Fashion-Gen: The Generative Fashion Dataset and Challenge

Negar Rostamzadeh [1] **Seyedarian Hosseini** [1 2] **Thomas Boquet** [1] **Wojciech Stokowiec** [1] **Ying Zhang** [1] **Christian Jauvin** [1] **Chris Pal** [3 1]

Abstract

We introduce a new dataset of 293,008 high definition (1360 x 1360 pixels) fashion images paired with item descriptions provided by professional stylists. Each item is photographed from a variety of angles. We provide baseline results on 1) high-resolution image generation, and 2) image generation conditioned on the given text descriptions. We invite the community to improve upon these baselines. In this paper we also outline the details of a challenge that we are launching based upon this dataset.

1. Introduction

Machine learning has recently been employed in many applications pertaining to the fashion industry. The use cases range from style matching (Bossard et al., 2012; Kalantidis et al., 2013; Liu et al., 2016), recommendation systems in e-commerce sites (Chen et al., 2012; Xiao et al., 2015; Kiapour et al., 2015; Chen et al., 2015; Simo-Serra et al., 2015), trend prediction, the ability for customers to virtually try on clothes (Han et al., 2017), and clothing type classification (Liu et al., 2012; Liang et al., 2016; Veit et al., 2015; Zhu et al., 2017b).

The availability of large-scale datasets such as DeepFashion (Liu et al., 2016) has fueled recent progress in applying deep learning to fashion tasks. However, there are still many aspects of the industry that computer vision methods have not been applied to. In this paper we explore the task of assisting fashion designers to share their ideas with others by translating verbal descriptions to images. Thus, given a description of a particular item, we generate images of clothes and accessories matching the description.

To explore these research directions we introduce here a new dataset of almost 300k high definition training images of clothes, and accessories accompanied by detailed design descriptions. Each description is provided by professional designers and contains fine-grained design details. Each product is photographed from multiple angles against a standardized background under consistent lighting conditions and annotated with matching items recommended by a stylist. See Figure 1 for examples.

In this paper we provide: 1) statistical details of the dataset, 2) detailed comparisons with existing datasets, 3) an introduction to the competition that we are launching on the task of text to image generation, with a brief explanation of the competition criteria and evaluation process, and 4) high-resolution image generation results using an approach based on the progressive growing of GANs (Karras et al., 2017), and text-to-image translation results using StackGAN-v1 (Zhang et al., 2017a), and StackGAN-v2 (Huang et al., 2017).

The paper is organized as follows: Section 2 discusses related work. Section 3 introduces the Fashion dataset, describes the collection procedure, and provides a statistical analysis of the dataset with details of our newly introduced challenge.[1] In Section 5, we describe baseline approaches and the evaluation process, including human evaluation. Section 6, concludes the paper and discusses future work.

2. Related Work

We first provide a summary of generative models used in text-to-image synthesis and then discuss related datasets.

2.1. Applications of Generative models

Generative Adversarial Networks (Goodfellow et al., 2014) have been used in a wide range of applications, including photo-realistic image super-resolution (Ledig et al., 2016; Sønderby et al., 2016), video generation (Denton & Fergus, 2018; Denton et al., 2017), inpainting (Belghazi et al., 2018), image-to-image translation (Isola et al., 2017; Zhu et al., 2017a; Taigman et al., 2016) and text-to-image syn-

[1]Element AI [2]MILA-University of Montreal [3]Ecole Polytechnique of Montreal. Correspondence to: Negar Rostamzadeh <negar@elementai.com>.

To be presented in the 35^{th} International Conference on Machine Learning ***Workshop on Theoretical Foundations and Applications of Deep Generative Models***, Stockholm, Sweden, PMLR 80, 2018.

[1]The competition is part of the first workshop of Computer Vision for Fashion, Art and Design at ECCV. The challenge website is https://fashion-gen.com/

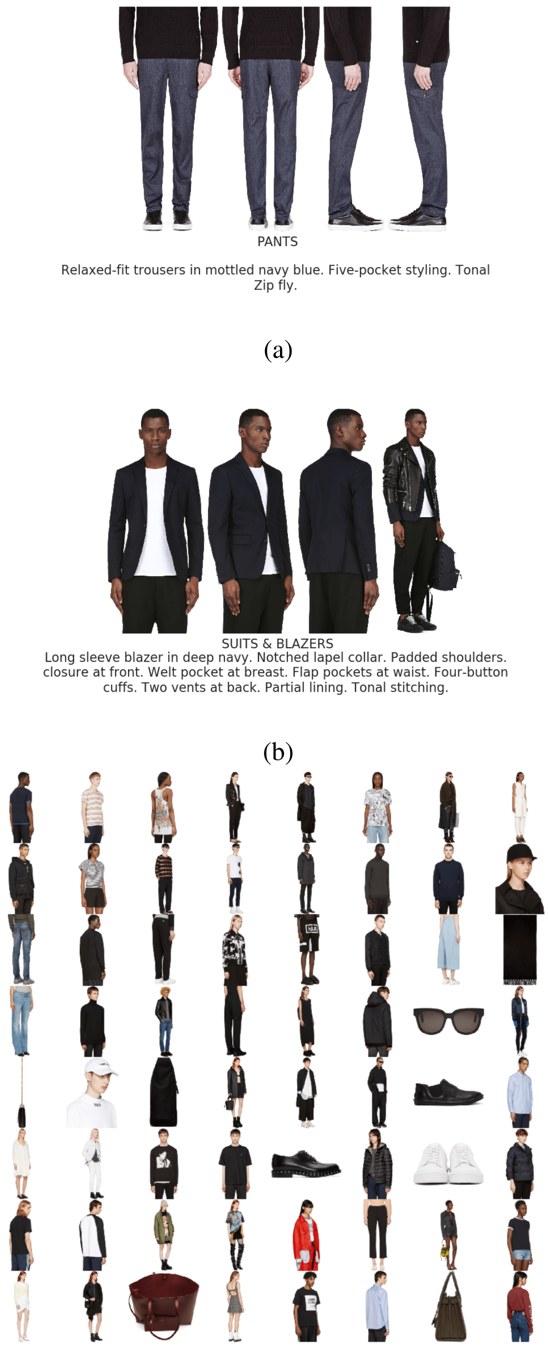

Figure 1. Pictures a, b and c present samples of the dataset. Each description is associated with all the images below it. And each item *ie. a, b* is photographed from different angles. We also provide each image's attributes, and its relationship to other objects in the dataset

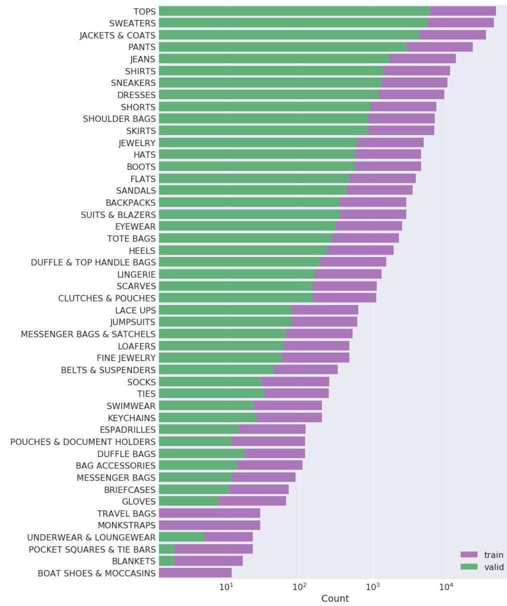

Figure 2. Distribution of the data per category. Note that the x axis is in log scale.

thesis (Zhang et al., 2017a; Huang et al., 2017; Reed et al., 2016a;b; Zhang et al., 2018).

Although state of the art generative models can already generate polished realistic images (Karras et al., 2017), conditional generation and translation tasks are still far from high quality. We hypothesize that this shortcoming is due to a shortage of large, clean datasets.

2.2. Related datasets

To the best of our knowledge none of the currently used datasets for text-to-image synthesis were collected specifically for the purposes of exploring the text-to-image synthesis task. Below we discuss existing datasets that have been used for text-to-image and attributes-to-image synthesis and focus on a specific set of attributes which are important for image synthesis.

Caltech-UCSD Birds-200-2011 (Reed et al., 2016a) was originally created for categorizing bird species, localizing their body-parts and classifying attributes. The dataset consists of 12k images , depicting 200 bird species with 28 attributes. More recent work employs this dataset for image synthesis tasks conditioned on text describing the attributes.
MS COCO (Lin et al., 2014) was originally created as a benchmark for image captioning. While some works use this dataset for text-to-image generation tasks, the generated images miss fine-grained details and only capture high-level information. This is due to the fact that the textual descriptions are very high-level.
Flowers Oxford-102 (Nilsback & Zisserman, 2008) con-

Table 1. Comparison of datasets

	Number of images	Resolution	Description	Binary attributes	Categories	Poses	Number of items
CelebA	202,599	43x55 to 6732x8984	no	40	no	multiple	10,177
CelebA-HQ	30,000	1024x1024	no	40	no	multiple	unknown
DeepFashion - Fashion Image Synthesis	78,979	300x300	multiple	1000	50	multiple	unknown
MS COCO	328,000	varying sizes	5 per image	no	80	single	unknown
Caltech-UCSD Birds-200-2011	11,788	varying sizes	no	312	200	single	unknown
Flowers Oxford-102	8189	varying sizes	no	no	102	single	unknown
Fashion dataset (ours)	325,536	1360x1360	yes	no	48	multiple	78850

sists of 102 categories of flowers and was proposed for the task of fine-grained image classification. (Reed et al., 2016a) collected 5 descriptions for each image in the dataset to augment it for the task of text to image generation.

CelebA (Liu et al., 2015) contains pictures of 10k celebrities, with 20 images per person (200k images in total). Each image in CelebA is annotated with 40 attributes.

DeepFashion (Liu et al., 2016) contains over 200k images downloaded from a variety of sources, with varying image sizes, qualities and poses. Each image is annotated with a range of attributes. This publicly available dataset was mainly employed for the task of cloth retrieval and classification. As an extension of the dataset on the task of text-to-image generation, 79k images from the dataset were later annotated with more descriptive text (Zhu et al., 2017b).

3. Our Fashion Dataset

The advantages of our new Fashion dataset over other contemporary datasets are as follows:

- The dataset consists of $293,008$ images ($260,480$ images for training, $32,528$ for validation, $32,528$ for test), which is larger than other available datasets for the task of text to image translation.
- We provide full HD images photographed under consistent studio conditions. There are no other datasets with comparable resolution and consistent photographing condition.
- All fashion items are photographed from 1 to 6 different angles depending on the category of the item. To our knowledge, this is the first dataset of this scale consisting of multiple angles of each item.
- Each product belongs to a main category and a more fine-grained category (*i.e: subcategory*). There are 48 main categories, and 121 fine-grained categories in the dataset. The name and density of each category is plotted in 2. Table 3 presents the number of images by category and subcategory.
- Each fashion item is paired with paragraph-length descriptive captions sourced from experts (*professional designers*). The distribution of the length of descriptions is presented in Figure 4.

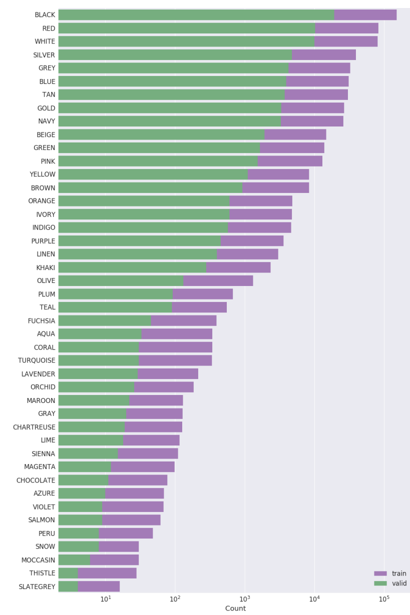

Figure 3. Distribution of the data based on colors. Note that the x axis is in log scale.

- For each item, we also provide metadata such as stylist-recommended matched items, the fashion season, designer and the brand. We also provide the distribution of colors extracted from the text description presented in Figure 3

4. Our Challenge

In addition to releasing a rich dataset, we are launching a challenge that uses our Fashion dataset for the task of text-to-image synthesis. To the best of our knowledge this is the first challenge on this task. Additionally, we encourage participants to take advantage of all information in the dataset, e.g. such as pose or category. We provide a framework that enables researchers to easily compare the performance of their models with an evaluation metric based on an *Inception Score* (Salimans et al., 2016). The inception model we use for the experiments we present in Section 5 was trained on the training set for classifying the images into the categories presented in Figure 2. For the final challenge evaluation we will also provide inception scores from a model trained on the test set. However, there are a number of issues to

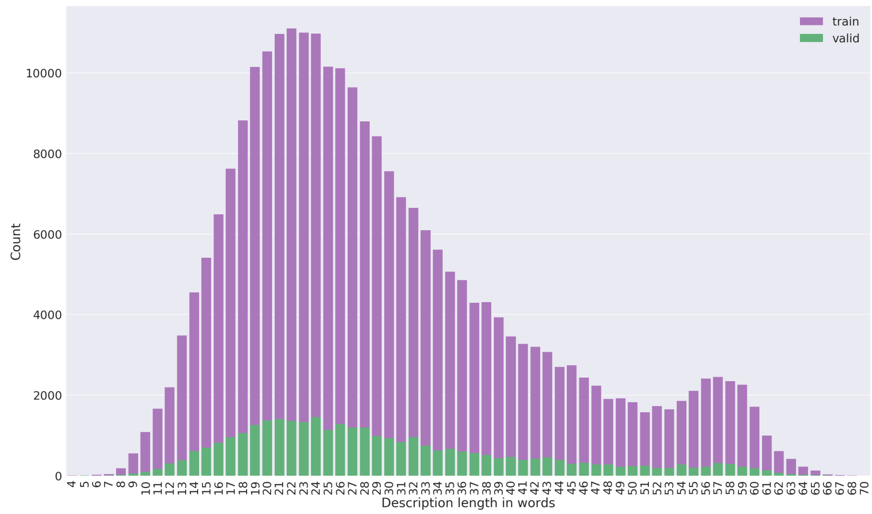

Figure 4. Distribution of the description lengths.

consider when using inception scores for evaluating generative models (Barratt & Sharma, 2018). For example, different implementations of the same model trained on the same dataset can result in significant differences in Inception scores. For these and other reasons our challenge will also provide a human evaluation as we outline below.

Our automated evaluation platform for the challenge computes and displays the Inception score for each submission and compiles the best scores in a leader-board. We provide a comprehensive template and an easy to use service to submit a docker container that runs code, and evaluate the performance on an Amazon Web Services cloud instance. Our test set, which won't be released, consists of descriptions of clothing items and is integrated at runtime in the challengers' docker container.

Human Evaluation setup:
Inception scores do not consider the correlation between text and the given image. As such, the competition results will also be evaluated by humans. Since inception scores also have other issues (Barratt & Sharma, 2018) as discussed above, the competition winner will be determined based on this human evaluation. During the human evaluation phase, a fixed subset of the test-set will be randomly selected and the corresponding images will be given to a human evaluation system. Each human-evaluator will be given a text and 5 images generated by each submission. The person's task will be to rank these sets of images into the first, second, and third best set with respect to the given text. Each task of this nature will be given to 10 different human-evaluators. The scores given to each image set will then be aggregated to compute final scores under the human evaluation.

5. Experiments with the Dataset

In this section, we present two sets of experiments: 1) Generating high-resolution images by using the progressive GAN (P-GAN) growing technique of Karras et al. (2017), and 2) text-to-image synthesis using StackGAN-v1 (Zhang et al., 2017a) and StackGAN-v2 (Zhang et al., 2017b).

5.1. Generating high-resolution images using P-GANs

The primary idea of Progressive Growing of GANs (Karras et al., 2017) is to grow the generator and discriminator gradually and in a symmetric manner in order to produce high-resolution images. P-GAN starts with very low-resolution images and each new layer of the model improves quality and adds fine-grained details to the image generated in the prior stage. Experiments on the CelebA dataset (Liu et al., 2015) showed promising results and we similarly employ P-GANs to generate 1024×1024 images using our fashion dataset as training data. To do this, we follow the same experimental setup and architectural details of the original P-GAN paper (Karras et al., 2017)[2]

[2]Using code provided by the authors of the P-GAN paper (Karras et al., 2017): https://github.com/tkarras/progressive_growing_of_gans

Figure 8 shows examples of images generated by P-GAN. The images exhibit global coherence and span a variety of poses and attributes ranging from color and category to accessory textures and characteristics of fashion designs.

In order to quantitatively evaluate the quality of our generated images, we compute the Inception score for the down-sampled version (256×256) of our generated images (See Table 2). The Inception score of the generated images using P-GANs is very close to the that of the original images, presented in Figure 5.

Figure 5. Images generated by the P-GAN approach (Karras et al., 2017)

5.2. Text-to-Image synthesis:

We employed two architectures: **StackGAN-v1** (Zhang et al., 2017a) and **StackGAN-v2** (Zhang et al., 2017b) to generate images conditioned on the their description.

StackGAN-v1 decomposes conditional image generation into two stages. First, the *Stage-I* GAN sketches a low resolution image (64×64) with the overall shape and colors of the image conditioned on the text and a random noise vector. Subsequently, the *Stage-II* GAN refines this low-resolution image conditioned on the results of the first stage and the same text embeddings, and generates a 256×256 image.

StackGAN-v2 follows a similar architecture consisting of multiple chained generators and discriminators. The input of each stage of the chain is the output of the previous stage. One of the major differences between StackGAN-v2 and StackGAN-v1 is that these stages are trained jointly, whereas in StackGAN-v1, they are trained independently.

In our experiments, we found that the method by which we encode the textual descriptions can indeed have a big impact on the quality of the generated images. Here, we discuss the text embedding that we applied.

Text embedding:
Both **StackGAN-v1** and **StackGAN-v2** condition the image generation process on φ_t, i.e. the text embedding of the corresponding image description generated from a pre-trained char-CNN-RNN encoder (Reed et al., 2016). It is important for the embedding of the description to correctly relate to the visual contents of the product image. We conducted our experiments using different encoders from a wide range of complexity, namely averaging word vectors, concatenating word vectors, a slightly modified encoder from the Transformer architecture (Vaswani et al., 2017) and a bidirectional LSTM (Schuster et al., 1997).

We experimented with both pre-training these models [3] and jointly training them with the GAN network. In the case of the Transformer's encoder and bi-LSTM, the text embedding φ_t is the output of the encoder of the Transformer, and the projected concatenation of the last hidden state of the forward and backward LSTM respectively. The final text embedding size for the Transformer is 1500 and 1024 for bi-LSTM.

We arrived at three conclusions based on our empirical experiments. First, we found that the bi-LSTM model achieves the highest category classification accuracy on the validation dataset in the pre-training process. As can be seen in Figure 9, the t-SNE (van der Maaten & Hinton, 2008) visualization of text embeddings shows relatively good separation of the categories. Secondly, we found that irrespective of the encoder architecture, pre-training the encoder model results in better correspondence between the descriptions and generated images. Finally, we found that overall, using the pre-trained bi-LSTM with fixed weights as the encoder leads to better results both visually and quantitatively.

The Inception scores reported in table 2 were obtained with the pre-trained bi-LSTM encoder (with fixed weights during the training of GAN).

Implementation details:
Throughout all the experiments, the descriptions were lowercased, tokenized and cleared of stop words[4]. We used the

[3]The pre-training step consisted of training the encoder to perform a classification task: given the item description predict its category.

[4]The python NLTK module was used to tokenize the descrip-

Figure 6. Images generated from the **StackGAN-v1** model with pre-trained bi-LSTM text encoder.

Figure 7. Images generated from the **StackGAN-v2** model with pre-trained bi-LSTM text encoder.

	Inception Score
Fashion Real data 256 × 256	9.71 ± 2.14
StackGAN-v1 (Zhang et al., 2017a)	6.50 ± 0.05
StackGAN-v2 (Zhang et al., 2017a)	5.54 ± 0.07
P-GAN (Karras et al., 2017)	7.91 ± 0.15

Table 2. Inception Scores on the validation set, i.e: trained on the Fashion train set.

first 15 tokens of the descriptions as the input sequence to the encoder model.

StackGAN-v1: We used the same overall architecture as (Zhang et al., 2017a) [5]. The first stage was trained for 80 epochs, and the second stage was trained for 185 epochs. The results can be seen in Figure 6.

StackGAN-v2: After careful experimentation, we ended up using the same architecture and hyper-parameters as (Zhang et al., 2017b) [6]. The results can be seen in Figure 7.

We can observe in the Table 2 that first of all, the Inception Score of the StackGAN-V1 is better than StackGAN-V2, while the quality of the images in the StackGAN-V2 is better and the reason is due to a significant mode-collapse that we were faced to in StackGAN-V2. Another interesting point, is the fact that most of the faces in StackGAN-V1 and StackGan-V2 are blurry. It suggests that since the images are conditioned on the text, the model is focusing more on clothing material than face information.

6. Conclusion

Recent progress in generative modeling techniques has great potential to give designers tools for rapidly visualizing and modifying ideas. While recent advances in generative models can be used to generate images of unprecedented realism, the quality of images generated from textual descriptions has so-far remained far from realistic. We believe that the lack of good datasets for this task has made it difficult to develop models for this task. In this paper, we have introduced a new Fashion themed text-to-image generation dataset, with high-quality images and extensive annotations provided by fashion experts. We provided results for 2 experiments: generating high-resolution images without providing textual descriptions as input, and generating realistic images conditioned on product description using the Fashion dataset as training data. We provide experiments with StackGAN-v1

tions by word and remove stop words

[5] We used the code provided by the authors of the StackGAN-v1 paper in github https://github.com/hanzhanggit/StackGAN-Pytorch

[6] We used the code provided by the authors of the StackGAN-v2 paper: https://github.com/hanzhanggit/StackGAN-v2

Figure 8. Generated images from the P-GAN approach (Karras et al., 2017)

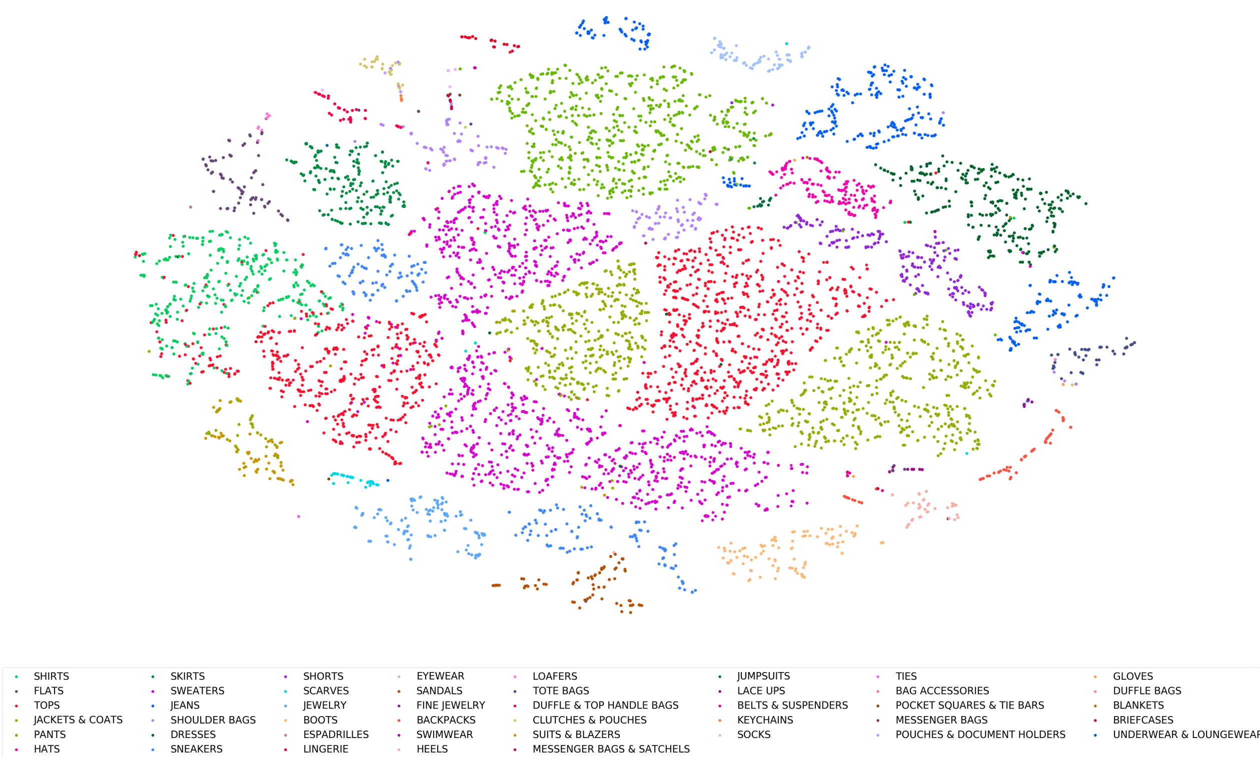

Figure 9. t-SNE visualization of the validation text embedding obtained by the bi-LSTM encoder

and StackGAN-v2 models using various text encoders.

To help stimulate further research on conditional generative models, we release our dataset as part of a challenge. Detailed submission instructions are provided and our API computes the inception score (trained on the Fashion dataset). Submissions with the highest quality images as judged by human evaluators will be selected as winners in the challenge organized around this new dataset.

7. Acknowledgement

We present our special thank to Alex Shee, for his help and support. We also thank Timnit Gebru and Archy de Berker, for assistance with comments that greatly improved the manuscript. We would also like to show our gratitude to Chelsea Moran, Valerie Becaert, Vincent Hoe-Tin-Noe, Misha Benjamin, Pedro Oliveira Pinheiro, David Vazquez, Francis Duplessis, Ishmael Belghazi, Caroline Bourbonniere and Xavier Snelgrove for their support and feedback during the course of this research. We would also like to thank SSENSE for open sourcing their data to the research community.

References

Barratt, S. and Sharma, R. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Belghazi, M. I., Rajeswar, S., Mastropietro, O., Rostamzadeh, N., Mitrovic, J., and Courville, A. Hierarchical adversarially learned inference. *arXiv preprint arXiv:1802.01071*, 2018.

Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., and Van Gool, L. Apparel classification with style. In *Asian conference on computer vision*, pp. 321–335. Springer, 2012.

Chen, H., Gallagher, A., and Girod, B. Describing clothing by semantic attributes. In *European conference on computer vision*, pp. 609–623. Springer, 2012.

Chen, Q., Huang, J., Feris, R., Brown, L. M., Dong, J., and Yan, S. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 5315–5324. IEEE, 2015.

Denton, E. and Fergus, R. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018.

Denton, E. L. et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pp. 4417–4426, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Han, X., Wu, Z., Wu, Z., Yu, R., and Davis, L. S. Viton: An image-based virtual try-on network. *arXiv preprint arXiv:1711.08447*, 2017.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 4, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Kalantidis, Y., Kennedy, L., and Li, L.-J. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pp. 105–112. ACM, 2013.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Kiapour, M. H., Han, X., Lazebnik, S., Berg, A. C., and Berg, T. L. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, pp. 3343–3351, 2015.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.

Liang, X., Lin, L., Yang, W., Luo, P., Huang, J., and Yan, S. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia*, 18(6):1175–1186, 2016.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., and Yan, S. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3330–3337. IEEE, 2012.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1096–1104, 2016.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pp. 722–729. IEEE, 2008.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016a.

Reed, S., Akata, Z., Lee, H., and Schiele, B. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49–58, 2016b.

Reed, S. E., Akata, Z., Schiele, B., and Lee, H. Learning deep representations of fine-grained visual descriptions. *CoRR*, abs/1605.05395, 2016. URL http://arxiv.org/abs/1605.05395.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Schuster, M., Paliwal, K. K., and General, A. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

Simo-Serra, E., Fidler, S., Moreno-Noguer, F., and Urtasun, R. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, volume 2, pp. 6, 2015.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., and Belongie, S. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pp. 4642–4650. IEEE, 2015.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699, 2015.

Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., and Metaxas, D. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 5907–5915, 2017a.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017b. URL http://arxiv.org/abs/1710.10916.

Zhang, Z., Xie, Y., and Yang, L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. *arXiv preprint arXiv:1802.09178*, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017a.

Zhu, S., Fidler, S., Urtasun, R., Lin, D., and Loy, C. C. Be your own prada: Fashion synthesis with structural coherence. *arXiv preprint arXiv:1710.07346*, 2017b.

Categories	Subcategories	Number of images
BACKPACKS	BACKPACKS	2858
BAG ACCESSORIES	BAG ACCESSORIES	110
BELTS & SUSPENDERS	BELTS & SUSPENDERS	335
BLANKETS	BLANKETS	17
BOAT SHOES & MOCCASINS	BOAT SHOES & MOCCASINS	12
BOOTS	ANKLE BOOTS	1975
	MID-CALF BOOTS	737
	CHELSEA BOOTS	562
	LACE-UP BOOTS	441
	TALL BOOTS	407
	ZIP UP & BUCKLED BOOTS	327
	DESERT BOOTS	30
	BIKER & COMBAT BOOTS	24
	WINGTIP BOOTS	12
BRIEFCASES	BRIEFCASES	71
CLUTCHES & POUCHES	POUCHES	640
	CLUTCHES	480
DRESSES	SHORT DRESSES	4819
	MID LENGTH DRESSES	3809
	LONG DRESSES	830
DUFFLE & TOP HANDLE BAGS	DUFFLE & TOP HANDLE BAGS	1533
DUFFLE BAGS	DUFFLE BAGS	120
ESPADRILLES	ESPADRILLES	119
EYEWEAR	SUNGLASSES	1706
	GLASSES	788
FINE JEWELRY	RINGS	215
	EARRINGS	200
	BRACELETS	36
	NECKLACES	24
FLATS	SLIPPERS & LOAFERS	1680
	BALLERINA FLATS	861
	LACE UPS & OXFORDS	684
	ESPADRILLES	630
GLOVES	GLOVES	66
HATS	CAPS & FLAT CAPS	1846
	BEANIES	1219
	STRUCTURED HATS	423
	FEDORAS & PANAMA HATS	367
	CAPS	360
	HEADBANDS & HAIR ACCESSORIES	174
	BEACH HATS	116
	AVIATOR	40
HEELS	HEELS	1931
JACKETS & COATS	COATS	7190
	JACKETS	6373
	BOMBERS	5864
	DOWN	5359
	DENIM JACKETS	3177
	LEATHER JACKETS	3080
	FUR & SHEARLING	1161
	BLAZERS	1111
	VESTS	794
	TRENCH COATS	685
	PEACOATS	241
JEANS	JEANS	13586
JEWELRY	BRACELETS	1926
	EARRINGS	1173
	RINGS	1014
	NECKLACES	794
	PINS	17
	BROOCHES	9
JUMPSUITS	JUMPSUITS	610
KEYCHAINS	KEYCHAINS	202
LACE UPS	LACE UPS	631

Categories	Subcategories	Number of images
LINGERIE	BRAS	636
	BRIEFS	259
	THONGS	252
	SLEEPWEAR	89
	BOY SHORTS	56
	TANKS	14
	ROBES	8
LOAFERS	LOAFERS	481
MESSENGER BAGS	MESSENGER BAGS	88
MESSENGER BAGS & SATCHELS	MESSENGER BAGS & SATCHELS	530
MONKSTRAPS	MONKSTRAPS	29
PANTS	TROUSERS	13590
	SWEATPANTS	5128
	LOUNGE PANTS	2565
	CARGO PANTS	677
	LEGGINGS	654
	LEATHER PANTS	616
POCKET SQUARES & TIE BARS	POCKET SQUARES & TIE BARS	23
POUCHES & DOCUMENT HOLDERS	POUCHES & DOCUMENT HOLDERS	121
SANDALS	FLAT SANDALS	1864
	HEELED SANDALS	937
	SANDALS	660
	FLIP FLOPS	54
SCARVES	KNITS	458
	SCARVES	432
	SILKS & CASHMERES	183
	FUR & SHEARLING	65
SHIRTS	SHIRTS	11398
SHORTS	SHORTS	7416
SHOULDER BAGS	SHOULDER BAGS	6952
SKIRTS	SHORT SKIRTS	3262
	MID LENGTH SKIRTS	3182
	LONG SKIRTS	465
SNEAKERS	LOW TOP SNEAKERS	7810
	HIGH TOP SNEAKERS	2620
	WEDGE SNEAKERS	48
SOCKS	SOCKS	250
SUITS & BLAZERS	BLAZERS	2560
	SUITS	171
	TUXEDOS	73
	WAISTCOATS	41
SWEATERS	CREWNECKS	13399
	SWEATSHIRTS	12331
	HOODIES & ZIPUPS	10388
	TURTLENECKS	3775
	CARDIGANS	3421
	V-NECKS	1261
	SHAWLNECKS	16
SWIMWEAR	BIKINIS	159
	SWIMSUITS	21
	ONE-PIECE	12
	COVER UPS	9
TIES	NECK TIES	240
	BOW TIES	13
TOPS	T-SHIRTS	33004
	SHIRTS	4563
	BLOUSES	3655
	TANK TOPS & CAMISOLES	3052
	POLOS	2117
	TANK TOPS	642
	BODYSUITS	526
	HENLEYS	205
TOTE BAGS	TOTE BAGS	2281
TRAVEL BAGS	TRAVEL BAGS	29
UNDERWEAR & LOUNGEWEAR	ROBES	17
	BOXERS	6

Table 3. Number of images per category and subcategory in the training set.