DesIGN: Design Inspiration from Generative Networks

Othman Sbai[1,2] Mohamed Elhoseiny[2] Antoine Bordes[2]
Yann LeCun[2,3] Camille Couprie[2]

[1] Université Paris Est, Ecole des ponts, Imagine
[2] Facebook AI Research, [3] New York University
{sbaio,elhoseiny,abordes,yann,coupriec}@fb.com

Abstract. Can an algorithm create original and compelling fashion designs to serve as an inspirational assistant? To help answer this question, we design and investigate different image generation models associated with different loss functions to boost novelty in fashion generation. The dimensions of our explorations include: (i) different Generative Adversarial Networks architectures that start from noise vectors to generate fashion items, (ii) a new loss function that encourages novelty, and (iii) a generation process following the key elements of fashion design (disentangling shape and texture). A key challenge of this study is the evaluation of generated designs and the retrieval of best ones, hence we put together an evaluation protocol associating automatic metrics and human experimental studies. We show that our proposed creativity loss yields better overall appreciation than the one employed in Creative Adversarial Networks. In the end, about 61% of our images are thought to be created by human designers rather than by a computer while also being considered original per our human subject experiments, and our proposed loss scores the highest compared to existing losses in both novelty and likability.

Keywords: fashion image generation, generative adversarial networks

1 Introduction

Artificial Intelligence (AI) research has been making huge progress in the machine's capability of human level understanding across the spectrum of perception, reasoning and planning [1], [2], [3]. Another key direction yet relatively understudied is creativity where the goal is for machines to generate original items with realistic, aesthetic attributes, usually in artistic contexts. We can indeed imagine AI to serve as inspiration for humans in the creative process and also to act as a sort of assistant able to help with more mundane tasks, especially in the digital domain. Previous work has explored writing pop songs [4], imitating the styles of great painters [5], [6] or doodling sketches [7] for instance.

There has also been a growing interest in generating images using GANs, given their ability to generate appealing images unconditionally [8], or conditionally like from text, class labels, and for paired and unpaired image translations [9–11]. However, it is not clear how *creative* such attempts can be considered

since most of them mainly tend to mimic training samples without expressing much originality. Creative Adversarial Networks (CANs) [12] have then been proposed to adapt GANs to generate original content (paintings) by encouraging the model to deviate from existing painting styles. Technically, CAN is a Deep Convolutional GAN (DCGAN) model [13] associated with an entropy loss that encourages novelty against known art styles. The specific application domain of CANs (art paintings) allows for very abstract generations to be acceptable but, as a result, does reward originality a lot without judging much how such enhanced creativity can be mixed with realism and standards.

In this paper we study how AI can generate creative samples for fashion. Fashion is an interesting domain because designing original garments requires a lot of creativity but with the constraints that items must be wearable. In contrast to most generative models works [14–16], the originality angle we introduce makes us go beyond replicating images seen during training. Fashion image generation opens the door for breaking creativity into design elements (shape and texture in our case), which is a novel aspect of our work in contrast to CANs. More specifically, this work explores various architectures and losses that encourage GANs to deviate from existing fashion styles covered in the training dataset, while still generating realistic pieces of clothing without needing any image as input. To the best of our knowledge, this work is the first attempt at incorporating creative fashion generation by explicitly relating it to its design elements.

Contributions. (1) We are the first to propose a novelty loss on image generation of fashion items with a specific conditioning of texture and shape, learning a deviation from existing ones. (2) We re-purposed automatic entropy based evaluation criteria for assessment of fashion items in terms of texture and shape; The correlations between the automatic metrics that we proposed and our human study allowed us to *draw some conclusions with useful metrics revealing human judgment*. (3) We proposed a shape conditioned model named Style GAN and a concrete solution to make it work in a non-deterministic way. Trained with creative losses, it results in a *novel and powerful model*. Our best models manage to generate realistic images with high resolution 512×512 using a relatively small dataset (about 4000 images). More than 60% of our generated designs are judged as being created by a human designer while also being considered original, showing that an AI could offer benefits serving as an efficient and inspirational assistant.

2 Models: architectures and losses

2.1 Network architectures

We experiment using two architectures: a modified version of the DCGAN model [13] for higher resolution output images, and our proposed styleGAN model as described below. In addition to its real/fake branch classification, the discriminator in each architecture is augmented with optional classification branches each for shape and texture classes.

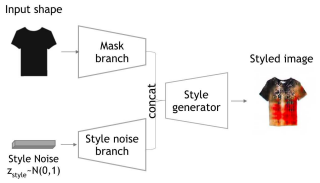

Fig. 1. From the segmented mask of a fashion item and different random vector z, our StyleGAN model generates different styled images.

Fig. 2. From the mask of a product, our StyleGAN model generates different styled image for each style noise.

GANs with optional classification loss. Let $\mathcal{D}$ be a dataset of N images. Following [10], we use shape and texture labels to learn a shape classifier and a texture classifier in the discriminator. Adding these labels improves over the plain model and stabilizes the training for larger resolution. We are adding to the discriminator network either one branch for texture or for shape classification, or two branches for both shape and texture classification and denote the extra classification output of the discriminator D_b. The additional loss is:

$$\mathcal{L}_{D\ \text{classif}} = -\sum_{x_i \in \mathcal{D}} \log(\text{softmax}(D_b(x_i)). \tag{1}$$

StyleGAN: Conditioning on masks. In this model, a generator is trained to compute realistic images from a mask input and noise representing style information (Fig. 1). We use the same discriminator architecture as in DCGAN with classifier branches that learn shape and texture classification on real images on top of real/fake prediction. Training styleGAN with two inputs is difficult, previous approaches of image to image translation such as pix2pix [17] and CycleGAN [11] create a deterministic mapping between an input image to a single corresponding one, i.e. edges to handbags for example or from one domain to another. To make sure that no input is being neglected, we add a ℓ_1 loss forcing the generator to output the mask itself in case of null style input z and thus ensure the impact of the shape in the generations as shown in Fig. 1.

2.2 Novelty losses

Because GANs learn to generate images very similar to the training images, we explore ways to make them deviate from this replication by studying the impact of two additional losses for the generator: the *CAN loss* (as used in [12]), and an *MCE loss* that encourage the generator to confuse the discriminator.

– **CAN loss**: As proposed in [12], the CAN loss is defined as

$$\mathcal{L}_{\text{CAN}} = -\lambda \left[\sum_i \sum_{k=1}^{K} \frac{1}{K} \log(\sigma(D_{b,k}(G(z_i)))) + \frac{K-1}{K} \log(1 - \sigma(D_{b,k}(G(z_i)))) \right], \tag{2}$$

where σ is the sigmoid function, and K the number of texture, shape, or both classes.

– **MCE loss**: We propose to use as alternative additional generator's loss the Multi-class Cross Entropy (MCE) loss between the class prediction of the discriminator and the uniform probability vector.

$$\mathcal{L}_{\text{MCE}} = -\lambda \sum_i \sum_k \frac{1}{K} \log \text{softmax}(D(G(z_i))). \tag{3}$$

Both MCE and sum of binary cross entropy losses encourage deviation from existing categories. However, our MCE criterion considers all classes globally in the softmax unlike the CAN loss which is based on a sum of K independent binary classification losses.

3 Results

Dataset. Unlike similar work focusing on fashion item generation [16, 15], we choose a dataset containing fashion items in uniform background allowing the trained models to learn features useful for creative generation without generating wearer faces and backgrounds. We augment the dataset of 4157 images by a factor 5 by jittering images with random scaling and translations. The images are classified into seven clothes categories: jackets, coats, shirts, tops, t-shirts, dresses and pullovers, and 7 textures categories: uniform, tiled, striped, animal skin, dotted, print and graphical pattern.

Automatic evaluation metrics. Evaluating the diversity and quality of a set of images has been tackled by scores such as the inception score and variants like the AM score [18]. We adapt both of them for two labels specific to fashion design (shape and texture) and supplement them by a mean nearest neighbor distance. Our final set of automatic scores contains 5 metrics : (1,2) shape score and texture score, each based on a Resnet-18 classifier [19] of shape or texture respectively. (3,4) shape AM score and texture AM score, based on the output of the same classifiers. (5) mean distance to 10 nearest neighbors score. We compute the mean distance for each sample to its retrieved k-Nearest Neighbors (NN), with $k = 10$, as the Euclidean distance between the features extracted from a Resnet18 pre-trained on ImageNet by removing its last fully connected layer.

Creating evaluation sets. We select for each setup (DCGAN or styleGAN trained with texture, shape, both or none novelty criterion) four saved models after a sufficient number of iterations. Our models produce plausible results after training for 15000 iterations with a batch size of 64 images.

Given a set of 10000 generations from a model, we extract different sets of images with particular visual properties such as (ii) high/low texture entropy, (iii) high/low NN distance to real images. We also explore random and mixed sets such as *low shape entropy* and *high nearest neighbors distance*. We expect such a set to contain plausible generations since low shape entropy usually correlates with well defined shapes, while high nearest neighbor distance contains unusual

designs. Overall, we have 8 different sets that may overlap. We choose to evaluate 100 images for each set.

Automatic evaluation results We set $\lambda = 5$ for the MCE loss, and $\lambda = 1$ for the CAN loss, as these parameters appeared to work best. All models were trained using the default learning rate 0.002 as in [13]. Our different models take about half a day to train on 4 Nvidia P100 GPUs for 256×256 models and almost 2 days for the 512×512 ones.

Table 1 presents shape and texture scores, AM scores (for shape and texture) and average NN distances computed for each model on 4 selected iterations. Our first observation is that the DCGAN model alone seems to perform worse than all other tested models with the highest NN distance and lower shape and texture scores. The value of the NN distance score may have different meanings. A high value could mean an enhanced "creativity" of the model, but also a higher failure rate. The two models having high shape score, AM shape score, AM texture score and NN distances scores are DCGAN with creativity losses models.

Method/Score	shape	tex.	AM sh	AM tx	NN
Dataset	6.25	3.76	20.4	12.6	5.65
GAN	4.70	2.74	13.3	8.92	**14.4**
GAN classif	**5.31**	2.86	**14.8**	**9.68**	13.1
CAN shape	**5.27**	2.77	**14.7**	8.92	13.1
CAN tex	**5.24**	3.01	**14.4**	**9.48**	**13.5**
CAN shTex	5.20	**3.24**	**14.7**	**10.0**	13.1
MCE shape	5.07	2.80	13.6	8.90	13.0
MCE tex	5.14	**3.33**	**14.4**	**9.30**	**13.6**
MCE shTex	4.98	3.04	13.3	**9.52**	13.2

Table 1. Quantitative automatic evaluation. High scores appear in bold.

Method/Human Method	over-all	shape nov.	shape comp.	tex. nov.	tex. comp.	real fake
DCGAN MCE shape	**3.78**	**3.58**	**3.57**	**3.64**	**3.57**	**60.9**
DCGAN MCE tex	**3.72**	**3.57**	3.52	**3.61**	**3.58**	**61.1**
StyleGAN CANtex	3.65	3.37	3.31	3.44	3.21	49.7
StyleGAN MCE tex	3.61	3.38	3.29	3.50	3.37	53.4
StyleGAN	3.59	3.28	3.21	3.27	3.15	47.2
DCGAN MCEshtex	3.49	3.40	3.24	3.40	3.31	**61.3**
DCGAN CANshtex	3.47	3.28	3.18	3.33	3.16	**63.8**
DCGAN classif	3.42	3.32	3.32	3.37	3.29	52.7
DCGAN CANtex	3.37	3.23	3.12	3.35	3.09	59.7
DCGAN CANshape	3.33	3.28	3.16	3.27	3.12	55.0
DCGAN	3.22	2.95	2.78	3.24	2.83	60.4

Table 2. Human evaluation ranked by decreasing overall score (higher is better).

Human evaluation. Each image was rated by 5 persons asked to answer 6 questions: Q1: how do you like this design overall on a scale from 1 to 5? Q2/Q3: rate the novelty of shape (Q2) and texture (Q3) from 1 to 5. Q4/Q5: rate the complexity of shape (Q4) and texture (Q5) from 1 to 5.Q6: Do you think this image was created by a fashion designer or generated by computer? (yes/no).

Table 2 presents the average score obtained by each model on each human evaluation question for the RTW dataset. From this table, we can see that using our novelty loss (MCE shape and MCE tex) performs better than the DCGAN baseline. While the two proposed models with MCE originality loss rank the best on the overall score, we observe that the preferred images have low nearest neighbor distance. This means that generations which are not close to their nearest neighbors are not always pleasant. It is indeed a challenge to obtain models able to generate novel (high nearest neighbor distance) and at the same time pleasant generations. However, we observe that the models that score better in the high nearest neighbors distance set are clearly the ones with our novelty loss(MCE). Fig. 3 shows how well our approaches worked on two axis: likability and real appearance. The most popular methods are obtained by the models employing an originality loss and

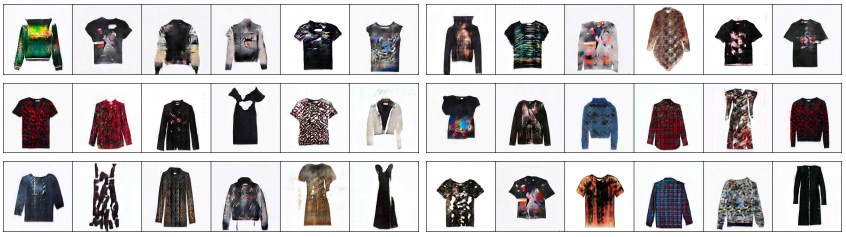

Fig. 4. Best generations as rated by annotators. Left: Q1: overall score, Q2: shape novelty, Q3: shape complexity; Right: Q4: tex. novelty, Q5: tex. complexity, Q6: Realism.

in particular our proposed MCE originality criterion, as they are perceived as the most likely to be generated by designers, and the most liked overall. We are greatly improving the state-of-the-art here, going from a score of 64 to more than 75 in likeability from classical GANs to our best model with shape creativity. We display images which obtained the best scores for each of the 6 questions in Fig. 4. Our proposed Style GAN (See Fig. 2) is producing competitive scores compared to the best DCGAN setups. In particular, StyleGAN with originality loss is ranked in the top-3.

We computed correlation scores between our automatic metrics and human ratings. The metric that correlates the most with the overall score is the NN distance. There is also a negative correlation of NN dist with real appearance.

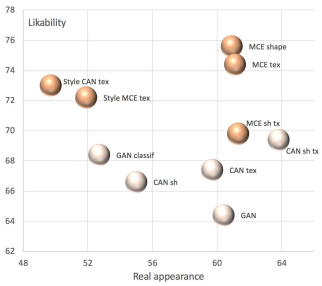

Fig. 3. Evaluation of the different models on the RTW dataset by human annotators on two axis: likability and real appearance. Our models reach nice trade-offs between real appearance and likability.

4 Conclusion

We introduced a specific conditioning of GANs on texture and shape elements for generating fashion design images. While GANs with such classification loss offer realistic results, they tend to reconstruct the training images. Using an MCE originality loss, we learn to deviate from a reproduction of the training set. We also propose a novel architecture named *StyleGAN model*, conditioned on an input mask, enabling shape control while leaving free the creativity space on the inside of the item. All these contributions lead to the best results according to our human evaluation study. We manage to generate accurately 512×512 images, however we seek for better resolution, which is a fundamental aspect of image quality, in our future work. Finally, while our results show visually pleasing textural novelty, it will be interesting to explore larger families of novelty loss functions, and ensure wearability constraints.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. ICCV (2017)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: CVPR. (2016)
3. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. Nature (2016)
4. Briot, J.P., Hadjeres, G., Pachet, F.: Deep learning techniques for music generation-a survey. arXiv:1709.01620 (2017)
5. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. (2016)
6. Dumoulin, V., Shlens, J., Kudlur, M., Behboodi, A., Lemic, F., Wolisz, A., Molinaro, M., Hirche, C., Hayashi, M., Bagan, E., et al.: A learned representation for artistic style. ICLR (2017)
7. Ha, D., Eck, D.: A neural representation of sketch drawings. ICLR (2018)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014)
9. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NIPS. (2016)
10. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: ICML. (2017)
11. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV (2017)
12. Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M.: Creative adversarial networks. In: ICCC. (2017)
13. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. ICLR (2016)
14. Date, P., Ganesan, A., Oates, T.: Fashioning with networks: Neural style transfer to design clothes. In: KDD ML4Fashion workshop. (2017)
15. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: Fashion synthesis with structural coherence. ICCV (2017)
16. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model of people in clothing. ICCV (2017)
17. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
18. Zhou, Z., Zhang, W., Wang, J.: Inception score, label smoothing, gradient vanishing and -log(d(x)) alternative. arXiv:1708.01729 (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 (2015)