REALISTIC MONOCULAR-TO-3D VIRTUAL TRY-ON VIA MULTI-SCALE CHARACTERISTICS CAPTURE

Chenghu Du[⋆], *Feng Yu*[⋆,†,∗], *Minghua Jiang*[⋆,†], *Yaxin Zhao*[⋆,‡], *Xiong Wei*[⋆], *Tao Peng*[⋆,†], *Xinrong Hu*[⋆,†]

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

ABSTRACT

3D virtual try-on receives widespread attention from scholars due to its great practical and commercial values. In prior methods, the fundamental problems lie in the limitations on texture retention during garment deformation and the lack of feature context capture during depth estimation. To address these problems, we propose a new 3D virtual try-on network via multi-scale characteristic capture (VTON-MC), which can produce an exact 3D model with the generated photo-realistic monocular image. The main processes are as follows: 1) predicting the human semantic-map and aligning the in-shop garment in the human pose using the appearance flow method, 2) synthesizing the human body and the warped garment to gain the image try-on result, and 3) estimating the human double-depth map of the image try-on result to reconstruct desired 3D try-on mesh by designed Depth Estimation Network (DEN). Extensive experiments on existing benchmark datasets demonstrate that VTON-MC outperforms state-of-the-art approaches efficiently.

Index Terms— 3d virtual try-on, appearance flow, depth estimation, semantic map, point cloud

1. INTRODUCTION

With the rapid development of "home clothing shopping," researchers open up two virtual try-on patterns to enhance consumers' shopping experience, which are image-based virtual try-on [1, 2, 3, 4] and 3D reconstruction-based virtual try-on[5, 6], respectively.

The image-based virtual try-on methods align the in-shop garment to the reference pose and synthesize the aligned garment with the human body. State-of-the-art (SOTA) image-based virtual try-on methods generally use geometric matching (Thin Plate Spline (TPS) algorithm [7]) to warp the in-shop garment. However, this method disregards the smooth stretching regular pattern of the fabric when the garment is warped in the actual fitting process. Generally, it aligns the contour with the reference pose at the cost of local overstretching. Ge et al. [4] propose a Disentangled Cycle consistency Try-On Network (DCTON), which produces highly realistic try-on images by disentangling important components of virtual try-on. However, it uses a spatial transformer network (STN) [8] that produces coarse unnatural results resulting in further visual defect of the results. To this end, M3D-VTON[6] proposes self-adaptive pre-alignment to help TPS get a more reasonable warping effect. Nevertheless, the warped results are undoubtedly coarse because the TPS uses the few transformation parameters obtained from the regressor to warp the entire image.

The 3D-based virtual try-on methods are divided into parametric and non-parametric models. SMPL [9] is an effective pattern for implementing parametric 3D try-on[5]. However, the limited detail sculpting makes the generated model lose some of its original appearance details. Zhao et al. [6] propose a monocular-to-3D virtual try-on network (M3D-VTON) that builds on the merits of both 2D and 3D approaches. It pioneers the monocular-to-3D virtual try-on method and gets impressive results. However, the depth estimation network needs to have the ability to correlate features locally and globally in an image to predict more accurate depth maps, which is where it falls short.

To address these challenges, we propose a 3D virtual try-on framework. It replaces the TPS of the coarse warping effect with an appearance flow method to enhance the quality of image try-on results. In addition, the proposed method improves the fineness of the model by having a deep evaluation network with local and global feature correlation capability. The paper's main contributions are summarized as follows: 1) we propose a 3D virtual try-on framework called VTON-MC. It successfully reconstructs the 3D try-on model with more mesh details based on the realistic image-based try-on result, 2) we propose a garment warping strategy. It achieves superior visual results of pixel-wise garment warping by appearance flow, and 3) we propose a depth estimation network called DEN, which can effectively extract and fuse global and local features in images to explore the cascade of feature contexts. It makes the predicted depth information more precise and natural.

2. METHODOLOGY

The general pipeline of 3D virtual try-on is divided into two stages: 1) generating a try-on image $\hat{I}$ by the image-based virtual try-on framework, and 2) reconstructing and rendering 3D person mesh $\hat{V}$ with the try-on image. Fig. 1 shows an overview of our pipeline.

2.1. Image-based Virtual Try-On

2.1.1. Segmentation Generation

The "image-to-image translation"[10] is an effective pattern to gain the desired try-on result by a semantic map. Therefore, the correctness of target semantic-map $\hat{s} \in R^{20\times H\times W}$ generation with the

∗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).

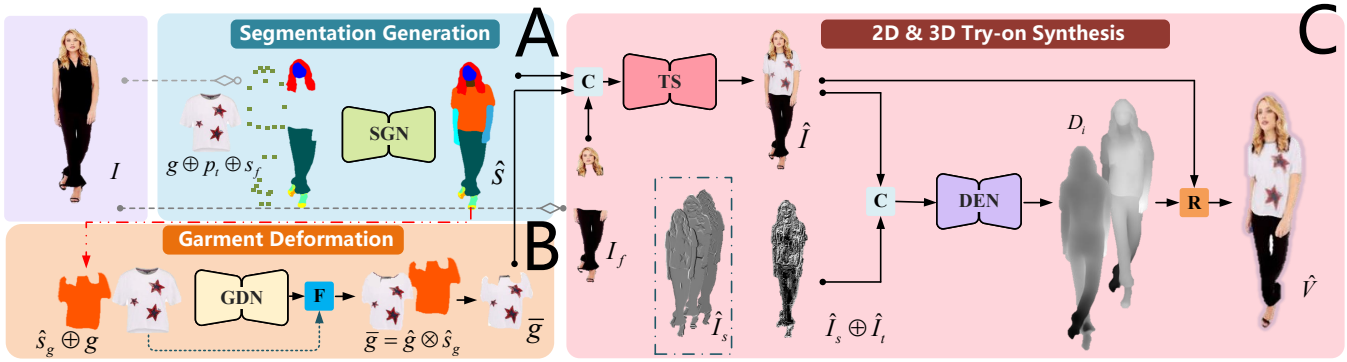

Fig. 1. Overview of the VTON-MC framework. The execution order is from **A** to **C**, where module **A** shows generating target semantic-maps $\hat{s}$. Module **B** shows warping in-shop garments g. Module **C** shows the synthesis process of 2D & 3D virtual try-on. **F** denotes the predicted appearance flow $\mathcal{F}_g$, **C** and $\oplus$ denote channel-wise concatenation, $\otimes$ denotes element-wise multiplication, and **R** denotes mesh's reconstruction process.

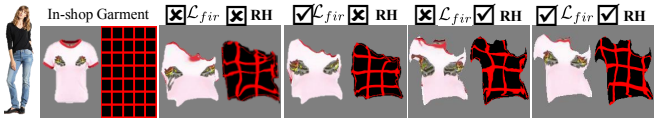

Fig. 2. Validation of the redundancy handling and flow interval restriction roles.

shape of the in-shop garment g is directly related to the final generation effect.

Note that before and after semantic estimation, the areas outside the garment, arms, and neck are unchanged. Therefore, invariant areas (e.g., head, legs) $s_f \in R^{17\times H\times W}$ in reference semantic map $s \in R^{20\times H\times W}$ are directly used as input. In-shop garment g as input to determine the shape of the garment area. In addition, reference pose map p_t is used as input to complement position information of the arms and neck. We adopt U-Net[11, 6] as the target semantic map generation network (SGN), and the pixel-level cross-entropy $\mathcal{L}_s$ [12] is used to optimize the SGN.

2.1.2. *Garment Deformation via Appearance Flow*

During the garment deformation phase, coarsely warped TPS will inevitably occur over-warping (garment boundary beyond the established garment area) and under-warping (garment not fully and reasonably aligned with the body) due to the sparse control points predicted by TPS. Inspired by the appearance flow[13], we propose a garment deformation network (GDN) based on it to solve the problem above. The appearance flow $\mathcal{F}_g \in R^{2\times H\times W}$ is a set of 2D coordinate vectors, and each vector represents the coordinates of the offset relationship corresponding between the in-shop garment g and the desired warped garment $\hat{g}$. The garment flow field changes for any given pose can be effectively predicted by learning garment shape changes. The desired warped garment $\hat{g}$ can then be gained by sampling the in-shop garment g regarding the corresponding dense flow field $\mathcal{F}_g$.

Redundancy Handling (RH). The distribution of fabric pixels in the garment g is smooth; however, no defined garment boundary can result in redundant pixels being sampled. It leads to redundant fabric occupying the other areas causes unnatural garment distortion. We perform a fixed-domain removal of the redundant fabrics in the sampled warped garment $\hat{g}$ before each iteration. The final aligned result $\bar{g}$ is obtained by element-wise multiplication $\otimes$ between the predicted garment semantic-map $\hat{s}_g$ and the warped garment $\hat{g}$. It can be formulated as: $\bar{g} = \hat{g} \otimes \hat{s}_g$.

Flow Interval Restriction. Since garment characteristics (e.g., patterns) do not have dense landmarks, the results of appearance flow estimation are flexible. Not restricting it can cause local over-warping. Inspired by [14], during training, we propose interval restriction loss $\mathcal{L}_{fir}$ to enable that the distance between flow coordinates remains uniform to smooth the warped garment. It can be formulated as:

$$\begin{aligned}\mathcal{L}_{fir} = &\sum_{i=-1,1}\sum_{x}\sum_{y}|\mathcal{P}_x(x+i,y)-\mathcal{P}_x(x,y)| \\ &+\sum_{j=-1,1}\sum_{x}\sum_{y}|\mathcal{P}_y(x,y+j)-\mathcal{P}_y(x,y)|\end{aligned} \tag{1}$$

where $\mathcal{P}_x$ and $\mathcal{P}_y$ are $x-$ and $y-$ coordinates of the flow field $\mathcal{F}_g$, the absolute difference $|a-b|$ is used to metric the distance between two adjacent nodes a and b.

GDN adopts U-Net[11] to generate a 2-channel appearance flow $\mathcal{F}_g$. It takes the in-shop garment g and the predicted garment semantic map $\hat{s_g}$ (provide the desired target pose and shape information) as input, and then iteratively optimizes the network by minimizing the difference between the ground-truth (GT) garment I_g and the warped garment $\bar{g}$. The pixel-level $\mathcal{L}_1$ loss and VGG perceptual loss $\mathcal{L}_{per}$ [15] are used to optimize the network. Finally, the overall loss of GDN is presented as:

$$\mathcal{L}_{GDN} = \lambda_1\mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{fir}\mathcal{L}_{fir} \tag{2}$$

where λ_* denotes the hyperparameter of the corresponding loss function, as shown in Fig. 2, redundancy handling and flow interval restriction have excellent auxiliary effects.

2.1.3. *Try-on Synthesizer*

Synthesizing the warped garment $\bar{g}$ with the fixed body parts I_f from the reference image I to get the try-on result $\hat{I}$ is the final step in the image-based virtual try-on. Recent works guide the synthesis process based on the distribution of the target semantic map $\hat{s}$. In our work, the garment has been aligned perfectly with the human body. Therefore, simply *"copy"* the warped garment and *"paste"* it into the corresponding garment area in the reference image can get the final result. The rest that needs to be addressed is that the gaps in the fabric joints need to be filled, and the missing arms part needs to be generated correspondingly. We adopt the strategy of "image-to-image translation"[10] and use U-Net as a try-on synthesizer (TS).

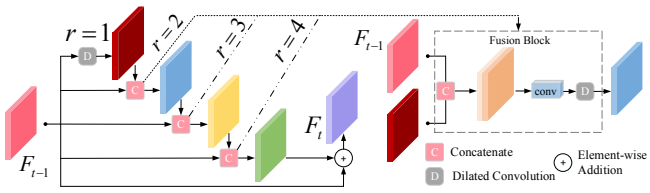

Fig. 3. Structure of stepped multi-scale fusion block (SMFB). r denotes the dilation rate of dilated convolution.

I_f, $\bar{g}$ and $\hat{s}$ as inputs, and the $\mathcal{L}_1$ loss and perceptual loss $\mathcal{L}_{per}$ as the total loss function to optimize the TS. The total loss function $\mathcal{L}_{TS}$ can be expressed as:

$$\mathcal{L}_{TS} = \lambda_1 \mathcal{L}_1 + \lambda_{per} \mathcal{L}_{per} \tag{3}$$

2.2. Monocular-to-3D Virtual Try-On

2.2.1. *Depth Estimation Network (DEN)*

We propose a depth estimation network to estimate the double-depth map D_i (front and back) of the human monocular image to effectively fuse global and local feature information. Its most distinctive characteristic is the inclusion of many stepped multi-scale fusion blocks (SMFB) (see Fig. 3). Inspired by [16], the SMFB is constructed with several multi-scale dilated convolutions with different receptive fields to capture more local detail information and global spatial feature. The input feature map of the SMFB is defined as F_{t-1}, then features of each dilated convolution in SMFB can be calculated as:

$$F_i = W^r \diamond F_{i-1} + b^r \tag{4}$$

where W and b denote parameters of dilated convolution, $\diamond$ denotes dilated convolutional operation. $r = 1, 2, ..., N$, N is dilation rate, F_i and F_{i-1} denote the output and input of dilated convolution. To further fuse the feature information, we combine the input F_{t-1} and the output of dilated convolution F_i. It can be calculated as:

$$F^f_{t-1} = W \star (F_{t-1} \uplus F_i) + b \tag{5}$$

where F^f_{t-1} is the new input of dilated convolution, $\star$ denotes convolutional operation, $\uplus$ denotes the connection of features in the direction of the channel. In SMFB, the last dilated convolution result F_i^{-1} is no longer connected to the input F_{t-1} in the channel direction. Instead, element-wise addition $+$ is performed to reproduce the properties that the standard residual structure [17] has.

$$F_t = F_{t-1} + F_i^{-1} \tag{6}$$

To constitute DEN, we add several SMFBs between the downsampling module and the upsampling module in the UNet-like network.

2.2.2. *3D Try-On Synthesizer*

The generated try-on image needs to find an efficient reconstruction method to get the 3D try-on mesh. PIFu [18] predicts a continuous inside/outside probability field of a clothed human to obtain a non-parametric model. However, it limits the model's accuracy because it uses low-resolution images to cover the contextual information. M3D-VTON [6] reconstructs a 3D mesh by predicting the human body's double-depth map (front and back) from another way. However, more contextual information is needed for the mesh to be accurately reconstructed, and the correct input conditions also determine the detailed performance of the reconstructed model.

In this work, same as [6], we predict the corresponding 3D point cloud by estimating the double-depth map D_i by the image-based try-on image $\hat{I}$ and then triangulate the point cloud by Poisson reconstruction [19] to obtain the 3D try-on mesh. We use the DEN to estimate the double-depth map to cover the more context of input condition information. To better sculpt the depth details, we use the Sobel operator for edge detection ($x-$, $y-$, and $xy-$) on image $\hat{I}$ to add the information of the brightness changes $\hat{I}_s$ in the a priori conditions. In addition, we introduce the Local Binary Pattern (LBP) operator to capture the local texture characteristics (e.g., garment pleats) to get the LBP map $\hat{I}_t$ as the input.

After obtaining the reconstructed 3D mesh from the double-depth map D_i, the image $\hat{I}$ is used directly for the frontal coloring of the mesh since it is aligned with the depth map [6]. After removing the face and neck areas in the image $\hat{I}$, we repair it using the fast matching method [20] and then color the back of the mesh by mirroring it [6] to obtain the 3D try-on mesh model $\hat{V}$.

During training, we take $\hat{I}$, $\hat{I}_s$ and $\hat{I}_t$ as the inputs of DEN, DEN is trained using the $\mathcal{L}_{depth}$ loss [6] to minimize the element-wise depth difference between the estimated result and GT. The $\mathcal{L}_{grad}$ loss [6] is used to capture geometric details differences to optimize the depth estimation. The total loss function $\mathcal{L}_{DEN}$ can be expressed as:

$$\mathcal{L}_{DEN} = \lambda_{depth} \mathcal{L}_{depth} + \lambda_{grad} \mathcal{L}_{grad} \tag{7}$$

3. EXPERIMENTS

The experimental datasets are from two benchmark virtual try-on datasets including VITON[1] and MPV-3D[6]. The VITON contains 16,253 image groups. Each group consists of a front-view female image of size 256×192, an in-shop garment image, and a reference semantic map. VITON contains 14,221 training data groups and 2,032 test data groups. The MPV-3D contains 6,566 garment-person image pairs (g, I) of size 512×320. And, each image has a corresponding front and back depth map. Unlike VITON, MPV-3D has full-body images while VITON has only half-body images. All experiments are executed on 2 Tesla V100 GPU with 16G RAM. By default, the learning rate for generators is 0.0001, and it is reduced linearly to 0 in half of the epochs with a batch size of 6. All experiments adopt ADAM optimizer[21], and the parameters are set as follow: $\beta_1 = 0.5, \beta_2 = 0.999$. In the loss function, $\lambda_1 = \lambda_{per} = 1, \lambda_{fir} = 40$ in $\mathcal{L}_{GDN}$, $\lambda_1 = \lambda_{per} = 1$ in $\mathcal{L}_{TS}$, and $\lambda_{depth} = \lambda_{grad} = 1$ in $\mathcal{L}_{DEN}$.

3.1. Qualitative Results

We perform a visual experimental comparison between baseline methods CP-VTON[2], ACGPN[3], M3D-VTON[6], and our method, as shown in Fig. 4. First, we perform image-based try-on experiments. The results show that the baseline methods suffer from texture confusion and misalignment due to over-warping and under-warping. It is attributed to the coarse warping mechanism of the TPS. During the try-on phase, the problem of texture confusion and loss occurred due to CP-VTON's lack of fixed reference image area and semantic map generation guidance. Moreover, ACGPN appeared to have defective try-on results generated by the incorrect semantic map. While M3D-VTON's pre-alignment strategy effectively pre-aligns the general shape, mechanical alignment without consideration of reasonable deformation angles results in loss of garment detail and unnatural-looking effects. In contrast, guided by the generated correct target semantic map, our GDN based on the

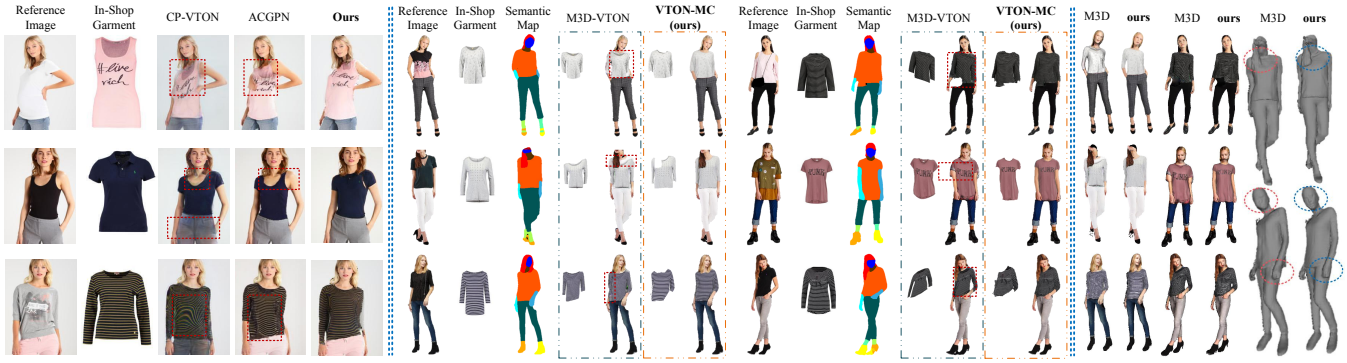

Fig. 4. Comparing our method with the SOTA methods (CP-VTON[2], ACGPN[3], and M3D-VTON[6]). The left side of the double dashed line shows the results on the VITON dataset, and the middle and right side shows the results on the MPV-3D dataset. The 3D reconstruction meshes are on the right side. The red dashed boxes represent the defects in the SOTA methods.

appearance flow has better garment alignment and effectively avoids the above phenomenon.

Table 1. Quantitative evaluation of different methods on VITON (top) and MPV-3D dataset (bottom). Partial data from [4].

Methods//VITON	**IS**[22]↑	**SSIM**[23]↑	**PSNR**↑	**FID**[24]↓
CP-VTON[2]	2.59± 0.13	0.72	16.956	24.45
ACGPN[3]	2.69± 0.12	0.81	23.067	15.67
DCTON[4]	2.85± 0.15	0.83	/	14.82
VTON-MC	2.80± 0.12	**0.89**	**25.236**	**8.93**
Methods//MPV-3D	**LPIPS**[25]↓	**SSIM**[23]↑	**PSNR**↑	**FID**[24]↓
M3D-VTON[6]	0.051	0.9435	25.238	15.74
VTON-MC$^{\times}$	0.036	0.9521	26.915	13.33
VTON-MC	**0.031**	**0.9590**	**28.859**	**11.93**
Real	0	1	N/A	0

Finally, we perform 3D try-on experiments on the MPV-3D dataset between M3D-VTON and our method shown in the right side of Fig. 4. Bad image-based try-on results are mapped directly to the 3D mesh model. Therefore, the defective image result affects the entire garment area of M3D-VTON, resulting in a mesh with defective visual effects. As can be seen from the detailed aspects of the mesh model, such as face and arms, the proposed SMFB can generate a detailed and accurate mesh model based on the fusion of feature contexts, which is where the proposed method has an advantage over M3D-VTON. The subsequent quantitative evaluation can reflect the generated model's specific accuracy.

3.2. Quantitative Results

To study the performance in the benchmark metrics for results. The experiment adopts Structural SIMilarity (SSIM)[23], Inception Score (IS)[22], Fréchet Inception Distance (FID)[24], Peak Signal to Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) [25] to measure generated image quality. Absolute Relative error (Abs.), Squared Relative error (Sq.) and Root Mean Squared Error (RMSE) are used to measure depth estimation results.

Table 1 summarizes the performance between the different methods on image-based virtual try-on. The data reflects that our results are much closer to the ground-truth distribution under all evaluation metrics. Furthermore, Table 2 summarizes the performance between the different methods on 3D virtual try-on. The data show that the proposed SMFB and the LBP map added to the input can reconstruct convincing mesh effects.

Table 2. Quantitative evaluation on MPV-3D. Partial data from [6]. The data were scaled to the same scale as in [6].

Methods	**Abs.**↓	**Sq.**↓	**RMSE**↓
Deephuman[26]	17.35	1.271	22.44
NormalGAN[27]	15.41	0.778	18.94
PIFu[18]	8.376	1.813	27.57
M3D-VTON[6]	7.880	0.385	11.27
VTON-MC‡ (w/o SMFB)	7.798	0.376	11.27
VTON-MC (w/ SMFB)	**7.382**	**0.339**	**10.66**

3.3. Discussion and Ablation Study

To evaluate the effectiveness of the multiple components in VTON-MC, we conduct several ablation experiments. In Table 1, we use VTON-MC to indicate image try-on results with the original semantic map and VTON-MC$^{\times}$ to indicate VTON-MC's results without the semantic map. In Table 2, VTON-MC represents the result with SMFB, and VTON-MC‡ represents VTON-MC without SMFB. Table 1 shows that after removing the target semantic-map, the image quality of try-on results degraded by lack of layout guidance. Table 2 shows that the results of depth estimation are degraded in accuracy by the lack of global and local feature fusion. Fig. 2 shows that the best visualization of the warped garment can only be obtained by introducing both redundancy handling and flow interval restriction.

4. CONCLUSION

In this work, we propose a 3D virtual try-on framework, VTON-MC. The study set out to generate highly accurate 3D try-on mesh models. In this paper, The GDN is proposed to change the traditional garment warping. It achieves superior aligned garment based on appearance flow and further improves the problem of error warping. The SMFB is proposed to capture more contextual information in the try-on images to refine the mesh model. Multi-scale dilated convolution helps the interfusion of different sensory field features. Compared with the most state-of-the-art algorithms, our framework can get the best visual effect and quantitative index. We plan to develop new methods to multi-pose 3D virtual try-on based on more stable reconstruction in the future.

5. REFERENCES

[1] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry 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] 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] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo, "Disentangled cycle consistency for highly-realistic virtual try-on," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16928–16937.

[5] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll, "Learning to transfer texture from clothing images to 3d humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7023–7034.

[6] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang, "M3d-vton: A monocular-to-3d virtual try-on network," *arXiv preprint arXiv:2108.05126*, 2021.

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

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, pp. 1–16, 2015.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei 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.

[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] 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.

[14] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018, pp. 607–623.

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

[16] 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.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2304–2314.

[19] Michael Kazhdan and Hugues Hoppe, "Screened poisson surface reconstruction," *Acm Transactions on Graphics*, vol. 32, no. 3, pp. 1–13, 2013.

[20] Alexandru Telea., "An image inpainting technique based on the fast marching method.," *J. Graphics, GPU, & Game Tools*, vol. 9, no. 1, pp. 23–34, 2004.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[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] 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.

[24] 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.

[25] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[26] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan, "A neural network for detailed human depth estimation from a single image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7749–7758.

[27] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu, "Normalgan: Learning detailed 3d human from a single rgb-d image," in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds. 2020, pp. 430–446, Springer International Publishing.