Image-Based Virtual Try-On Network

Pham Anh Huy[1] and Do Duc Vinh[1]

[1]Students specializing in Artificial Intelligence (AI) at FPT University

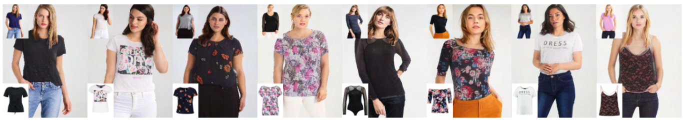

Fig. 1. Illustrative outcomes produced by the Image-Based Virtual Try-On Network (IVTON). Each instance displays the initial input image at the upper left corner, accompanied by the intended clothing in the lower left corner.

***Abstract*—The groundbreaking Image-Based Virtual Try-On Network (IVTON) represents a significant leap forward in the realm of deep learning for fashion technology. By harnessing the power of conditional Generative Adversarial Networks (GANs), IVTON redefines the way individuals engage with clothing in the digital realm, seamlessly superimposing garments onto user-provided images to create a remarkably authentic and immersive virtual fitting experience. With its intricate encoder-decoder architecture enhanced by attention mechanisms, IVTON captures the nuances of how clothing drapes and conforms to the wearer's body, resulting in a strikingly accurate portrayal of virtual attire. The rigorous process of assembling a diverse dataset, meticulously annotated with detailed garment attributes and poses, empowers IVTON to generalize across an array of clothing types and poses, while strategic training techniques and finely-tuned loss functions optimize its performance. Empirical validations underscore IVTON's remarkable proficiency in generating photorealistic virtual try-ons, poised to reshape the landscape of online fashion shopping by fostering a heightened sense of customer confidence and satisfaction.**

I. Introduction

In recent years, the intersection of deep learning and fashion technology has yielded innovative solutions that are transforming the way we engage with clothing in the digital age. One of the pioneering advancements in this domain is the Image-Based Virtual Try-On Network (IVTON), a revolutionary system that leverages the capabilities of conditional Generative Adversarial Networks (GANs) to create an unprecedented virtual fitting experience. This paper presents a comprehensive exploration of IVTON, an intricate architecture that seamlessly integrates garments onto user-provided images, resulting in a remarkably authentic and immersive representation of virtual attire.

The fusion of computer vision and generative modeling has opened up new avenues for enhancing the fashion industry's online presence. IVTON, through its sophisticated encoder-decoder framework enhanced by attention mechanisms, bridges the gap between real-world clothing and the digital realm by accurately capturing the way garments drape and conform to the wearer's body. This level of fidelity in rendering virtual clothing on user images is expected to revolutionize online fashion shopping, fostering a heightened sense of customer confidence and satisfaction by allowing users to virtually try on clothing items before making a purchase.

To achieve its remarkable proficiency, IVTON relies on a meticulously curated and annotated dataset - VITON. This dataset, composed of diverse clothing types and poses, serves as the foundation for training the network to generalize effectively across a wide spectrum of fashion items. The incorporation of detailed garment attributes and poses ensures that IVTON comprehends the subtleties of various clothing styles and how they interact with different body movements. Furthermore, strategic training techniques and finely-tuned loss functions play a pivotal role in optimizing IVTON's performance, enhancing its ability to generate photorealistic virtual try-ons that resonate with the user's personal style and physique.

Empirical validations conducted on IVTON underscore its efficacy in producing photorealistic virtual try-ons. By evaluating the network's output against real-world clothing images, these validations establish IVTON's capacity to generate accurate and visually compelling representations. This breakthrough technology stands poised to reshape the landscape of online fashion shopping, transcending the limitations of static images and inaccurate size estimations that have traditionally hindered customers' digital interactions with clothing items.

In the subsequent sections of this paper, we delve into the architectural intricacies of IVTON, discussing its encoder-decoder design, attention mechanisms, and the role of conditional GANs in facilitating the seamless integration of virtual attire. Additionally, we detail the construction and annotation of the dataset, shedding light on the strategies employed to ensure diversity and comprehensiveness. Finally, we present the results of empirical validations, shedding light on IVTON's potential to revolutionize the fashion industry's digital landscape.

In summary, IVTON represents a significant milestone in the convergence of deep learning and fashion technology. By blending cutting-edge generative modeling techniques with a carefully curated dataset, IVTON redefines the virtual fitting experience, paving the way for enhanced customer engagement and satisfaction in the realm of online fashion shopping.

II. Related Work

In recent times, image-based virtual try-on techniques have emerged as an attractive alternative to conventional try-on solutions that depend on 3D modeling and specialized computer graphics pipelines. For instance, pioneering efforts by GANs models approached the virtual try-on task as an image analogy challenge and yielded promising outcomes. However, the generated images displayed limited photorealism due to the absence of explicit (clothing) deformation modeling. To address this limitation, we implemented a two-stage strategy named VITON, employing a coarse-to-fine image generation method. They incorporated a Thin-Plate Spline (TPS) transformation [2] to align the desired clothing image with the target subject's pose. Building upon this, Wang et al. [1] enhanced the approach with CP-VTON, introducing a Geometric Matching Module (GMM) that allowed learning TPS clothing transformations [2] in an end-to-end manner, resulting in impressive try-on results similar to [5].

Further progress refined the geometric matching stage using diverse mechanisms. CP-VTON+ [6], for instance, improved the human mask input to the GMM, while VTNFP [7] designed a detailed person representation for the GMM input. In line with this body of research, our approach with IVTON presents a unique matching module based on simplified inputs that can be reliably estimated even in the presence of significant appearance variability. We accomplish this by conditioning the module on body parts only, leveraging the capabilities of modern human-parsing models.

Numerous solutions have also been proposed to enhance the quality of generated try-on outcomes during the image synthesis phase. For instance, MG-VTON [4], ACGPN [9], and VITON-HD [8] suggested using secondary neural networks to generate clothing segmentations that match the target garment. These additional sources of information are then integrated into the generator. S-WUTON [10] and PF-AFN [16] employed a knowledge distillation framework between teacher-student models to mitigate the need for error-prone intermediate processing steps often seen in existing try-on methods. FE-GAN [5] and VITON-HD [8] followed recent advancements in image synthesis, introducing generators with conditional normalization layers to enhance the quality and realism of the synthesized try-on results.

Similar to these techniques, our approach with IVTON also employs a sophisticated image generator featuring conditional normalization layers during the synthesis stage. However, we harness contextual information to guide the generation process. Moreover, we employ three powerful discriminators in an adversarial training setup to maximize the utility of available contextual information, ultimately enhancing the realism of the generated outcomes.

III. Proposed Methods

We present a two-step procedure involving: warping the clothing to match the desired contour and shape, and then aligning the individual's body shape with the contoured clothing. Formally, we give an image of human body as an input, $I \in \mathbb{R}^{w \times h \times 3}$, and the target clothing, $C \in \mathbb{R}^{w \times h \times 3}$. The objective of the model is to generate an image that achieves synchronization between the body shape of individual I and the clothing item C they are wearing, $I_c \in \mathbb{R}^{w \times h \times 3}$.

A. *The Body-Part Geometric Matcher (BPGM)*

The initial step in the process involves estimating the parameter theta for Thin-Plate Spline (TPS) [2] transformation is used to warp the input clothing according to the human pose image I. The BPGM takes a target clothing C and the body segmentation $S \in \{0,1\}^{w \times h \times d}$ - are generated using DensePose model from [3] and contain d = 25 channels (classes), each corresponding to a different body part. - and then produces a warped image C_w as an output. The BPGM still relies on the design of Geometric Matching Module (GMM)[34] and consists of two encoding modules: E_1 and E_2. The E_1 takes input C and generate the corresponding feature vector $e_1 \in \mathbb{R}^{w_f \times h_f \times d_f}$. Similarly, E2 takes the body segmentations S and outputs corresponding feature vectors $e_2 \in \mathbb{R}^{w_f \times h_f \times d_f}$. Note that w_f and h_f represent the dimensions of the encoding, and h_f signifies the number of output channels. Next, the feature representations are normalized channel-wise to unit L2 norm, then flattened and organized into a matrix $\Psi_E \in \mathbb{R}^{d_f \times w_f h_f}$, which serves as the basis for computing the correlation matrix Corr:

$$Corr = \Psi_{E_1}^T \Psi_{E_2} \in \mathbb{R}^{(w_f h_f) \times (w_f h_f)}$$

The $Corr$ matrix is then fed into the Regressor, R, which predicts the parameter θ (with $2n^2$ dimensions) that relates to the x and y shifts within an n×n grid, determining how the clothing item C is transformed. There are three lost functions are used to learn the parameters of the BPGM:

- **A target shape loss** (L_{shp}): that motivates the deformation process to adjust the desired clothing into a form that aligns with the pose of the person in image I.

$$L_{shp} = \|M_w - M_c\|_1 = \|T_\theta(M_t) - M_c\|_1$$

  where M_t and M_w are binary masks corresponding to the original (C) and warped target clothing (C_w). M_c is a binary mask corresponding to the clothing area in the input image (generated by the DensePose model [3]) and T_θ denotes the TPS transformation parameterized by θ.
- **An appearance loss** (L_{app}): that compels the visual representation of the deformed clothing C_w within the body region M_b to closely resemble the original input image I.

$$L_{app} = \|C_w \odot M_b - I_b\|_1$$

  where $\odot$ denotes Hadamard product and M_b is a binary mask of the body area (a channel in S), and $I_b = I \odot M_b$.
- **A perceptual loss** (L_{vgg}): that guarantees that both the desired clothing and its transformed version maintain identical semantic content within the body region.

$$L_{vgg} = \sum_i^n \lambda_i \|\phi_i(C_w \odot M_b) - \phi_i(I \odot M_b)\|_1$$

  where $\phi_i(\cdot)$ is a feature map generated before each (of the n = 5) max-pooling layer of a VGG19 [11] model (pretrained on ImageNet), and λ_i is the corresponding weight.

Among the aforementioned losses, L_{shp} intends to align the overall clothing area, while L_{app} and L_{vgg} are tailored to precisely match the graphics on the clothing itself. This is done without imposing the requirement on the BPGM matcher to align sleeves, which frequently result in unrealistic modifications that the generator later employs. To include, the loss function for BPGM is:

$$L_{BPGM} = \lambda_{shp}L_{shp} + \lambda_{app}L_{app} + \lambda_{vgg}L_{vgg}$$

where λs are the corresponding balancing weights. The parameters of the BPGM are acquired through learning using a dataset consisting of input images I and corresponding images of target clothing C.

B. *The Context-Aware Generator (CAG)*

The second step involves the Context-Aware Generator (CAG), which is responsible for producing the final virtual try-on image I_c. Before delving into the discussion, I would like to clarify the following pieces of information. After the fisrt step, we got a body segmentations S and C_w as the deformed clothing. we jointly refer to all inputs of the generator as Image Context (IC) and we define it as following: $IC = S \oplus I_m \oplus C \oplus C_w$ (with $I_m = I \odot M_c$, as a mask cloting area of input image I).

The context-aware generator cosists of a sequence of ResNet blocks [17] and (2×) upsampling layers augmented with what refered to as Context-Aware Normalization (CAN) operations. The proposed CAN layers are designed to leverage the information from the image context IC and supply the generator with those essential details. This process is carried out at different resolutions to make sure that the generator is normalized at different levels of guaranularity and efficiently spreading details about the intended semantic layout and desired appearance of the generated output across the generator. Each ResNet block has two inputs: the image context IC and the activation map from the previous model layer. The ResNet blocks consist of a sequence of batch-normalization and convolutional layers repeated twice with CAN operations preceding the convolutional layers. The output of batch-normalization is denoted as X_{BN}, then CAN operation can formally be defined as: $X_{CAN} = X_{BN} \odot \gamma + \beta$, where $\odot$ denotes the Hadamard product. γ and β respectively represents the scale and bias parameters with the same dimensionality as X_{BN}.

There are two types of loss functions are designed to learn the parameters of generator:

- **A perceptual loss** (L_{per}): that stimulates the generator to produce a virtual try-on result as close as possible to the input image I in terms of semantics.

  $$L_{per} = \sum_{i}^{n} \tau_i \|\phi_i(I'_c) - \phi_i(I)\|_1$$

  where $\phi_i(.)$ are feature maps produced by a pretrained VGG19 model [11] before each of the n=5 max-pooling layer, τ_i is the i-th balancing weight, and I'_C is the try-on result generated with the matching target clothing.
- Three **adversarial losses** defined throught three discriminators, each focused on authentically generating distinct components of the ultimate try-on image:
  - **The segmentation discriminator**(D_{seg}): ensure realistic body-part generation by predicting (per-pixel) segmentation maps S and their origin (real or fake). With the input image (I or I_C), D_{seg} returns a tensor $w \times h \times (d+1)$ where the first d channels contain segmented body parts and the (d+1)-st channel encodes whether a pixel is from a real or generated data distribution. D_{seg} is trained by minimizing a (d+1)-class cross-entropy loss:

    $$\begin{aligned} L_{D_{seg}} &= -\mathbb{E}_{(I,S)}[\sum_{k=1}^{d} \alpha_k (S_k \odot logD_{seg}(I)_k] \\ &= -\mathbb{E}_{I_C}[logD_{seg}(I_C)_{d+1}] \end{aligned}$$

    where the first term applies to the input images I and penalizes the first d output channels, while the second term penalizes the last remaining channel generated from the synthesized image I_C. α_k is a balancing weight calculated as the inverse frequency of the body-part in the given channels of the segmentation map S: $\alpha_k = h_w / \lfloor S_k \rfloor$, where $\lfloor . \rfloor$ is a cardinality operator. The corresponding adversarial loss (L_{seg}) for the generator is finally defined as:

    $$L_{seg} = -\mathbb{E}_{(I_C,S)}[\sum_{k=1}^{d} \alpha_k (S_k \odot logD_{seg}(I_C)_k)].$$
  - **The matching discriminator** (D_{mth}): motivate the generator to produce output images featuring the intended target clothing by predicting whether the target garment C corresponds to the clothing worn in either I or I_C.

    $$L_{mth} = -\mathbb{E}_{(I_C,C)}[logD_{mth}(I_C, C)].$$
  - **The patch discriminator** (D_{ptc}): contributes towards realistic body-part generation by focusing on the appearances of local patches, $P = \{p_0, \ldots, p_m\}, p_i \in \mathbb{R}^{w_p \times h_p \times 3}$, centered at m = 5 characteristic body-parts, the neck, and both upper arms and forearms. The discriminator is trained to distinguish between real and generated body areas based on the following objective:

    $$\begin{aligned} L_{D_{ptc}} &= -\mathbb{E}_{P_{real}}[logD_{ptc}(P_{real})] \\ &= -\mathbb{E}_{P_{fake}}[log(1 - D_{ptc}(P_{fake}))] \end{aligned}$$

    where the P_{real} and P_{fake} e correspond to patches extracted from the real and generated images I and I_C, respectively. The generator loss then takes the following form:

    $$L_{ptc} = -\mathbb{E}_{P_{fake}}[logD_{ptc}(P_{fake})].$$

The overall loss function of the generator is the sum of all individual losses:

$$L_G = \lambda_{per}L_{per} + \lambda_{seg}L_{seg} + \lambda_{mth}L_{mth}L_{mth} + \lambda_{ptc}L_{ptc},$$

where the λs denote hyperparameters that determine the relative importance of each loss term.

IV. Experiments and results

A. Datasets

For our experiments, we utilized two prominent datasets, VITON and MPV, to comprehensively evaluate the performance of the IVTON.

VITON Dataset [12]: The VITON dataset is a widely recognized benchmark for evaluating virtual try-on solutions. It consists of 14,221 training and 2,032 testing pairs of images, each comprising subjects and target clothing items. The images are of resolution 256 × 192 pixels. To ensure fair comparisons and eliminate duplicate images, we filtered out duplicates from both the training and test sets. After this filtering process, the training set comprised 8,586 image pairs, and the test set comprised 416 unique image pairs not seen during training.

MPV Dataset [13]: The MPV dataset serves as another crucial resource for our experiments. It contains 35,687 person images (256 × 192 pixels) adorned with 13,524 unique garments. Unlike VITON, MPV exhibits a greater degree of appearance variability, including variations in zoom level and viewpoint. To maintain consistency, we prefiltered the MPV images to retain primarily (close to) frontal views. The final train and test subsets of MPV consist of 17,400 paired person-clothing images and 3,662 unpaired images.

B. Implementation Details

IVTON is realized through Python using the PyTorch framework. Throughout the processing pipeline, numerous modules are constructed based on ResNet-like blocks, each comprising two convolutional layers followed by a ReLU activation and a trainable shortcut connection. We detail the architectural specifics for the core components of IVTON below.

Body-Part Geometric Matcher (BPGM): The BPGM consists of two encoders, E_1 and E_2, each comprising five stacked convolutional layers. These layers are succeeded by a downsampling operation, a ReLU activation function, and batch normalization. The feature regressor R employs four convolutional layers, each followed by ReLU activation and batch normalization. To acquire the thin-plate spline transformation parameters θ, an 18-dimensional linear output layer is utilized.

Context-Aware Generator (CAG): The CAG employs ResNet blocks, augmented with context-aware normalization before each convolutional layer. Six such blocks are utilized, each paired with a (2×) upsampling layer. Contextual inputs are resized to match the resolution of each block's input. An exponential moving average (EMA) is implemented over generator weights, utilizing a decay value of 0.9999, akin to the approach in [18].

Data	Model	FID↓	LPIPS↓ ($\mu \pm \sigma$)
VITON	CP-VTON	47.36	0.303 ± 0.043
	CP-VTON+	41.37	0.278 ± 0.047
	ACGPN	37.94	0.233 ± 0.047
	PF-AFN	27.23	0.237 ± 0.049
	IVTON	20.48	0.182 ± 0.044
MPV	S-WUTON	8.188	0.161 ± 0.070
	PF-AFN	6.429	n/a
	IVTON	5.292	0.102 ± 0.53

TABLE I
Quantitative comparison of IVTON and other models in terms of FID and LPIPS scores - lower is better, as also indicated by the corresponding arrows.

Discriminators: The matching discriminator D_{mth} encompasses two encoders, one for clothing (C) and the other for the inverse clothing (I_C), comprising six ResNet blocks each. The encoders' outputs are concatenated and fed to a linear layer producing the final output. The patch discriminator D_{ptc} comprises four ResNet blocks, organized in an encoder architecture, followed by a fully-connected layer. The segmentation discriminator D_{seg} adopts a UNet [19] encoder-decoder architecture, featuring a total of 12 ResNet blocks.

Training Details: For training, the ADAM optimizer [20] is employed with distinct learning rates: lr_{BPGM} = 0.0001 for the BPGM, lr_G = 0.0001 for the generator, and lr_D = 0.0004 for the discriminators. All weights are set to 1, except for λ_{vgg} = 0.1 and λ_{per} = 10. The geometric matcher is trained for 30 epochs, while the generator undergoes 100 epochs in all configurations.

C. Quantitative Results

We will analyze FID [14] and LPIPS [15] over processed VITON and MPV datasets to demonstrate the performance of IVTON. Pretrained models, such as CP-VTON [1], CP-VTON+ [6], ACGPN [9], PF-AFN [16] and S-WUTON [10], will also be used to compare results with IVTON.

In Table 1, we compare IVTON with other popular models. It should be noted that pretrained PF-AFN model on MPV are not publicly available, so we borrowed results from [16].

Clearly, IVTON stands out by achieving remarkable improvements over all other models across both datasets. On the VITON dataset, it substantially enhances performance by decreasing the FID score by 24.8% compared to the second-best model and the LPIPS measure by 23.2%. Similar patterns of improvement are also evident on the MPV dataset, where IVTON consistently delivers significant reductions in both FID and LPIPS scores in comparison to its closest competitors. We attribute these impressive outcomes to the streamlined geometric matching process implemented in IVTON, coupled with the incorporation of diverse contextual cues during the final stages of image synthesis.

D. Qualitative Results

Next, we dive into a comprehensive performance evaluation of IVTON. It is important to note that due to the unavailability of a pretrained PF-AFN model for the MPV dataset, we solely

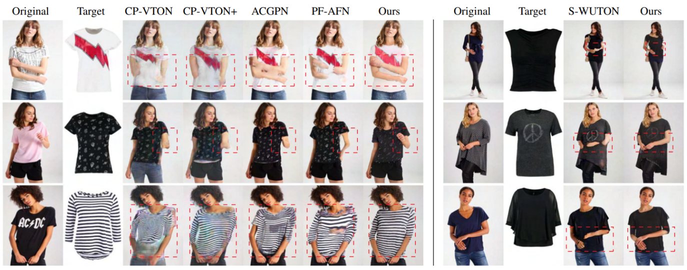

Fig. 2. Comparison of IVTON (ours) and recent state-of-the-art models on VITON (left) and MPV (right) datasets. IVTON excels in arm, hand, and on-shirt graphics synthesis. Best viewed electronically, zoomed-in for details.

engage in a comparison between IVTON and S-WUTON on this dataset.

Visual Comparisons: IVTON in generating highly convincing virtual try-on results, particularly excelling in the synthesis of intricate hand and on-shirt graphics. Our approach consistently produces results that align well with the realism and authenticity sought after in virtual fitting experiences. Even when challenged with subjects imaged in challenging poses and intricate arm/hand configurations, IVTON consistently delivers visually appealing and coherent results.

Comparison with Competing Models: In Figure 2, we present visual examples that illustrate the outcomes of IVTON in comparison to other competing models.In terms of virtual try-on performance on the VITON dataset, the results showcase IVTON as a leading contender. PF-AFN emerges as a strong competitor, generating impressive results. However, as demonstrated by the presented examples, PF-AFN sometimes struggles to preserve essential attributes like arms, initial body shape, and subject pose, areas where IVTON excels. In contrast, our model demonstrates enhanced proficiency in maintaining the integrity of these critical features.

Among the evaluated alternatives, including CPVTON, CP-VTON+, and ACGPN, these models exhibit less convincing results and often falter in retaining specific non-transferable image parts and textures from the target garment. On the MPV dataset, S-WUTON similarly faces challenges in preserving arms and body shape, while C-VTON triumphs in both areas. The exemplary performance of IVTON can be attributed to the meticulous body-part segmentation process employed and the judiciously designed discriminators that ensure the authenticity and realism of the generated images.

In conclusion, the visual evidence provided solidifies the effectiveness of C-VTON in generating impressive and compelling virtual try-on results, outperforming its competitors in critical areas such as pose preservation and intricate graphic synthesis.

E. Limitations

Several limitations warrant consideration in the assessment of IVTON's performance. Among these, the masking procedure used to generate image context, as well as the presence of loose-fitting clothing in the input images, contribute to certain instances of less convincing virtual try-on results produced by IVTON. Additionally, the model's inability to discern between the front and backside of the target garment exacerbates the challenges in achieving optimal outcomes. These limitations collectively give rise to issues such as unrealistic and indistinct garment edges, inaccurately synthesized clothing styles, and inadequately rendered neck areas. It is worth noting that analogous limitations are also observed in the performance of competing models.

V. Conclusion

In this paper, we introduced a new way for people to try on clothes virtually using images. Our method, called IVTON, showed impressive results in various tests on different datasets. It did better than the current best methods.

IVTON uses special parts to understand the shape and style of clothes and then generates realistic images of people wearing those clothes. It's good at keeping the right poses and designs, making the virtual clothes look real.

While there are still some things to improve, IVTON has shown that it can make trying on clothes online much better. This could lead to happier online shoppers and change how people shop for clothes on the internet.

As technology continues to improve, IVTON's achievements could bring exciting changes to how we shop for clothes online, making it more fun and reliable.

References

[1] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward Characteristic-Preserving Image-based Virtual Try-on Network. In European Conference on Computer Vision (ECCV), pages 589–604, 2018.

[2] J. Jean Duchon. Splines Minimizing Rotation-Invariant SemiNorms in Sobolev Spaces. In Constructive Theory of Functions of Several Variables, pages 85–100. Springer, 1977.

[3] Rıza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In Computer Vision and Pattern Recognition (CVPR), pages 7297–7306, 2018.

[4] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards Multi-Pose Guided Virtual Try-on Network. In International Conference on Computer Vision (ICCV), pages 9026–9035, 2019.

[5] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional Neural Network Architecture for Geometric Matching. In Computer Vision and Pattern Recognition (CVPR), pages 6148–6157, 2017.

[6] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. CP-VTON+: Clothing Shape and Texture Preserving Image-based Virtual Try-on. In Computer Vision and Pattern Recognition Workshops (CVPR-W), page 11, 2020.

[7] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Preservation. In International Conference on Computer Vision (ICCV), pages 10510–10519, 2019.

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14131–14140, June 2021.

[9] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards Photo-Realistic Virtual Try-on by Adaptively Generating-Preserving Image Content. In Computer Vision and Pattern Recognition (CVPR), pages 7850–7859, 2020.

[10] Thibaut Issenhuth, Jer´ emie Mary, and Cl ´ ement Calauz ´ enes. 'Do Not Mask What You Do Not Need to Mask: A ParserFree Virtual Try-On. In European Conference on Computer Vision (ECCV), pages 619–635, 2020.

[11] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations (ICLR), 2015.

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An Image-based Virtual Try-on Network. In Computer Vision and Pattern Recognition (CVPR), pages 7543–7552, 2018.

[13] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards Multi-Pose Guided Virtual Try-on Network. In International Conference on Computer Vision (ICCV), pages 9026–9035, 2019.

[14] 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 Advances in Neural Information Processing Systems (NIPS), pages 6626–6637, 2017.

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Computer Vision and Pattern Recognition (CVPR), pages 586–595, 2018.

[16] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-Free Virtual Try-on via Distilling Appearance Flows. In Computer Vision and Pattern Recognition (CVPR), pages 8485–8493, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In Computer Vision and Pattern Recognition (CVPR), pages 770– 778, 2016.

[18] Vadim Sushko, Edgar Schonfeld, Dan Zhang, Juergen Gall, ¨ Bernt Schiele, and Anna Khoreva. You Only Need Adversarial Supervision for Semantic Image Synthesis. In International Conference on Learning Representations (ICLR), 2020.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. UNet: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and ComputerAssisted Intervention (MICCAI), pages 234–241, 2015.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for Stochastic Optimization. In International Conference on Learning Representations (ICLR), 2015.