Generating High-Resolution Fashion Model Images Wearing Custom Outfits

Gökhan Yildirim Nikolay Jetchev Roland Vollgraf Urs Bergmann
Zalando Research
{gokhan.yildirim,nikolay.jetchev,roland.vollgraf,urs.bergmann}@zalando.de

Abstract

Visualizing an outfit is an essential part of shopping for clothes. Due to the combinatorial aspect of combining fashion articles, the available images are limited to a predetermined set of outfits. In this paper, we broaden these visualizations by generating high-resolution images of fashion models wearing a custom outfit under an input body pose. We show that our approach can not only transfer the style and the pose of one generated outfit to another, but also create realistic images of human bodies and garments.

1. Introduction

Fashion e-commerce platforms simplify apparel shopping through search and personalization. A feature that can further enhance user experience is to visualize an outfit on a human body. Previous studies focus on replacing a garment on an already existing image of a fashion model [5, 2] or on generating low-resolution images from scratch by using pose and garment color as input conditions [8]. In this paper, we concentrate on generating high-resolution images of fashion models wearing desired outfits and given poses.

In recent years, advances in Generative Adversarial Networks (GANs) [1] enabled sampling realistic images via implicit generative modeling. One of these improvements is Style GAN [7], which builds on the idea of generating high-resoluton images using Progressive GAN [6] by modifying it with Adaptive Instance Normalization (AdaIN) [4]. In this paper, we employ and modify Style GAN on a dataset of model-outfit-pose images under two settings: We first train the vanilla Style GAN on a set of fashion model images and show that we can transfer the outfit color and body pose of one generated fashion model to another. Second, we modify Style GAN to condition the generation process on an outfit and a human pose. This enables us to rapidly visualize custom outfits under different body poses and types.

2. Outfit Dataset

We use a proprietary image dataset with around 380K entries. Each entry in our dataset consists of a fashion model wearing an outfit with a certain body pose. An outfit is composed of a set of maximum 6 articles. In order to obtain the body pose, we extract 16 keypoints using a deep pose estimator [10]. In Figure 1, we visualize a few samples from our dataset. The red markers on the fashion models represent the extracted keypoints. Both model and articles images have a resolution of 1024×768 pixels.

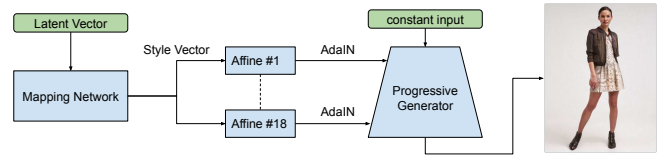
Figure 1: Samples from our dataset (red markers represent the extracted keypoints).

3. Experiments

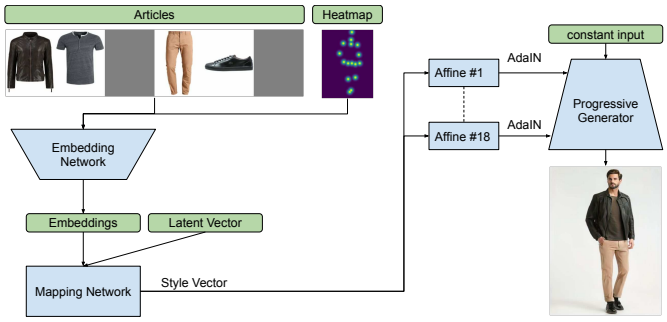
The flowchart for the unconditional version of Style GAN is illustrated in Figure 2(a). We have 18 generator layers that receive an affinely transformed copy of the style vector for adaptive instance normalization. The discriminator is identical to the original Style GAN. We train this network for around four weeks on four NVIDIA V100 GPUs, resulting in 160 epochs.

In the conditional version, we modify Style GAN with an embedding network as shown in Figure 2(b). Inputs to this network are the six article images (in total 18 channels) and a 16-channel heatmap image that is computed from 16 keypoints. The article images are concatenated with fixed ordering for semantic consistency across outfits. We can see this ordering in Figure 1. If an outfit does not have an article on a particular semantic category, it is filled with an empty gray image. The embedding network creates a 512-dimensional vector, which is concatenated with the latent vector in order to produce the style vector. This model is also trained for four weeks (resulting in 115 epochs). The discriminator in the conditional model uses a separate network to compute the embedding for the input articles and

heatmaps, which is then used to compute a final score using [9].



(a) Unconditional



(b) Conditional

Figure 2: The flowcharts of our (a) unconditional and (b) conditional GANs.

3.1. Unconditional

In Figure 3, we illustrate images that are generated by the unconditional model. As we can see, not only the articles, but also the human body parts are realistically generated at the maximum resolution of 1024×768 pixels.

During the training, one can regularize the generator by switching the style vectors for certain layers. This has the effect of transferring information from one generated image to another. In Figure 4, we illustrate two examples of information transfer. First, we broadcast the same source style vector to layers 13 to 18 (before the affine transformations in Figure 2) of the generator, which transfers the color of the source outfit to the target generated image, as shown in Figure 4. If we copy the source style vector to earlier layers, this transfers the source pose. In Table 1, we show which layers we broadcast the source and the target style vectors to achieve the desired transfer effect.

	Color Transfer	Pose Transfer
Source	13-18	1-3
Target	1-12	4-18

Table 1: Layers to broadcast the style vector.

3.2. Conditional

After training our conditional model, we can input a desired set of articles and a pose to visualize an outfit on a human body as shown in Figure 5. We use two different outfits in Figure 5(a) and (b), and four randomly picked body poses to generate model images in Figure 5(c) and (d), respectively. We can observe that the articles are correctly rendered on the generated bodies and the pose is consistent across different outfits. In Figure 5(e), we visualize the generated images using a custom outfit by adding the jacket from the first outfit to the second one. We can see that the texture and the size of the denim jacket are correctly rendered on the fashion model. Note that, due to the spurious correlations within our dataset, the face of a generated model might vary depending on the outfit and the pose.

In our dataset, we have fashion models with various body types that depend on their gender, build, and weight. This variation is implicitly represented through the relative distances between extracted keypoints. Our conditional model is able to capture and reproduce fashion models with different body types as shown in the fourth generated images in Figure 5. This result is encouraging, and our method might be extended in the future to a wider range of customers through virtual try-on applications.

3.3. Quantitative Results

We measure the quality of the generated images by computing the Fréchet Inception Distance (FID) score [3] of the unconditional and conditional GANs. As we can see from Table 2, the unconditional GAN produces higher quality images, which can be observed by comparing Figure 3 and Figure 5. The conditional discriminator has the additional task of checking whether the input outfit and pose are correctly generated. This might cause a trade-off between image quality (or 'realness') and the ability to directly control the generated outfit and pose.

	FID Score	Training Epochs
Unconditional	5.15	115
Conditional	9.63	115

Table 2: FID Score for the models.

4. Conclusion

In this paper, we proposed two ways to generate high-resolution images of fashion models. First, we showed that the unconditional Style GAN can be used to transfer the style/color and the pose between generated images via swapping the style vectors at specific layers. Second, we modified Style GAN with an embedding network, so that we can generate images of fashion models wearing a custom outfit with a give pose. As future work, we plan to improve the image quality and consistency of the conditional model on more challenging cases, such as generating articles with complicated textures and text.

Figure 3: Model images that are generated by the unconditional Style GAN.

Figure 4: Transferring the colors of an outfit or a body pose to a different generated model.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*. 2014.

[2] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. *CVPR*, 2017.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*. 2017.

[4] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *ICCV*, 2017.

[5] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *ICCV Workshops*, 2017.

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2017.

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, 2019.

[8] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. In *ICCV*, 2017.

[9] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018.

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

(a) Outfit #1

(b) Outfit #2

(c) Generated model images with outfit #1

(d) Generated model images with outfit #2

(e) Generated model images with outfit #2 and the jacket from outfit #1

Figure 5: Two different outfits (a) and (b) are used to generate model images in (c) and (d). (e) The jacket from outfit #1 is added to outfit #2 to customize the visualization.