Vastr-GAN: Versatile Apparel Synthesised from Text using a Robust Generative Adversarial Network

Dhruvi Lodhavia*
IIT Gandhinagar, India
dhruvi.lodhavia@iitgn.ac.in

Hetvi Shastri*
IIT Gandhinagar, India
hetvi.shastri@iitgn.ac.in

Palak Purohit *
IIT Gandhinagar, India
palak.purohit@iitgn.ac.in

Ronak Kaoshik
IIT Gandhinagar, India
kaoshik.ronak@iitgn.ac.in

Nipun Batra
IIT Gandhinagar, India
nipun.batra@iitgn.ac.in

ABSTRACT

The development of the fashion industry has increased the demand for customised and meticulously designed clothes. This poses a challenge to fashion designers who need to create novel clothing designs based on the requirements specified by the customers. This work presents a generative adversarial network (GAN) based text-to-image synthesis model for fabricating intricate Indian apparel designs. We introduce an architecture that strategically combines multiple trained GAN models for a streamlined text-to-image generation. Existing fashion datasets with elaborate image descriptions cater to western fashion only. We have extracted traditional Indian images like kurtis, kurtas, etc., and then combined with an existing dataset to create an Indian Fashion dataset of around 16000 images with their corresponding text descriptions. On carrying out elaborate testing on our dataset we have achieved good visual results that can capture the details given in the text descriptions.

CCS CONCEPTS

• **Computing methodologies** → *Ensemble methods*; *Neural networks*; *Unsupervised learning*.

KEYWORDS

Text-to-image synthesis, Indian Fashion, GAN, Text encoder,Classifier

ACM Reference Format:
Dhruvi Lodhavia, Hetvi Shastri, Palak Purohit, Ronak Kaoshik, and Nipun Batra. 2022. Vastr-GAN: Versatile Apparel Synthesised from Text using a Robust Generative Adversarial Network. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD) (CODS-COMAD 2022), January 8–10, 2022, Bangalore, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3493700.3493721

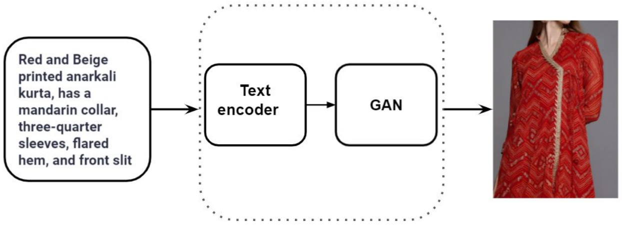

Red and Beige printed anarkali kurta, has a mandarin collar, three-quarter sleeves, flared hem, and front slit

Figure 1: Vastr-GAN can generate realistic images of Indian apparel, given a texual query

1 INTRODUCTION

India is a multicultural country with varied traditions and rituals which introduces diversity in our clothing style. There are numerous factors such as geographical location, social standard, age and culture which impact Indian fashion. Creativity in fashion wear is an amalgamation of various factors like colour combination, relevance to trends, and style. Therefore, it is a tedious task for designers to design according to customised demands.

One way to generate novel designs is via a textual input description fed to a generator of a Generative Adversarial Network (GANs) [2, 13, 14] . GANs are now being used widely to produce relevant output images based on an input pictorial and/or textual query. The basic framework of Vastr-GAN, as shown in Figure 1 demonstrates the text-to-image synthesis pipeline. The fashion industry is one sector that could primarily benefit from the effectiveness of GANs. One of the most recent techniques for text-to-image synthesis is DF-GAN [10]. It has shown significant results in terms of the realistic quality and resolution of the output image for the CUB Dataset [12]. Hence, for our work we have used DF-GAN baseline implementation.

Clothes commonly worn in India range from casual wear like t-shirts and shirts to traditional Indian attire, which includes kurtas and kurtis. For the purpose of dealing with diversity in the data, we trained multiple GAN models on different parts of the dataset. However, the final output is only one of these images, selected on the basis of the resemblance of the image to the category desired by the user. For the purpose of choosing one final output, we classified the synthesized images from each of these GAN models using an object detection classifier. Here, we have used the classifier to find the confidence scores of each image for the category of clothes desired. Since we have strategically combined different models

[*] These authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CODS-COMAD 2022, January 8–10, 2022, Bangalore, India
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8582-4/22/01…$15.00
https://doi.org/10.1145/3493700.3493721

trained on different parts of the dataset, we have termed it as an ensemble-like approach. The output of this ensemble framework is the image which most accurately resembles the class specified by the user, as shown in Figure 2.

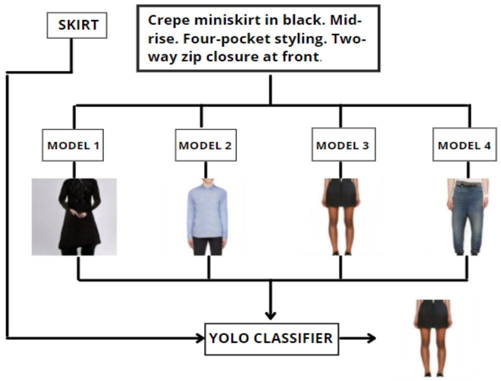

SKIRT
Crepe miniskirt in black. Mid-rise. Four-pocket styling. Two-way zip closure at front.
MODEL 1
MODEL 2
MODEL 3
MODEL 4
YOLO CLASSIFIER

Figure 2: Complete Framework including Ensemble Approach

We have curated a novel Indian fashion dataset with elaborate text descriptions that can potentially be used in future works relevant to Indian fashion. To the best of our knowledge, Vastr-GAN is the first such implementation that allows the addition of new classes in the dataset without retraining the GAN on the entire dataset from scratch. The newly added classes can be trained separately on the GAN and using the object detection classifier, we get a single output image from multiple models. Our work can be extended to other imbalanced and/or diverse datasets as well. The results obtained from our model have distinctive features and are consistent with the given textual description. We evaluated our results using Frechet Inception Distance (FID) and obtained an FID score of 55.9 on our dataset. [1]

2 RELATED WORK

We have described the work related to our paper in two significant domains below - text to image generation using GANs for different datasets and creating an Indian Fashion dataset [2, 7, 11].

Generating semantically consistent, photo-realistic images using textual descriptions is a popular research problem. Recently, GANs are being deployed as a means of solving this task. Reed et al. [8] uses GANs to generate 64*64 images according to the textual features but this method lacks intricate details and image clarity. StackGAN [14] uses two stages with multiple discriminator and generator blocks due to which the final output is dependent on the intermediate layers, which makes the model inefficient as noise gets accumulated at each stage. The work by Hao Dong et al. [1] uses a source image along with a target text description. It generates images that match the textual description and the output features are consistent with the input image. However, the resolution of the output image is highly compromised.

We now discuss previously done work related to Indian fashion. The Atlas dataset [11] consists of 186,150 images for clothing taxonomy but the dataset is imbalanced and does not emphasize on detailed captions. The IndoFashion dataset [7] consists of a 106K images with 15 different categories. However, lack of elaborate descriptions makes its usage limited to classification tasks. Generative Fashion for Indian Clothing [2] uses a dataset consisting of 12K images extracted from e-commerce websites. The dataset used has small length captions and is limited in size. We have extracted our own dataset for a wide range of classes with images having long and descriptive captions. The details have been included in Section 4.1. Generative Fashion for Indian Clothing [2] use GANs for generating images but they provide an image and a textual query as an input. Vastr-GAN unlike the prior work [2] takes in just a textual query as an input and provides an image as its output for Indian Fashion Dataset.

3 MODEL OVERVIEW

Vastr-GAN consists of three main parts: text encoder, Generative Adversarial Network, and object detection classifier. This section discusses the reason behind choosing a particular model for the three parts of our pipeline mentioned above.

3.1 Text Encoder

DF-GAN used in our implementation takes in a sentence vector and a random noise vector as its input. The text encoder extracts semantic vectors from captions and converts text to a sentence vector. In DF-GAN, the authors have used the text encoder implemented in Attn GAN [13]. It learns visually discriminative word features using Bi-directional Long Short Term Memory (Bi-LSTM) network. The attention mechanism in the text encoder computes the region context vector by finding attention weights corresponding to each word. Hence, we have used the text encoder of Attn GAN [13] for our implementation.

3.2 Generative Adversarial Network (GAN)

The basic structure of a Generative Adversarial Network (GAN) consists of a generator and a discriminator. For text-to-image synthesis, the generator generates fake images corresponding to the sentence vector. The discriminator distinguishes between real and fake images.

We thoroughly analysed existing GAN architectures for our application. The Stack GAN [14] framework consists of a 2-stage GAN architecture which makes it challenging to optimise the losses for both the stages simultaneously whereas the DF-GAN [10] uses a 1-stage GAN network. The salient features of DF-GAN is its Matching Aware Gradient Penalty (MA-GP) which improves the image quality, and deep fusion blocks which fuses text and image more effectively. Thus, we have used DF-GAN for our implementation.

3.3 Object detection Classifier

The object detection classifier extracts image features and using them, it recognises the class of the instance. For this purpose, we are using yolov5 [4] . It is implemented with PyTorch and has shown high values of mean average precision (mAP) and excellent performance over custom datasets.

[1] Link to Github repository : https://github.com/Dhruvi-Lodhavia/Vastr-GAN.git

Table 1: Our collected Indian fashion dataset has a large number of categories containing different number of images in each class. This results in a biased dataset which inspires the idea of the ensemble approach.

Class	Number of images	Gender
Kurtis	Women	849
Nehru-Jacket	Men	1118
kurtis	Men	1365
Jeans	Both	2000
Pants	Both	2000
Shorts	Both	2000
Skirts	Both	2000
Shirts	Both	2000
T-shirts	Both	2500
Total		15832

4 APPROACH

4.1 Dataset Collection

As described in Section 2, the unavailability of descriptive captions in existing datasets introduced the need to create our own dataset. We have obtained Traditional Indian clothes from the e-commerce website Myntra [6] by web-scraping using the selenium library. Along with the images, corresponding two to three lines of long text descriptions were also scraped. We extracted a total of 3332 images from Myntra. For the casual fashion apparel, we extracted 12500 images from the Fashion-gen dataset [9]. Table 1 lists the dataset distribution in classes and gender.

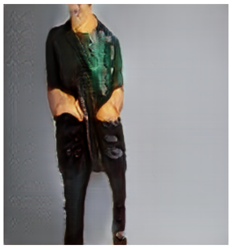


Figure 3: Image generated from short (1 line) texual information is not able to capture finer details. This inspired us to select the dataset with descriptive and long textual information

A major challenge faced was that GANs would not be able to capture the fine details in clothes in a short one line textual query. Initial training showed poor results for such captions as shown in Figure 3. Hence, we have chosen images with atleast 2-3 lines of description.

4.2 Motivation for training the Text encoder

The pretrained text encoder, trained on CUB [12] and COCO [5] dataset are not familiar with words frequently used in fashion - both Indian and Western - such as kurta, nehru-jacket, pant and jeans. If the sentence vector given as an input to the GAN is not trained well on the words used in Indian Fashion, it will lead to low text-image semantic consistency. We trained a text encoder using a bidirectional LSTM network on our custom dataset to solve this problem.

Table 2: Division of our Dataset into 4 models for training using visual diversity

Model Number	Classs Name	No. of images
Model 1	Shirt, Tshirt	4500
Model 2	Skirt, Shorts	4000
Model 3	Pant, Jeans	4000
Model 4	Kurtis, Kurtas, Nehru-jacket	3332

4.3 Ensemble-Inspired Approach

As seen in Table 1, there is an imbalance in the number of images per class, and there are significant differences in the features of traditional Indian wear like kurtis, and Indian casual wear like t-shirts. Further, the captions of the Indian casual wear dataset have more lines in their text description. Considering these factors, training the DF-GAN for all the classes simultaneously does not lead to good results. We observed that the model got skewed towards classes with highly descriptive captions and more images.

To tackle this problem, we derived the approach of ensemble learning which could cover the diverse characteristics of each class and finally give the result. For grouping the classes, we considered the visual diversity between different classes. As shown in Figure 4, we did the first branching based on the type of clothes, second branching based on whether it involves the upper or the lower body. Finally, we did the third branching according to the style. Thus, the dataset has been divided into 4 classes as shown in Table 2. Finally, we trained the resulting four models using DF-GAN. All four models receive the user given textual input. These models generate four different images for the given input caption. Using the YOLOv5 [4] model, we perform image classification on these four synthesised images. The final output is the image which has a maximum confidence score corresponding to the input class.

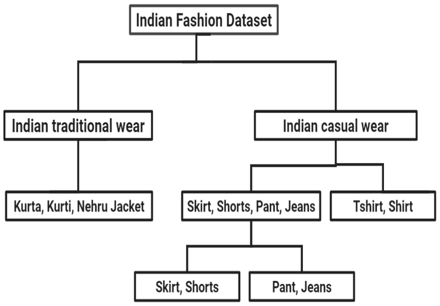
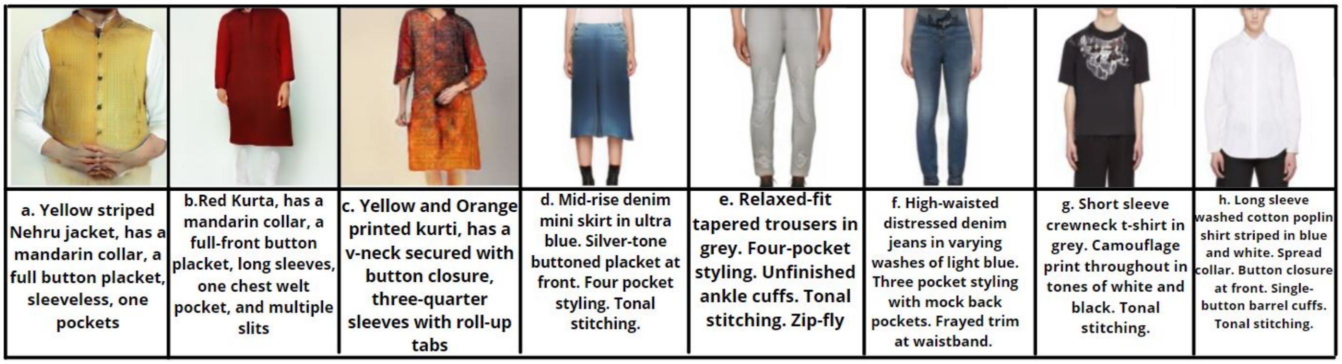


Figure 4: Hierarchical distribution of dataset for training different models based on visual diversity

Ensemble approach as presented in Figure 2, makes the framework more robust since it compares the output from multiple diverse and independent models, and returns the one which best resembles the desired class. There is also increased extendability,



a. Yellow striped Nehru jacket, has a mandarin collar, a full button placket, sleeveless, one pockets

b.Red Kurta, has a mandarin collar, a full-front button placket, long sleeves, one chest welt pocket, and multiple slits

c. Yellow and Orange printed kurti, has a v-neck secured with button closure, three-quarter sleeves with roll-up tabs

d. Mid-rise denim mini skirt in ultra blue. Silver-tone buttoned placket at front. Four pocket styling. Tonal stitching.

e. Relaxed-fit tapered trousers in grey. Four-pocket styling. Unfinished ankle cuffs. Tonal stitching. Zip-fly

f. High-waisted distressed denim jeans in varying washes of light blue. Three pocket styling with mock back pockets. Frayed trim at waistband.

g. Short sleeve crewneck t-shirt in grey. Camouflage print throughout in tones of white and black. Tonal stitching.

h. Long sleeve washed cotton poplin shirt striped in blue and white. Spread collar. Button closure at front. Single-button barrel cuffs. Tonal stitching.

Figure 5: Result images obtained on testing the Vastr-GAN models for various input text descriptions as shown below each image

since for inclusion of more classes or expansion of the dataset, another such DF-GAN model can be trained and included in the ensemble. The reductions of classes per model reduces the training time to a great extent and also decreases the chances of a skewed output due to biases in the dataset. Since each model is trained for a lesser number of classes, generation of sharper features for those classes becomes more probable. Hence, using this approach, we are able to leverage the diversity of generated images in each individual model, without compromising the resemblance of the image to the desired class.

5 EVALUATION

5.1 Experimental Settings

We trained the the descriptions of the images on the text Encoder for 600 epochs with a batch size of 48. We divided our custom dataset in the ratio 1:5 into test and train datasets and trained the four DF-GAN models with the corresponding batch size and epoch as mentioned in Table 3. The object detection classifier was trained on our dataset with pretrained checkpoints in YOLOv5x model. We performed the training for all our networks on Nvidia Tesla T4 GPU.

5.2 Metrics

The output images obtained can be analysed visually and by using metrics such as Frechet Inception Distance (FID) and Inception Score (IS). FID is used for measuring the quality of the generated images by comparing the statistics of generated and real images. This is done by extracting image features from an intermediate layer of the Inception model. A lower FID means better image quality and also diversity in the images generated [3]. FID Score has been used as an evaluation metric in DF-GAN, Stack GAN and Attn-GAN. Hence, we have used the same evaluation metric for measuring the performance of our trained model. The GANs mentioned above are not implemented on fashion datasets and therefore we have not included their results for comparison.

Table 3: DF-GAN Training Data

Attribute	Model 1	Model 2	Model 3	Model 4
Number of Epochs	215	210	160	450
Batch Size	24	24	24	24
Training Time	10 days	6 days	7 days	7days

Table 4: FID score of our 4 models

Model Number	FID Score
Model 1	36.7
Model 2	41.4
Model 3	56.4
Model 4	89.0

5.3 Results and Analysis

Figure 5 shows the clothing images generated for various input descriptions. The final FID score has been calculated as the average of the FID scores for each model. The FID scores for the 4 DF-GAN models have been shown in Table 4 . The value of the FID score of our whole dataset is 55.9. It can be seen in Table 4 that Model 4 has a higher FID score in comparison to the other models. This is due to presence of complicated Indian traditional designs and lack of descriptive captions on e-commerce websites for the images used in this model. This results in lesser diversity and poor quality as compared to other models.

6 LIMITATIONS AND CONCLUSION

Vastr-GAN does not generate images with proper text-image semantic consistency if the image captions are not descriptive enough as seen in Figure 3. This problem increases further in the generation of traditional Indian clothes using GANs since generally, they are more detailed and have more elaborate patterns than casual western clothes. However, the generated images have a high correlation with their captions and are diverse for different captions. This suggests that the model is capable of generating objects that have variety in terms of features and appearance.

REFERENCES

[1] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 5706–5714.

[2] Harshil Jain, Rohit Patil, Utsav Jethva, Ronak Kaoshik, Shaurya Agarawal, Ritik Dutta, and Nipun Batra. 2021. Generative fashion for Indian clothing.

[3] Neal Jean. 2018. *Fréchet Inception Distance | Neal Jean.* https://nealjean.com/ml/frechet-inception-distance/

[4] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. 2021. *ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations.* https://doi.org/10.5281/zenodo.4679653

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[6] Amar Nagaram. 2007. Myntra. www.myntra.com

[7] Pranjal Singh Rajput and Shivangi Aneja. 2021. IndoFashion: Apparel Classification for Indian Ethnic Clothes.

[8] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*. PMLR, 1060–1069.

[9] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-Gen: The Generative Fashion Dataset and Challenge. arXiv:1806.08317 [stat.ML]

[10] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. 2021. DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis. arXiv:2008.05865 [cs.CV]

[11] Venkatesh Umaashankar, Aditi Prakash, et al. 2019. Atlas: A Dataset and Benchmark for E-commerce Clothing Product Categorization.

[12] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.

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

[14] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. arXiv:1612.03242 [cs.CV]