MLVU Final Report
A Comprehensive Image-based Virtual Try-on Network: Bringing Real-World Images into Use

Dongsig Kang 2020-21915
Eojin Kim 2021-28110
Hyunjong Kim 2021-20021

Abstract

The purpose of Image-based virtual try-on is to synthesize target clothing into a proper area of the person. A number of studies have been conducted in this area and performed well on organized images. However, previously proposed Image-based virtual try-on models have several challenges regarding untidy realistic images and model learning only on tops. First, previous models have mostly used the VITON [1] dataset as train and test data. Images in this dataset have a white background and are aligned in the middle with only clothes and human images. However, the realistic images are not aligned in the middle, and the background exists, so the results were very poor when applying the existing models. Second, previous models have implemented models with a focus on top, and some have implemented models to make it impossible to train bottoms. To alleviate these challenges, we improved existing models by using DeepFashion2 dataset [2] and built a comprehensive model that works for tops and bottoms.

1. Introduction

The online clothing market has a greater commercial advantage than the traditional clothing market, but it has a disadvantage of not being able to try it on physically. To provide a shopping environment that is close to reality, virtual try-on techniques have received much attention because they can provide information similar to trying on real products. This technique helps users make decision quickly about whether or not to buy the clothes. Existing studies have produced 3D models and synthesized them directly through graphical manipulation, which requires a lot of time and labor.

Recently, more economical methods have begun to emerge to implement virtual try-on techniques with images themselves, rather than converting to 3D information. Given the image of a person and the image of the clothes user wants to wear, various studies have been conducted on basic pipeline models that modify the image of clothes based on human posture, body shape and characteristics of clothes and synthesize them.

However, most studies are conducted by transforming clothes and synthesizing them into images only for tops, and most of the studies on other areas are based on 3D transformation. Try-on technique for skirts, pants and bottom has also been proposed, but simply passing through a single network during the inference process does not produce good results and requires additional learning to achieve desired results, which actually takes a lot of time to achieve.

Most models in try-on field learn a module that warps clothes to fit the human form. For this learning, they mainly use pose maps with segmented images by person part. However, it is very difficult to apply existing models because the pose estimation is not done well for the lower body for pants or skirts. Most pose estimations are carried out using key points of the upper body, such as arms, chest, shoulders, neck and face, and the image of a person wearing a bottom is difficult to estimate the pose because most of the images show only the lower body. Furthermore, even if the body is fully visible, pose estimation is difficult in the case of skirts because it is difficult to detect the leg which is key point for pose estimation.

Therefore, we seek to create a network that allows people to wear different types of clothing, such as pants and skirts, rather than a virtual try-on network that only works well on tops. In addition, we want to create a network for progressing learning using realistic dataset (DeepFashion2) [2] and implement a network that is desirable to apply to real images.

2. Related Works

2.1. Generative Adversarial Networks (GAN)

GAN has led to tremendous advances in image synthesis [3, 4, 5] and processing [6, 7]. GAN consists of a gener-

ator and a discriminator. Generator generates realistic images to deceive the discriminator, which learns to distinguish synthesized images among images in reality [8]. Although GAN is used in many domains, it has resulted in tremendous performance gains, especially in the field of image synthesis [3, 4, 5].

To utilize additional information, such as text [9], and attributes [10], in generating image process, conditional generative adversarial networks (cGANs) were proposed. Many cGANs conditioned on input images have been proposed to generate high-resolution images [11, 12]. However, these models had the problem of generating blurry images when dealing with very large spatial variations or deformation between input and target images.

2.2. Fashion Analysis and Synthesis

Various tasks on fashion have received considerable attention because there are many applicable fields in the real world. Examples of current tasks include clothing compatibility and matching learning [13], clothing landmark detection [2], and fashion image analysis [14, 15]. Virtual try-on field is also one of the main challenges in fashion.

2.3. Virtual Try-On Approaches

Most studies in the field of virtual try-on are based on 3D graphics models which render the output images via the precise control of geometric transformations or physical constraints [16, 17, 18, 19]. Using these 3D models, they produce good results for Virtual Try-On. However, due to its low efficiency, high computation resource and heavy labor, a new methodology has been found and with the development of deep neural networks it has become an important area that has recently resurfaced.

Without 3D transition of images, maintaining human pose and identity information was very important to synthesize clothes into humans, and most recent studies have approached how to learn this information. For example, CP-VTON [20] and CP-VTON+ [21] used similar two-stage frameworks and made the original Thin-Plate Spline (TPS) transformation [22] learnable based on a convolutional network for geometric matching [23]. Although the Virtual Try-On results changed more naturally, the results were still not good when there was high occlusion or large transformation was needed.

To alleviate these problems, ACGPN [24] was presented. CP-VTON only focuses on the clothes, leading to coarse and blurry bottom clothes and posture details. Therefore, ACGPN added an additional semantic generation module (SGM) to generate a semantic frame of spatial layout. Although the performance has improved, it was still not good enough to apply to real-world datasets.

Prior methods are heavily based on human parsing and pose. However, wrong segmentation for human images would lead to bad results on Virtual Try-on. To solve this problem, PF-AFN(Parser-Free Virtual Try-on) [25] was presented. PF-AFN treats the fake images produced by the parser-based network as input of the parser-free student network, which is supervised by the original real person image in a self-supervised way.

3. Method

We adopt ACGPN because the model predicts and generates exactly where to synthesize, and would be best suited to our method. We made two separate models for each top and bottom generation based on ACGPN. Model differs at which mask to make in semantic generation module (SGM). Each semantic generation module makes preserving mask and warping mask which is the region of target clothes that have to warp. In spatial transform network (STN), target cloth warps to warping mask. Finally, content fusion module (CFM) uses the previous network's outputs and reference image with information from the clothing part removed as an input and generates the final synthesized image. We constructed the models for top and bottom generation each but for consistency, the following explanation will focus on bottom clothes generation. Main difference between each part is where to mask and preserve. By replacing legs to arms and bottom to top will explain for the top generation.

3.1. Semantic Generation Module (SGM)

If the shape of the clothes you are wearing is different from the shape of the target clothing, creating a clothing mask when you are wearing the target clothing can preserve body parts of the person [24]. To generate a new mask applying target cloth, change the label of lower body parts (bottom clothes, legs) identically from reference mask M and make a fused mask M_F. Next, target cloth c, pose map p [2], fused mask M_F are concatenated to generate warping mask M_w. Once again, concatenating M_w, M_F, p we synthesized mask M_S by $G2$.

For training, we used the structure of the conditional generative adversarial network (cGAN) [26] with a generator using the U-Net [27] structure and a discriminator using pix2pixHD [11]. Loss function of cGAN is:

$$L_{GAN} = E_{x,y}[log(D(x,y))] + E_{x,z}[log(1 - D(x, G(x,z))]$$

where x indicates the input and y indicates the ground truth mask by reference mask. z is noise sampled from standard normal distribution.

Overall Loss function is as follows,

$$Loss = \lambda_1 L_{GAN} + \lambda_2 L_{PCE}$$

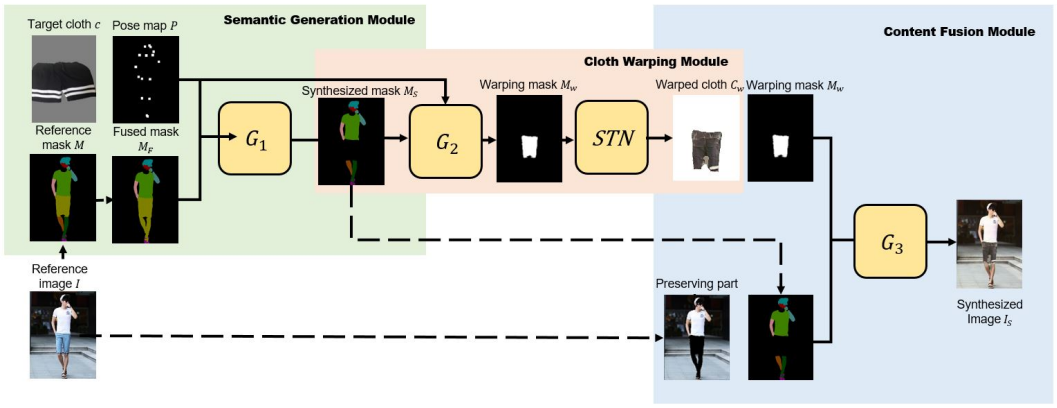

Figure 1: The overall architecture of our method for pants.

where λ_1 and λ_2 are trade-off parameters and L_{PCE} is pixel-wise cross entropy loss for training to achieve more accurate segmentation results.

3.2. Cloth Warping Module (CWM)

By applying thin-plate splines [28] to spatial transformation network [23], transforms target cloth c to the mask of the clothing part M_w. Additionally, using the second-order difference constraint on cloth warping module ensures robust results for the complex text and rich color [24].

3.3. Content Fusion Module (CFM)

The final synthesized image should remain unchanged except for the lower body part where we want to proceed with overall networks from the reference image. For generating synthesized image, we used warped cloth c_w, cloth agnostic reference image I_p, synthesized masking result M_S and warping mask M_w as an input with U-Net structured network and trained the network with L_1 loss.

4. Experiment

4.1. Dataset

Our dataset was constructed by extracting images from DeepFashion2 dataset [2] and organizing them into the form our model needs. Compared to the VITON dataset [1] used in many other virtual try-on networks, DeepFashion2 dataset includes more real-world unprocessed images. Image sizes are not constant, numerous backgrounds appear, and occlusion often occurs. DeepFashion2 dataset contains 391K images for training, 34K images for validation, and 67K images for test. An image can have several items and each item has annotations including category, segmentation, occlusion, and viewpoint. Also, an identical clothing item can appear in several images.

For training, we had to make a dataset which contains pairs of an image of a clothing item and an image of a person wearing the same item. First, we extracted the list of items with viewpoint label 'no wear' and 'frontal', respectively. Then, we combined two lists, leaving only the items that have both no wear image and frontal image. For items that have several no wear images, we just left one with the lowest occlusion value. Finally, by pairing the no wear image with all the frontal images of the same item, the pairs we needed were obtained. No wear image is used as the clothing image, while frontal image is used as the person image. We took these steps for both top items and bottom items, where top items refer to the items with category la-

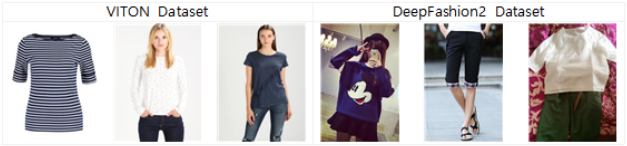

Figure 2: Examples of VITON dataset and DeepFashion2 dataset.

Figure 3: Examples of VITON dataset and DeepFashion2 dataset.

bel 'short sleeve top', 'long sleeve top', 'vest', 'sling' and bottom items refer to the items with category label 'shorts', 'trousers', 'skirt'.

Using the aforementioned method, we extracted our training set and test set each from the original DeepFashion2 training set and validation set. For top items, the training set and test set each contains 27116 pairs and 4354 pairs. For bottom items, the training set and test set each contains 11093 pairs and 1994 pairs.

Then, using the polygon segmentation annotation of the DeepFashion2 dataset [2], we created cloth data without a background for the no-wear dataset.

By preprocessed pose estimation [29] and human parsing [30], we created cloth agnostic human representation for semi-supervised learning. Unlike VITON [1], DeepFashion2 dataset is not well-organized, so we often failed to detect people in pose estimation. For better training we excluded data which detected none. We used 18334 pairs for top, and 5969 pairs for bottom in the training procedure.

4.2. Implementation Details

Architecture. The model contains semantic generation module (SGM), cloth warping module (CWM) and content fusion module (CFM). Generator of $G1$, $G2$, $G3$ uses the structure of U-Net [27] and spatial transformation network [23] consists of five convolution layers with U-Net and a discriminator with pix2pixHD [11]. Because DeepFashion2 dataset has various resolutions, resized to resolution of VITON, 256 × 192.

Training. We trained 20 epochs for each top and bottom network with the pretrained network using VITON. We set hyperparameters as batch size 16, $\lambda_r = \lambda_s = 0.1$, $\lambda_1 = \lambda_2 = 1$. Learning rate is initialized as 0.0002 and used Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$.

Testing. We tested with the validation set of DeepFashion2. For fair comparison, we made test pairs for top and bottom and evaluated them in quantitative and qualitative ways.

4.3. Qualitative Results

Since recent networks are already producing remarkable results with VITON dataset, we wanted to see if we could make progress on the results with trickier images. Also, since DeepFashion2 dataset includes images of both top and bottom clothing items while VITON dataset is restricted to tops [1], we could make an attempt on virtual try-on with bottom clothing items.

Regarding top clothing items, we tested our model on DeepFashion2 dataset and compared the results to the model trained with VITON dataset. As shown in the left side of Figure 4, the model trained with VITON dataset struggled detecting the exact area where the target clothing should be put on. Also, lots of distortions and blurry points occurred. In contrast, our model showed better performance on rendering the target clothing on to the person, with less distortions and misalignments. The right side of Figure 4 shows an example of synthesizing different clothing items to the same person. The model trained with VITON dataset has shown to be extremely unstable when the target clothing is not well-aligned, and when the object size of the target clothing differs from the one initially wearing. Our model showed much more stability, performing better in warping the target clothing into the body shape of the reference person. It has shown the possibility of utilizing a wider range of images in virtual try-ons.

In the case of bottom clothing items, we focused on generating more natural-looking images. As shown in in the left side of Figure 5, our model generated some promising results, successfully synthesizing bottom clothing items to the reference person. A notable fact is that the network managed to straighten the folded jeans into the person's body shape, as it can be seen in the second row. The right side of Figure 5 shows an example of synthesizing different clothing items on the same person. On average, the model worked better with shorts, trousers and short skirts, than long skirts.

	PF-AFN	ACGPN	Our Model (preserving background)
FID (top)	71.678	213.576	27.057
FID (bottom)	101.980	110.892	40.662

Table 1: Quantitative results with FID score.

4.4. Quantitative Results

In virtual try-on tasks, there are no correct labels, so we adopted Frechet Inception Distance (FID) [12] for quantitative evaluation. It calculates distance between distance of feature using inception model pretrained with ImageNet [31]. Lower scores indicate higher quality of the results. We did not adopt Inception Score (IS) [32] because it is known that models which are not trained with ImageNet give incorrect results [25].

For baseline, we used PF-AFN trained with VITON dataset, ACGPN and our model trained with DeepFashion2 dataset. In order to compare models fairly, we made test pairs which include target clothes and reference images to generate an image where the person in the reference image is wearing target clothes. As shown in Table 1, our model showed the lowest score in both top and bottom, among the three.

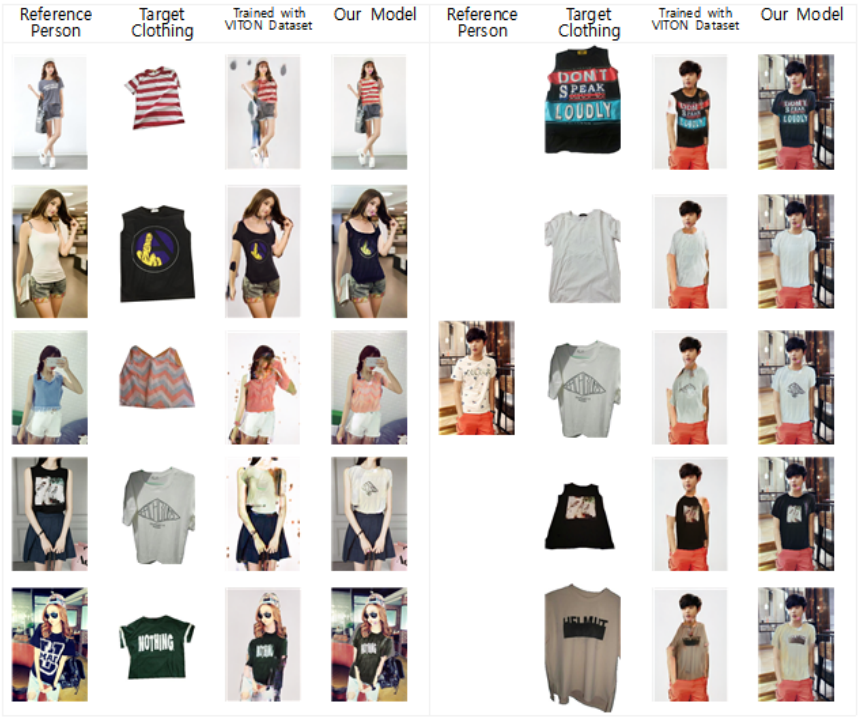

Figure 4: Qualitative comparisons with top clothing items.

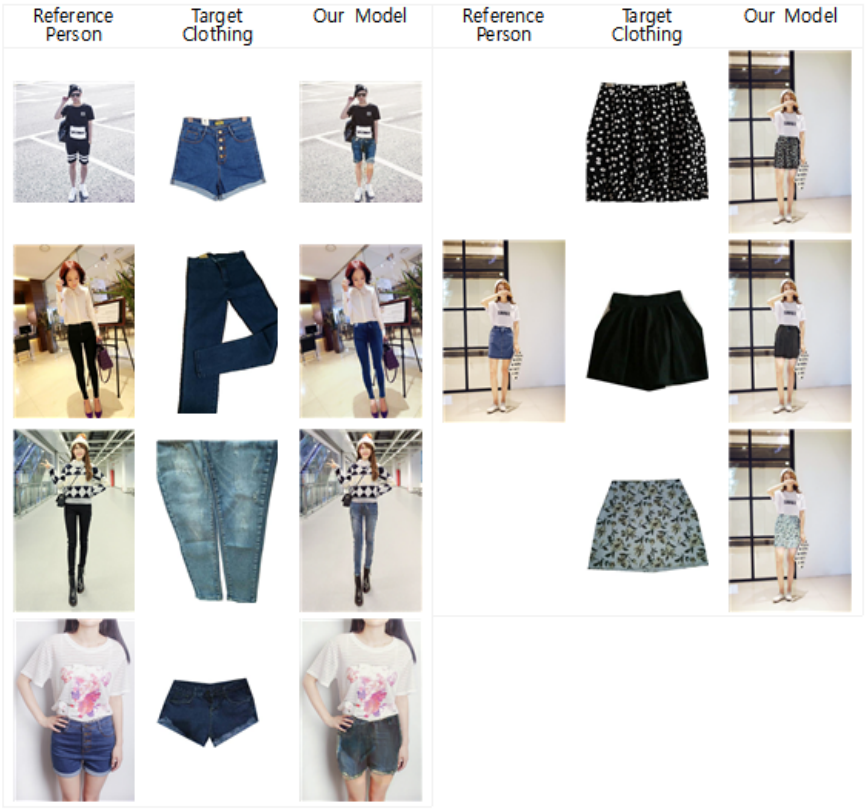

Figure 5: Virtual try-on results with bottom clothing items.

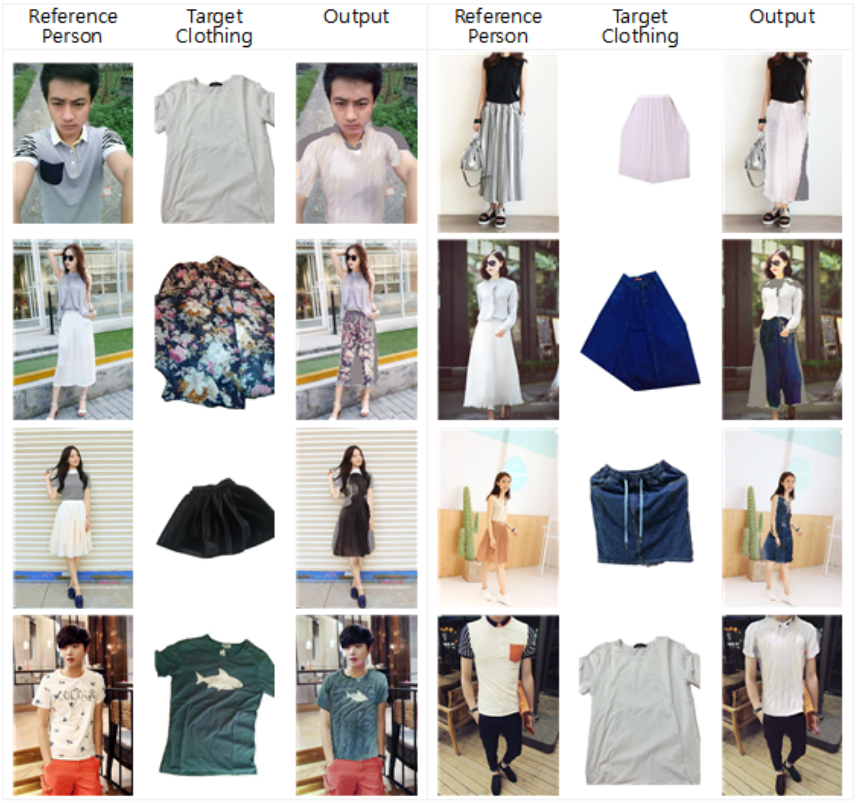

Figure 6: Examples of failure cases.

5. Discussion

5.1. Failure Cases

During experiments, we observed some phenomena that hinder producing acceptable results. These phenomena are shown in Figure 6.

First, when the shape or the size of the target clothing doesn't match that of the original one, the network often filled the missing part with monotonic gray color. Our model succeeded in warping the shape smoothly when the gap was manageable, but the real-world dataset contained images that couldn't be covered with the current network. Examples are shown in the first row.

Second, the network sometimes generated images with weird trousers shapes, even when the target clothing is a skirt and the reference person is wearing a skirt. Examples are shown in the second row. The general expectation is a person wearing a new skirt, but this kind of failure happened. This might be due to the fact that the bottom training set contained all shorts, trousers and skirts, and they were trained all together. This phenomenon may be improved if different models are applied according to the specific type of bottom clothing, but this will require an additional classification network.

Third, the network put the target clothing on wrong areas in some cases. The third row shows examples where short skirts are stretched to the whole body. Indeed, the quality of the synthesized images was disappointing. This kind of failure occurred when the parsing result was not accurate.

Finally, some kind of twisted texture, which doesn't exist in the target clothing image, often appeared on the surface of the synthesized cloth. This particularly happened with top clothing items. Examples are shown in the fourth row. We are not sure about the exact cause, but predicting it to be due to the unstable nature of generative adversarial networks [8].

5.2. Failed Attempts

We attempted to synthesize try-on images in various settings. First, we used PF-AFN which has the most superior try-on result [25]. However, most of the results processed with DeepFashion2 dataset were unrealistic. We assumed that it is because the model has been trained with no background and images are aligned in the center. Also, DeepFashion2 data with bad quality are hard to detect pose will produce bad results. Furthermore, we assumed appearance flow in PF-AFN, which is pixel sensitive in spatial way, may struggle with various poses.

Figure 7: Synthesized results of PF-AFN with VITON cloth data (Reference image from VITON, DeepFashion2, DeepFashion2 background removed).

Figure 8: Synthesized results with PF-AFN with different target clothes from DeepFashion2.

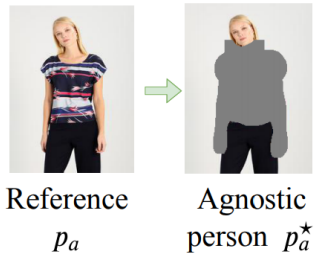

Figure 9: Agnostic person representation from WUTON [33].

5.3. Future Work

As aforementioned, we often had failure of pose estimation and lost considerable data for training. Regarding Issenhuth et al. [33], cloth agnostic human representation without pose estimation had better results. In real-world data, following this agnostic representation in Figure 9 from WUTON will make it possible to train without excluding any data, leading to better results.

Also, we had some failure in human parsing. By applying distillation, we can make direct inference from reference images without making pose estimation and human parsing [25, 33]. It can produce better results regardless of the quality of the parsing result of the reference image.

6. Conclusion

In this work, we proposed an improved ACGPN for realistic images and bottom clothes. To make ACGPN perform well in the realistic image, the training was implemented using DeepFashion2 dataset preprocessed for learning above of the pretrained model using VITON data. Also, because previous ACGPN implemented Virtual Try-On only for the top, we changed the model to make it applicable to the bottom. Qualitative and quantitative experiments show that our new model outperforms previous models for realistic data and bottoms.

References

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7543–7552.

[2] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

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

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[6] Y. Jo and J. Park, "Sc-fegan: face editing generative adversarial network with user's sketch and color," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1745–1753.

[7] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[9] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[10] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4030–4038.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.

[13] T. Iwata, S. Wanatabe, and H. Sawada, "Fashion coordinates recommender system using photographs from fashion magazines," in *IJCAI*, vol. 22, no. 3. Citeseer, 2011, p. 2262.

[14] X. Han, Z. Wu, W. Huang, M. R. Scott, and L. S. Davis, "Compatible and diverse fashion image inpainting," *arXiv preprint arXiv:1902.01096*, 2019.

[15] W.-L. Hsiao, I. Katsman, C.-Y. Wu, D. Parikh, and K. Grauman, "Fashion++: Minimal edits for outfit improvement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5047–5056.

[16] R. Brouet, A. Sheffer, L. Boissieux, and M.-P. Cani, "Design preserving garment transfer," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. Article–No, 2012.

[17] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen, "Synthesizing training images for boosting human 3d pose estimation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 479–488.

[18] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "Drape: Dressing any person," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1–10, 2012.

[19] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual fitting by single-shot body shape estimation," in *Int. Conf. on 3D Body Scanning Technologies*. Citeseer, 2014, pp. 406–413.

[20] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 589–604.

[21] M. Minar, T. Tuan, H. Ahn, P. Rosin, and Y. Lai, "Cp-vton+: Clothing shape and texture preserving image-based virtual try-on," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, vol. 2, no. 3, 2020, p. 11.

[22] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[23] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6148–6157.

[24] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7850–7859.

[25] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," *arXiv preprint arXiv:2103.04559*, 2021.

[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[29] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[30] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[32] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *arXiv preprint arXiv:1606.03498*, 2016.

[33] T. Issenhuth, J. Mary, and C. Calauzènes, "Do not mask what you do not need to mask: a parser-free virtual try-on," *arXiv preprint arXiv:2007.02721*, 2020.