Diffusion Models for Outfit Rendering: Novel Conditioning Architectures for Subject-driven Generation

Anonymous ECCV submission

Paper ID 100

Abstract. Generating images of digital fashion models dressed with a curated outfit has various applications especially when these fashion models can be conditioned on different poses, body sizes, etc. In this paper, we propose novel conditioning architectures for diffusion models for generating curated outfits to be rendered on a digital human in predefined pose. The conditioned outfits are fed through information pathways including learned deepset embedding and cross-attention with pose skeleton, allowing for a strong conditioning signal for subject-driven generation. Such an *outfit renderer* a) allows to scale fashion imagery to millions of outfit combinations b) enables unprecedented access to creative control over studio content generation c) provides high level of personalization because users could explore or complete outfits on-demand and also in their own likeness d) a stepping stone towards 2D virtual try on which unlike 3D virtual try-on does not require dedicated hardware.

Keywords: Diffusion models, outfit renderer, cross-attention

1 Introduction

Generating images of fashion models wearing a curated outfit in desired poses has the potential to disrupt the fashion e-commerce industry. A tool that allows circumventing the need for a human model physically wearing the outfit combination would allow unprecedented creative control over studio fashion imagery. Such a tool enables retail firms to onboard their articles to online platforms more easily. Synthesizing images with desired content is a challenging computer vision task. Generative models have been widely utilized to address this problem. In recent years, diffusion models have achieved state-of-the-art performance on various tasks including conditional image generation [16], [7], [17], [18], [11], [3]. Text-to-image models have been shown to generate high quality images [15], [12], [13], but they lack the ability to drive the image towards a specific input conditional image. [1] show that models can be forced to generate only over a certain region in the image by providing a mask, but this still does not allow for generating a given conditional image.

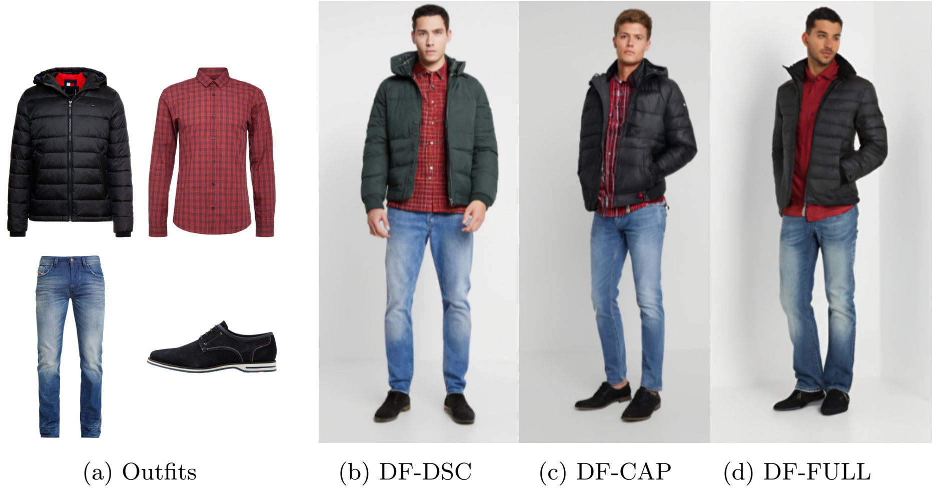


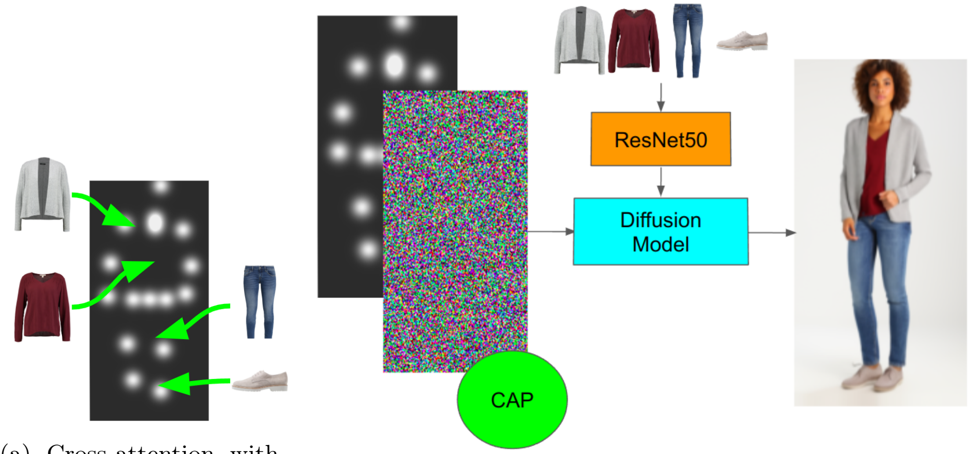
Fig. 1: Sample results from our method DF are shown here. a) shows the input conditional outfits to the DM. b-d) shows resulting images from the different methods. We find that results from DF-FULL are more faithful to the input conditions capturing the fine textural details.

Subject-driven generation with diffusion models (DM) has gained more popularity recently [14], [4]. However, the above methods assume some amount of correlation of contexts with the pretrained large text-to-image models. This is not suitable for problems such as studio fashion photography. The challenge here is to not only generate the human in a predetermined pose, but also be strictly faithful while overlaying the given conditional outfits on top of the human. The final image should capture the full body of the human in a given pose and also represent the fluidity of the outfits overlaid.

Previous works on the outfit renderer problem primarily used a generative adversarial networks (GAN) based architecture [9], [21], but this has often resulted training instability and also catastrophic mistakes during inference. In this work, we present Dif-Fashion (DF), a diffusion model-based fashion outfit renderer. We effective show ways of utilizing guided DM for our particular use-case of subject-driven generation of fashion images. We demonstrate several key mechanisms in the architecture of DM that allow for reconstructing the input garment image on the rendered digital human.

Our main contributions in this paper can be briefly summarized as follows:

- Diffusion model-based outfit renderer—generating a digital human in a desired pose and wearing a curated set of outfits, while being faithful to the texture and semantics of the given outfits.
- Composing multiple outfits with deep set embeddings.
- Blob attention to capture the fluidity of the outfit on the digital human.
- Simulating classifier-free guidance with article dropping during training.



(a) Cross-attention with Pose (CAP)

(b) Overall architecture

Fig. 2: a) Illustration of the allowed information flow in our cross-attention mechanism is shown here. The attention values in the dot product (shown as green arrows here) will effectively copy information from the different spatial regions of the four outfits and allow more information conditioning for the diffusion process. b) The overall architecture is shown here. The input pose, CAP output and the noise are concatenated and fed as input to the DM, while the deepset embeddings from ResNet50 are fed through the adaptive group normalization layers in the autoencoding structure.

Being faithful to the given input(s) images is a non-trivial problem. We show that a diffusion model-based outfit renderer produces images that accurately capture the semantics as well as the fluidity of the outfits while generating high quality digital humans paving the way for a revolution in fashion studio photography. Sample results from our different methods can be found in Figure 1.

2 Dif-Fashion (DF)

DF utilizes conditional DM for the outfit renderer use-case. The goal of this work is to generate a digital human in a desired pose and wearing a curated set of outfits. The DM here is trained using pairs of outfits and a real human model wearing these outfits $(A_k, \boldsymbol{x})$, where $\boldsymbol{x}$ is the image of a real human model and A_k is the variable number of outfits. We feed the input conditions of pose and outfits to the DM.

2.1 Pose Conditioning

To be able to generate digital humans with a desired pose, we train the DM by conditioning it on the pose. 16 keypoints extracted for the ground-truth images from a pretrained pose estimating method [19]. We pass the pose $P \in \mathbb{R}^{16 \times H \times W}$, where H and W are the height and width of the images, by concatenating them as 16 channels with the input to the autoencoding structure of the DM.

2.2 Outfit Conditioning

The set of outfits should influence and strongly condition the final image generated by the DM. A set of K outfits $\{A_k\}_{k=0}^{K}$, each is an image tensor of size $\mathbb{R}^{C \times H \times W}$. We have explored two approaches to making this conditioning work.

Deepset Conditioning (DSC) In order to process a variable number of outfits in the outfit, we use a deepset architecture [22]. First, we process each outfit packshot image by an encoder in the form of a Resnet50 model [5] pretrained on the ImageNet dataset [2] and further finetuned on the downstream dataset. Each outfit A_k is thus represented by an embedding $a_k \in \mathbb{R}^{1000}$. Then the deepset processes these and makes the final outfit representation $a' \in \mathbb{R}^{1000}$, where $a' = \frac{1}{K}\sum_i^K a_k$. This deepset embedding a', is then combined with the typical time embedding in the adaptive group normalization layer of autoencoding structure in the DM allowing for faithfulness of the generated image to the variable number of input outfits. A schematic of this architecture shown in Figure 2b.

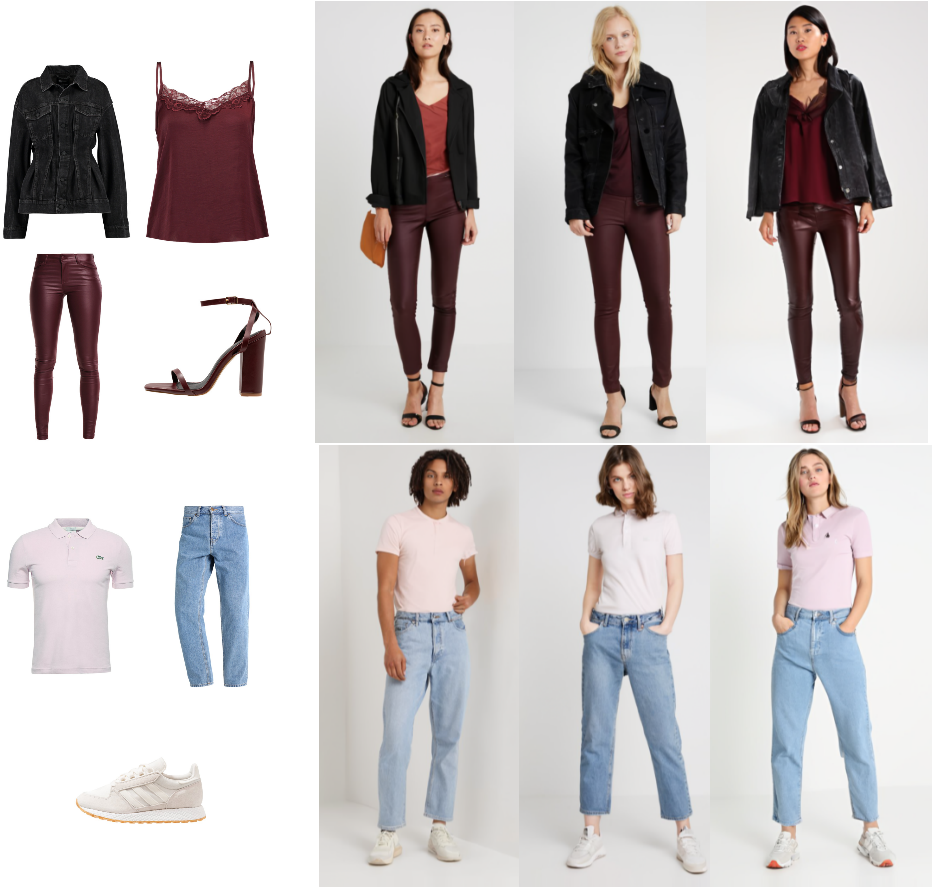
Cross-Attention with Pose (CAP) The previous mechanism acts as a feature extractor collapsing the conditioning outfit images to a vector performing extreme compression of the image information. To avoid loss of fine textural details, we also consider an additional pathway that can pass more information from the outfits, by directly leveraging the image spatial structure.

We propose to treat the set $\{A_k\}_{k=0}^{K}$ as a single image $\hat{A} \in \mathbb{R}^{K \times C \times H \times W}$, by concatenating spatially the images and making a composite image. Afterwards, we calculate dot-attention [20] with key and value tensors X^k, X^v that are functions of $\hat{A}$, and query tensor X^q which depends on the pose of the generated human. The output of the attention X^a can be directly concatenated as channels to the input noise of the diffusion process. A schematic of this cross-attention mechanism is shown in Figure 2a.

An advantage of this structure is that it is permutation-invariant and can use input outfits $\{A_k\}_{k=0}^{K}$ of any size K. The dot-attention mechanism will copy information from each of the spatial regions of $\hat{A}$ and pass it to the right positions according to the query X^q. In practice, parts of the query will activate with different regions of the $\hat{A}$, e.g. the legs keypoint will copy a trouser image A_k, the torso keypoint will copy from a shirt image A_k, the feet will activate from the shoe packshot image. All of these regions are spatially combined and the trainable attention parameters will adapt to facilitate the right activations.

2.3 Classifier-free Guidance

Recently, classifier-free guidance (*cf*-guidance) [8] has been shown to significantly improve the quality of the generated image, albeit with some loss in diversity. [8]



(a) Outfits (b) DF-DSC (c) DF-CAP (d) DF-FULL

Fig. 3: Sample results from our method DF are shown here. a) shows the input conditional outfits to the DM. b-d) shows resulting images from the different methods. We find in both top row as well as the bottom row that results from DF-FULL are more faithful to the input conditions capturing the fine textural details.

proposed to add dropout during training, inducing the learning of the unconditional distribution and thus allowing to weigh the different components of the Bayesian formulation.

For the case of outfit renderer, this can be achieved by random dropout of the outfits during training. This is similar to random dropout of the conditional embeddings during training, as shown by [8]. Dropping an outfit forces the DM to learn the unconditional distribution of a digital human wearing a random outfit to fill the missing part of the outfit. Thus, during inference, the number of outfits can be varied, allowing to estimate the conditional as well as the unconditional distributions required to perform *cf*-guidance.

3 Experiments

3.1 Dataset & Training

In this work, we use a proprietary dataset consisting of 1.6 million images. In the ground-truth pairs of $(A_k, \boldsymbol{x})$, every image $\boldsymbol{x}$ is associated with at least one and a maximum of 6 outfits. The images are fixed at a maximum resolution of 512×256. The test dataset consists of images of outfits taken by a subsidiary of Zalando with different studio settings. This train-test split allows to check the generalization performance of our method.

We utilized the code repository of [3] for the design and training of DM. The code was modified to fit the architectural changes proposed in Section 2. We trained DF using the following conditional information pathways to render the outfits on the generated digital human:

- DF with deepset embeddings as conditioning (DF-DSC).
- DF with deepset embeddings and designed to perform cross-attention with pose (DF-CAP).
- DF with both the above information pathways and with classifier-free guidance (DF-FULL).

3.2 Qualitative Evaluation & User Study

Evaluating the diversity of images generated in literature has been primarily performed by FID [6] and precision recall [10] metrics. However, in the case of outfit renderer, these metrics do not have a useful meaning as having a low FID score but poor faithfulness to the original input outfits is considered a failure case. Hence, we rely on qualitative evaluation in the form of visual inspection with a user study.

We posed two questions to the users as shown in Table 1. The first question focuses on the overall quality of the digital human generated, while the second question looks at the reconstruction of the outfit on the final image. This survey was filled out by $N = 10$ users who looked at three images at once as shown in Figure 3 b-d). The results shown in Table 1, displayed as percentages indicates that the users preferred the method DF-CAP for a more realistic digital human and DF-FULL for faithful outfit reconstruction. Since the digital human generation is modeled here using the unconditional distribution, DF-CAP achieves a better digital human generation, as reported by users in Table 1. On the other hand, DF-FULL is simply DF-CAP with *cf*-guidance, which provides additional weight to the conditional distribution during inference. Thus the final image is driven towards the conditioned outfits, with a loss in diversity. For the outfit renderer, this loss in diversity is beneficial as the final image is more faithful to the outfits. Hence, users preferred DF-FULL for the second question on outfit reconstruction in Table 1. This effect can also be observed in Figure 3. DF-DSC images in the top and bottom row show some inconsistencies in rendering the correct outfit—color of the outfit in the top row and collar of the outfit in the bottom row. DF-CAP improves upon DF-DSC, while DF-FULL reconstructs more fine details of the conditioned outfits including color and texture.

The first and foremost requirement for enabling a meaningful virtual try-on (and its 2D variant, outfit renderer) experience is the ability to faithfully reconstruct input garments on a body. DF-FULL can more accurately reconstruct the outfits due to its ability to retain their high-frequency details with the help of dual information pathways as well as employing *cf*-guidance.

Additional images of comparisons of the different methods as well as further images generated by our best performing model DF-FULL can also be found in Appendix **??**.

Table 1: Results of user study on qualitative evaluation of the different methods. The percentage of user preference for each question is shown here.

Question	DF-DSC	DF-CAP	DF-FULL
Which digital human looks more realistic?	0.295	0.43	0.275
Which image reconstructs the outfit more accurately?	0.168	0.160	0.672

4 Conclusion

The main difficulty in rendering a given outfit on a digital fashion model is to accurately reconstruct the input outfits on the body of the digital model. This is in particular a challenging task for outfits with complex patterns as those patterns rarely appear in the training set. Here we showed that with a careful design of the image encoder (to avoid information loss) and classifier-free diffusion guidance at inference time we can accurately dress the digital fashion model with the input outfits. This in turn will enable us to relieve our studio from many repetitive and time consuming tasks and let them focus their resources on creating more creative and artistry contents.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition. pp. 248–255 (2009)
3. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
4. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. pp. 6840–6851 (2020)
8. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
9. Jetchev, N., Bergmann, U.: The conditional analogy GAN: Swapping fashion articles on people images. In: International Conference on Computer Vision Workshops. pp. 2287–2292 (2017)
10. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019)
11. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171 (2021)
12. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
14. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
15. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)
16. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265 (2015)
17. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Neural Information Processing Systems (2019)
18. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Neural Information Processing Systems (2020)
19. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), `http://arxiv.org/abs/1706.03762`
21. Yildirim, G., Jetchev, N., Vollgraf, R., Bergmann, U.: Generating high-resolution fashion model images wearing custom outfits. In: International Conference on Computer Vision Workshops. pp. 0–0 (2019)
22. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. Advances in neural information processing systems **30** (2017)