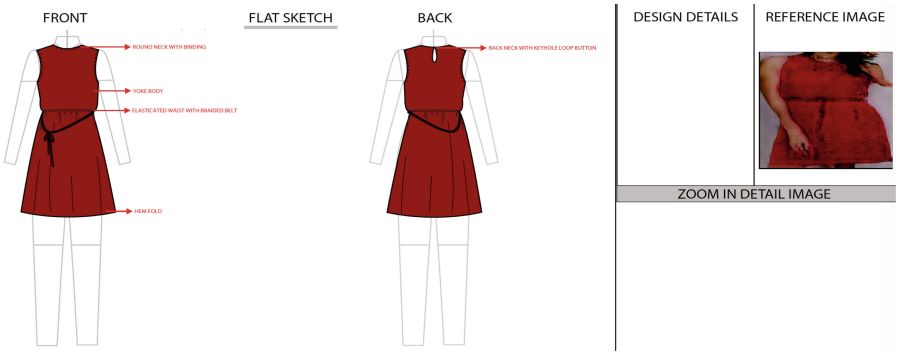

Fig. 15. Tech pack for dress. The reference image refers to generated images from our approach. Flat sketch is the actual image which goes for manufacturing. We observe the how the elasticated belt is added making the generated image aesthetic.

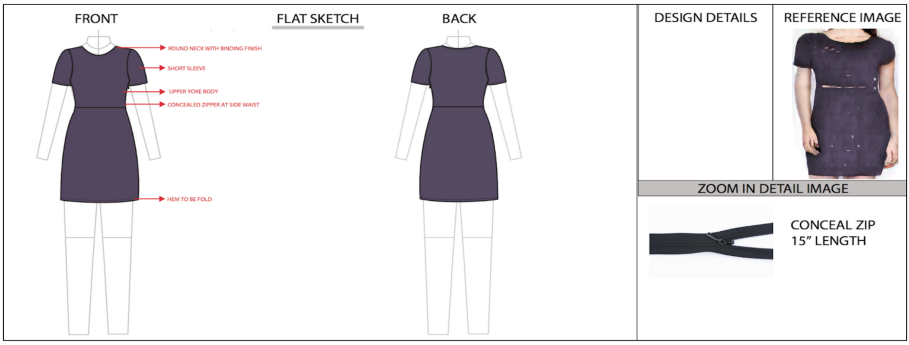

Fig. 16. Tech pack for dress. The reference image refers to generated images from our approach. Flat sketch is the actual image which goes for manufacturing. Addition of concealed zip (not visible) adds to aesthetics of style.

5 Conclusion

In this work, we have propose a contextual way of generating fashion articles using taxonomy/text. We have explored generation using coarse high level attributes in a multi-label setting as well as fine grained text descriptions in an attention setting. Both approaches aid in providing a better assortment in terms of addressing popular demand and the longer tail of diverse tastes and have an interesting byproduct of generating text to regions of interest mapping in images that can be further exploited for automated generation. We use the inception score as a metric for model evaluation and note that both models are close, with the attention model performing slightly better in some cases where the articles needed design details.

Our work has far reaching consequences for automated design generation, which has so far focused on automatically generating by training on given images as in [6], or training by separating components of a fashion article into texture, colour, and shape as in [15]. We exploit all fashion descriptions to auto generate images by following a region of interest detection based attention GAN framework. We plan to extend our mapping to include other crawled images/descriptors, social media content, and other user generated content on our platform to make the mappings richer, and the generative process catering to a wide range of fashion tastes.

References

1. The bhils - bhil art. http://bhilart.com/
2. Date, P., Ganesan, A., Oates, T.: Fashioning with networks: neural style transfer to design clothes. Machine Learning meets fashion, KDD (2017)
3. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
4. Kang, W.C., Fang, C., Wang, Z., McAuley, J.: Visually-aware fashion recommendation and design with generative image models. In: ICDM (2017)
5. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 2642–2651 (2017)
6. Osone, H., Kato, N., Sato, D., Muramatsu, N., Ochiai, Y.: Crowdsourcing clothes design directed by adversarial neural networks. In: 2017 Workshop Machine Learning for Creativity and Design. NIPS (2017)
7. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
8. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)
9. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Advances in Neural Information Processing Systems, pp. 217–225 (2016)
10. Rostamzadeh, N., et al.: Fashion-Gen: the generative fashion dataset and challenge. arXiv preprint arXiv:1806.08317 (2018)
11. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
12. Torres, TJ.: Deep style: inferring the unknown to predict the future of fashion (2015). https://multithreaded.stitchfix.com/blog/2015/09/17/deep-style/
13. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
14. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: ICCV (2017)
15. Yildirim, G., Seward, C., Bergmann, U.: Disentangling multiple conditional inputs in GANs. arXiv preprint arXiv:1806.07819 (2018)

16. Zhang, H., et al.: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. arXiv preprint (2017)
17. Zhang, H., et al.: StackGAN++: realistic image synthesis with stacked generative adversarial networks. arXiv preprint arXiv:1710.10916 (2017)
18. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: fashion synthesis with structural coherence. arXiv preprint arXiv:1710.07346 (2017)