MULTI-POSE VIRTUAL TRY-ON VIA SELF-ADAPTIVE FEATURE FILTERING

Chenghu Du[1], *Feng Yu*[1,2,†], *Minghua Jiang*[1,2], *Xiong Wei*[1], *Tao Peng*[1,2], *Xinrong Hu*[1,2]

[1]School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China
[2]Engineering Research Center of Hubei Province for Clothing Information, Wuhan, China
duceh_lzy@163.com, {yufeng, minghuajiang, wx_wh, pt, hxr}@wtu.edu.cn

ABSTRACT

With the growing trend of virtual try-on, multi-pose tasks attract researchers due to their higher commercial value. Prior methods lack an effective geometric deformation to maintain the original image details resulting in many details loss in the head and garment. To address this problem, we propose a new multi-pose virtual try-on network, which can fit a garment to the corresponding area of a person in arbitrary poses. First, the target pose's body-semantic distribution is predicted by the target pose point. Second, the in-shop garment and human body are warped based on a human pose to solve the unnatural alignment and the lack of body details by the Deformation Module (DM). Finally, the human body in the given pose and garment is fine generated by the Filtering Synthesis Network (FSN). Compared to state-of-the-art methods with objective experiments on the MPV dataset, the proposed method achieves the best performance in metrics and the rich details in visual results.

Index Terms— multi-pose virtual try-on, semantic segmentation, pose transfer, feature filtering, appearance flow

1. INTRODUCTION

With the rapid growth of the intelligent garment industry, buying garments online is all the rage. Virtual try-on algorithms that enable online dressing gradually gain attention from consumers and researchers. Existing virtual try-on methods[1, 2, 3, 4] such as CP-VTON[2] and ACGPN[3] can only perform single pose virtual try-on tasks, i.e., they can only transfer the in-shop garment to the appropriate area of the human body in a fixed pose; hence, the methods are no longer effective in the face of multi-pose virtual try-on demands.

In recent years, some researchers have expanded on multi-pose virtual try-on. Its steps are using Generative Adversarial Networks (GANs) [5] to transfer the original image to the target pose [6, 7, 8]. Afterward, the aligned in-shop garment is fused to produce the final result. Retaining the body's appearance details is a great challenge because the body undergoes a certain degree of deformation of the individual parts when the pose is changed. Dong et al.[9] propose a 4-stage coarse-to-fine network called MG-VTON. It generates a semantic map from the target pose and warps the in-shop garment to obtain a coarse result from the semantic map. Then, it refines the garment region of the coarse result to generate the final refined result. The face obtained by MG-VTON is blurred because the MG-VTON only refines the garment region and does not further process face details. To address this problem, Wang et al.[10] propose a 4-stage coarse-to-fine network called TB-VTON. Unlike MG-VTON, TB-VTON uses a garment mask to optimize the garment region at the coarse result generation stage. Moreover, in the final stage, TB-VTON introduces a facial refinement network to improve the clarity of the resulting face.

Both of the above approaches attempt to generate realistic results. The critical challenge is to preserve the details of the deformed in-shop garments (e.g., logos, patterns) and bodies (e.g., face, hands). We propose a multi-stage framework that uses self-adaptive feature filtering to generate realistic try-on results. Firstly, unlike other virtual try-on approaches, we use an Appearance Flow Garment Alignment Network from our previous basic work [4] to warp the in-shop garment, which deformed garment fits the body pose reasonably well. Secondly, we design a Filtering Synthesis Network (FSN) that adaptively captures effective information from input conditions to obtain a realistic result.

Our work has three main contributions as summarized as follows:

- We propose a multi-pose virtual try-on framework called MV-TON to get superior try-on results.
- We use a robust deformation module called DM to generate naturally warped garment and preliminary pose transfer results, which effectively addresses loss of detail and over-warping.
- We propose a pose transfer network called FSN, to capture the valuable condition information in the inputs and enhance the learning of spatial transformation to generate the try-on results with realistic details.

2. METHODOLOGY

The entire framework is shown in Fig. 1, and we describe our architecture in more detail below.

2.1. Human Segmentation Estimation (HSE)

The synthesis of try-on images needs to be guided by the semantic map, hence ensuring the correctness of the semantic map generation will directly affect the final generation effect. The task of human segmentation estimation is to generate a desired semantic map $\hat{I}_p$ with the shape of the target pose T_p. We mix the upper limb, garment, and neck in the reference semantic map into one as an integrated region $I_{\tilde{p}}$ [4]. In the target semantic-map $\hat{I}_p$, the integrated region is re-segmented according to the in-shop garment T_c. We

[†]Feng Yu is corresponding author. Email: yufeng@wtu.edu.cn.

This work was supported by the Young Talents Programme of the Scientific Research Program of the Hubei Education Department (Project No.Q20201709), research on the key technology of flexible intelligent manufacturing of clothing based on digital twin of Hubei key research and development program (Project No.2021BAA042), open topic of engineering research center of Hubei province for clothing information (Project No.900204).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

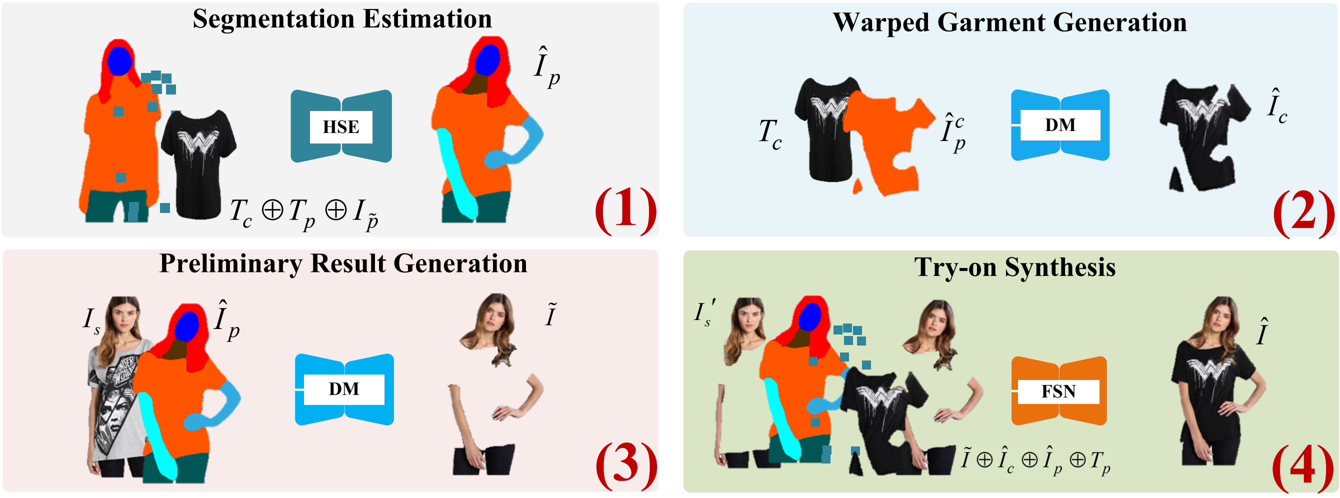
Segmentation Estimation
$T_c \oplus T_p \oplus I_{\tilde{p}}$
HSE
$\hat{I}_p$
(1)
Warped Garment Generation
T_c
$\hat{I}_p^c$
DM
$\hat{I}_c$
(2)
Preliminary Result Generation
I_s
$\hat{I}_p$
DM
$\tilde{I}$
(3)
Try-on Synthesis
I_s'
$\tilde{I} \oplus \hat{I}_c \oplus \hat{I}_p \oplus T_p$
FSN
$\hat{I}$
(4)

Fig. 1. Overview of the MVTON framework. Framework execution order from (1) to (4). It consists of 4 parts: (1) the target semantic map $\hat{I}_p$ in the target pose T_p is generated, (2) the warped garment $\hat{I}_c$ is generated according to the garment semantic-map $\hat{I}_p^c$ and the in-shop garment T_c via the DM, (3) the DM is used to generate the preliminary result $\tilde{I}$ without the garment part, and (4) the desired result $\hat{I}$ is reconstructed by FSN analyses human representations P_t.

adopt U-Net[11] as the generation network and optimize it with the pixel-level cross-entropy $\mathcal{L}_s$ [12].

2.2. Garment & Human-body Deformation Module

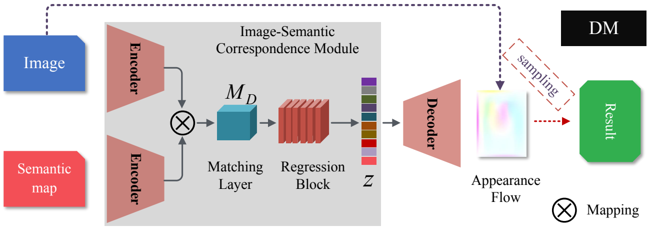

Fig. 2. The structure of Garment & Human-body Deformation Module[4]. The mapping structure between the semantic map and the image generates a stable and efficient appearance flow for warping the given image.

The transfer of the in-shop garment T_c to the target body without losing details is a primary challenge in the virtual try-on tasks. Current methods (e.g., MG-VTON[9], TB-VTON[10]) use the Thin Plate Spline (TPS)[13] to warp the garment in target pose T_p. However, the images generated by this method often result in an excessively warped appearance and loss of detail. In our previous basic work[4], we used a deformation module (see Fig. 2) called Appearance Flow Garment Alignment Network[4], which can warp in-shop garments by corresponding appearance flow information [14] between the semantic map and the garment image.

Also, we found that the results obtained from the preliminary deformation of the original human body using this module can provide important preparatory information for the subsequent pose transfer of the human body. Therefore, we use the original image and the target semantic-map as the input of deformation module[4] to generate the a priori deformation information.

We use the total loss $\mathcal{L}_{total}^{DM}$ consists of pixel-level $\mathcal{L}_1$, VGG perceptual loss $\mathcal{L}_{per}$ [15] and regularization correction loss $\mathcal{L}_c$ [4] to optimize the DM, which can be represented as:

$$\mathcal{L}_{total}^{DM} = \lambda_1 \mathcal{L}_1 + \lambda_{per} \mathcal{L}_{per} + \lambda_c \mathcal{L}_c \tag{1}$$

where λ_* denotes the hyperparameter of the corresponding loss function.

2.3. Filtering Synthesis Network (FSN)

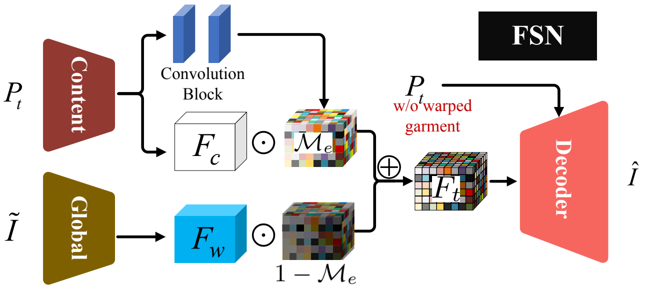

Fig. 3. Overview of Filtering Synthesis Network.

After getting the warped garment, the remaining body parts will be transferred in the target pose T_p. MG-VTON uses a coarse-to-fine framework, where a coarse result is generated using Warp-GAN[16]. Then it further refines the garment part to improve the visual effect. However, MG-VTON lacks optimization of the area outside the garment, which results in defective results. TB-VTON refines the garment area when generating preliminary results and optimizes the face with a specialized network. However, since it lacks the spatial information of the head as a condition, the results generated by TB-VTON often do not retain the characteristics of the original appearance.

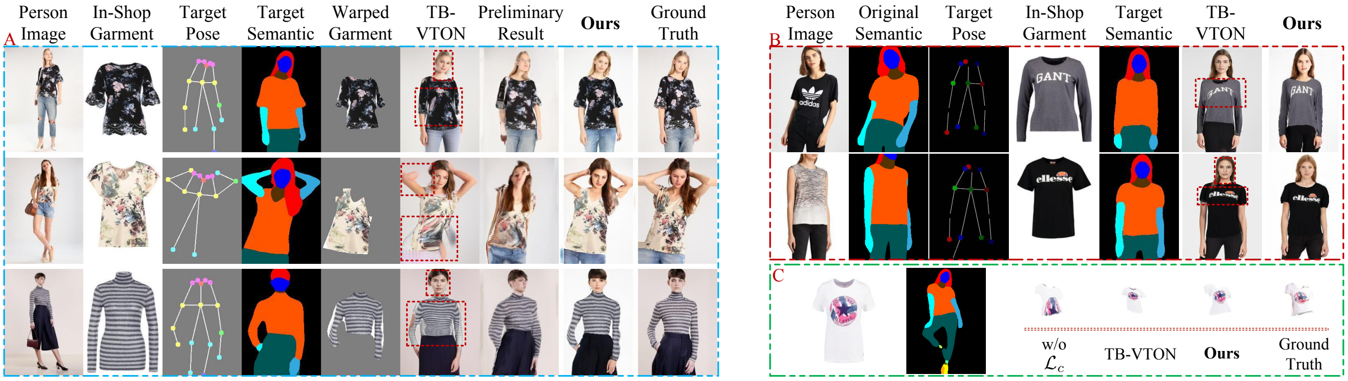

Fig. 4. Visual comparison between MVTON and TB-VTON. Area **A** is the results of pose transfer, area **B** is the results of virtual try-on, and area **C** is the effect of garment alignment. The red dotted boxes represent defects.

To solve the problems above, we design a Filtering Synthesis Network to generate a satisfactory result. As shown in Fig. 3, the encoding part consists of two encoders with the same structure (UNet-like) named content encoder and global encoder. The decoding part consists of seven Spatially-Adaptive Instance Normalization ResBlocks (SAINR) [17]. Between the encoder and decoder in FSN, we introduce a filter matrix (FM) $\mathcal{M}_e \in R^{\mathbf{C\times H\times W}}$ to capture useful information.

We use the content encoder to extract the content-information features F_c of the human representation P_t and the global encoder to extract the global information (e.g., color) features F_w of the preliminary result $\tilde{I}$ without garment to narrow down the solution space. The human representation P_t consists of the reference image without garment I'_s (for obtaining body information), the target pose T_p (for locating facial organs), the warped garment $\hat{I}_c$ (for virtual try-on), and the target semantic-map $\hat{I}_p$ (for guiding the synthesis).

Then F_c is further convoluted to predict a filter matrix $\mathcal{M}_e$, which is used to adaptively filter the effective features between F_c and F_w to compose F_t with all latent information of the desired result. The specific process can be calculated as:

$$F_t = \mathcal{M}_e \odot F_c + (1 - \mathcal{M}_e) \odot F_w \tag{2}$$

where $\odot$ denotes element-wise multiplication, $+$ denotes element-wise addition, and $-$ denotes element-wise subtraction.

The final result $\hat{I}$ can be generated by decoding F_t with FSN's decoder. During decoding, we remove $\hat{I}_c$ from P_t as input to spatially-adaptive instance normalization for capturing the spatial information, positioning information and content information to enhance generation quality. It provides facial localization information and the body's spatial distribution information corresponding to the semantic map for image reconstruction.

During the training phase, the total loss $\mathcal{L}^{FSN}_{total}$ consists of $\mathcal{L}_1$, $\mathcal{L}_{per}$, $\mathcal{L}_{CX}$ [18] and $\mathcal{L}_{adv}$ [19]. It can be expressed as:

$$\mathcal{L}^{FSN}_{total} = \lambda_1\mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{CX}\mathcal{L}_{CX} + \lambda_{adv}\mathcal{L}_{adv} \tag{3}$$

3. EXPERIMENTS

The experiments adopt the MPV dataset[9], which consists of 14,754 garment-person image pairs of size 256 × 192. The MPV contains 12,410 training data groups and 2,344 test data groups. All experiments are carried out on 5 Tesla V100 GPU with 32G RAM. By default, the learning rates is 0.0001 for the generator and the discriminator (same as pix2pixHD[19]) and is linearly reduced to 0 in half the epochs left with a batch size of 4. The framework adopts ADAM optimizer[20] and sets $\beta_1 = 0.5, \beta_2 = 0.999$. During training, the hyperparameters of the $\mathcal{L}^{DM}_{total}$ are set to $\lambda_1 = \lambda_{per} = \lambda_c = 1$, the hyperparameters of the $\mathcal{L}^{FSN}_{total}$ are set to $\lambda_1 = 10, \lambda_{per} = 5, \lambda_{CX} = \lambda_{adv} = 1$.

3.1. Qualitative Results

We conduct a qualitative comparison between the proposed method and the state-of-the-art (SOTA) benchmark method, TB-VTON. Firstly, to verify the effectiveness of DM, we visually compare the warping effect on garments. As shown in Fig. 4 (C). TPS-based TB-VTON produces an unnatural, under-aligned result due to the small number of control points used to warp the entire garment. DM is remarkably effective in controlling garment warping by learning the mapping relationship between the garment semantic-map $\hat{I}^c_p$ and the in-shop garment. The in-shop garment is realistically warped by sampling through a dense flow field. Since the overlapping parts in the semantic map do not exist, redundant pixels are not mapped. It illustrates that the occlusion problem is effectively optimized.

The garment is warped correctly can produce a fine visual result. Recovery of the head, arms, and lower body is also essential as the body transfers to the target pose. As shown in Fig. 4 (A and B), although the TB-VTON refines explicitly for the face, the lack of information on the positioning of the facial organs (e.g., eyes, nose) also leads to a certain lack of facial detail. We use FSN to capture spatial and positioning information, effectively making the details such as the head (e.g., eyes, mouth) and hands in the generated results more realistic. Furthermore, although the preliminary results obtained by DM are heavily distorted, it nevertheless provides sufficient a priori information for FSN to optimize the results of pose transfer. In addition, the filter matrix enables adaptive capture of useful features between coarse warping results and human representations to restore feature details (e.g., hair color, hairstyle).

In summary, the visual results show that MVTON performs well in the multi-pose virtual try-on task. It maintains better detail in the appearance of clothing, faces, etc.

3.2. Quantitative Results

We use Structural SIMilarity (SSIM) [21], Inception Score (IS)[22], Fréchet Inception Distance (FID)[23], and Peak Signal to Noise Ra-

Table 1. Quantitative evaluation between our method and TB-VTON on the MPV dataset. mask- represents the result of removing the background. MVTON† represents the preliminary result $\tilde{I}$, and MVTON represents the result with the original semantic map. A lower FID indicates a better effect.

	IS	mask-IS↑	SSIM↑	mask-SSIM↑	PSNR↑	mask-PSNR↑	FID↓	mask-FID↓
TB-VTON[10]	2.7974±0.1267	2.6854±0.1428	0.6726	0.6913	16.2522	17.5957	22.2014	18.7796
MVTON†	**2.9895±0.1468**	3.0284±0.1712	0.6988	0.7398	16.6161	17.7960	31.0222	19.8827
MVTON	2.9818±0.1831	**3.0509±0.1483**	**0.7842**	**0.7927**	**20.1536**	**20.7856**	**11.9856**	**10.9005**
Real	3.0410±0.1435	3.0564±0.1821	1	1	N/A	N/A	0	0

tio (PSNR) for quantitative evaluation. To reduce the background effect, we calculate mask-SSIM, mask-IS, mask-PSNR, and mask-FID by removing the background. IS is used to assess the sharpness and diversity of the generated image, with higher IS values indicating sharper results and a greater variety of colors and effects. SSIM is a fully referenced image quality evaluation metric that measures image similarity in brightness, contrast, and structure, respectively, with higher values indicating less image distortion. PSNR is used to assess the quality of the resulting image after compression compared to the original image, with higher PSNR values representing less distortion and higher quality when the image is compressed. Because IS only considers the quality of the generated samples and does not consider the influence of the actual data, the FID is used to calculate the distance between the actual samples and the generated samples in the feature space. So that a lower FID value means that the images have higher quality and diversity, as shown in Table 1, our method can achieve convincing scores in the quantitative assessment by adding details such as garment details and human faces. The coarse result $\tilde{I}$ produces the highest calculated IS value as a result of the preservation of background integrity. However, in other indicators, our method achieved the best performance.

3.3. Ablation Study

To demonstrate the necessity and effectiveness of each module in the framework, we remove the filter matrix (FM), SAINR, $\mathcal{L}_1$ loss and $\mathcal{L}_{CX}$ loss in the framework separately. Then we record the corresponding IS, SSIM, PSNR, and FID values, as shown in Table 2. After removing each module, we visually compare the results, as shown in Fig. 5.

The ablation study consists of five parts, which are: 1) FSN without FM, 2) FSN without SAINR, 3) FSN without $\mathcal{L}_1$, 4) FSN without $\mathcal{L}_{CX}$, and full FSN. In summary, target features can be extracted efficiently and accurately by adding the FM. It can bring the results closer to the ground truth. SAINR can also control the generation of complex spatial relationships such as human faces and arms. $\mathcal{L}_1$ effectively enhances the pixel-level details of the generated results. $\mathcal{L}_{CX}$ can maintain the realism of the generated images is retained by comparing the feature context between the generated images and ground-truth. The qualitative and quantitative evaluations show that our method performs better than the existing state-of-the-art methods.

4. CONCLUSION

This paper proposes a multi-pose virtual try-on framework, MVTON, to generate photo-realistic virtual try-on results. First, we use the highly efficient DM to warp the in-shop garment to align the body accurately with the appearance flow. Moreover, we use DM to warp the human body to obtain preliminary results to provide a priori information for FSN. Then, we propose a try-on network, FSN, to generate a multi-pose virtual try-on image with rich original details. Our method proves to be highly efficient in quantitative and qualitative evaluation in terms of quality improvement compared to state-of-the-art methods through experiments by capturing and processing spatial and content information. In the future, we will further improve image generation effectiveness and explore high-performance, lightweight solutions.

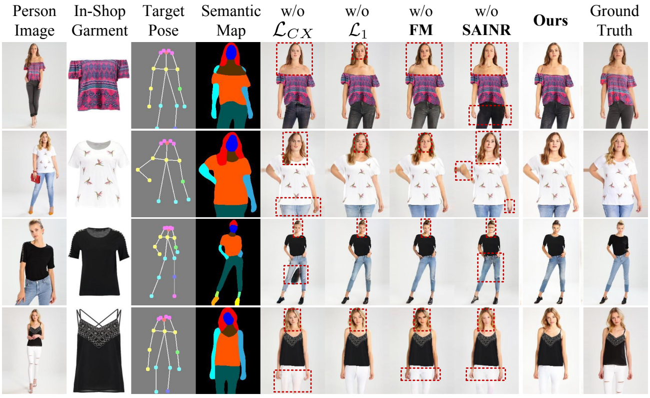

Fig. 5. Ablation study results obtained by different methods, the dashed box represents the defects.

Table 2. Ablation study of our framework. A lower FID indicates a better effect.

	IS	SSIM↑	PSNR↑	FID↓
w/o **FM**	2.9587±0.1894	0.7737	19.5557	13.5334
w/o **SAINR**	2.9151±0.1974	0.7561	18.8569	19.6676
w/o $\mathcal{L}_1$	2.9542±0.1934	0.7785	19.6299	13.0465
w/o $\mathcal{L}_{CX}$	**3.0205±0.1300**	0.7707	19.3102	14.1618
Ours	2.9818±0.1831	**0.7842**	**20.1536**	**11.9856**
Real	3.0410±0.1435	1	N/A	0

5. REFERENCES

[1] Nikolay Jetchev and Urs Bergmann, "The conditional analogy gan: Swapping fashion articles on people images," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2287–2292.

[2] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 589–604.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7847–7856.

[4] Chenghu Du, Feng Yu, Minghua Jiang, Ailing Hua, Xiong Wei, Tao Peng, and Xinrong Hu, "Vton-scfa: A virtual try-on network based on the semantic constraints and flow alignment," *IEEE Transactions on Multimedia*, 2022 (in press).

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser 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, 2021.

[7] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu, "Distribution-aware coordinate representation for human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7091–7100.

[8] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5385–5394.

[9] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9026–9035.

[10] Jiahang Wang, Tong Sha, Wei Zhang, Zhoujun Li, and Tao Mei, "Down to the last detail: Virtual try-on with fine-grained details," in *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, p. 466–474, Association for Computing Machinery.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, vol. 9351, pp. 234–241.

[12] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 932–940.

[13] Jean Duchon, *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, pp. 85–100, 1977.

[14] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.

[15] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] Yichun Shi, Debayan Deb, and Anil K Jain, "Warpgan: Automatic caricature generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10762–10771.

[17] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan 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.

[18] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 768–783.

[19] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan 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.

[20] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*.

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," 2016, vol. 29, pp. 2234–2242.

[23] 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," *Advances in neural information processing systems*, vol. 30, pp. 6627–6638, 2017.