






MFAR-VTON: Multi-scale Fabric Adaptive Registration for Image-Based Virtual Try-On

Shuo Tong  and Han Liu  

School of Automation and Information Engineering, Xi'an University of Technology,
Xi'an 710048, Shaanxi, China
liuhan@xaut.edu.cn

Abstract. Image-based virtual try-on technology provides a better shopping experience for online consumers. However, existing methods face challenges in effectively capturing high-level semantic information and achieving accurate registration between clothing and the body, particularly in complex body poses or target garments. To address these issues, we propose MFAR-VTON, a novel framework that incorporates a multi-scale enhanced adaptive clothing registration strategy and possesses matching filtering capabilities. Our method enables the generation of highly precise clothing alignment results, leading to seamless integration of try-on images. Additionally, we introduce a deformation energy constraint that effectively preserves intricate garment details. Experimental results demonstrated that MFAR-VTON achieves state-of-the-art performance in terms of accuracy and realism.

Keywords: MFAR-VTON · Virtual try-on · Adaptive clothing registration

1 Introduction

Image-based virtual try-on technology aims to seamlessly replace a person's clothing with in-shop target garment. The virtual try-on pipeline has evolved from initial two-stage [1, 5, 6] to the mainstream three-stage models [2, 14, 15]. The former includes the clothing warping and try-on modules. ACGPN [2] proposes a three-stage model, which introduce a post-try-on segmentation prediction module to provide improved guidance. However, both pipelines strongly depend on the clothing warping stage. Despite advancements in generating realistic try-on results, challenges persist in accurately aligning the clothing and preserving fine clothing details. In traditional garment deformation module, the extracted features are directly fed into the correlation layer to predict Thin-Plate Splines (TPS) parameters, which can only estimate low-complexity parameter transformations, generate rough clothing alignment. Moreover, previous methods compute the features correlation descriptors at a single scale, leading to the loss of some important semantic information.

To address these challenges, we propose the Multi-scale Fabric Adaptive Registration try-on network (MFAR-VTON). In this framework, we have designed a novel geometric matching module called the Multi-scale Neighborhood Consensus Warp Module

(MNCWM). It can extract more comprehensive and rich contextual information, effectively describing the details and shape features of the clothing by performing correlation matching of global semantic patterns across multiple feature scales. Inspired by the NCNet [3], our registration network employs 4D convolutions to refine the correlation feature maps, effectively filtering out geometrically inconsistent correlations, preventing incorrect matches between the clothing and the human body. Furthermore, we design a new end-to-end fabric deformation energy smoothing loss to address the issue of unconvincing distortion in clothing textures. In summary, the main contributions of this work are as follows:

1. We propose a novel image-based virtual try-on network called the Multi-scale Fabric Adaptive Registration Virtual Try-On Network (MFAR-VTON), which generates accurate clothing alignment results and state-of-the-art try-on results.
2. We propose a geometric alignment module, named the Multi-scale Neighborhood Consensus Warp Module (MNCWM), which incorporates multiple scales and employs matching filtering techniques to achieve seamless integration of clothing with body parts.
3. We introduce a fabric warping constraint that enhances the capability to handle complex textures on garments.
4. Experimental results demonstrate that our method outperforms the state-of-the-art approaches in both qualitative and quantitative evaluations.

2 Related Work

2.1 Trainable Image Alignment

Based on deep learning, trainable image alignment methods rely on geometric models, such as affine and thin-plate spline transformations, to estimate geometric transformation parameters through pairwise feature matching. Rocco et al. [3] proposed the Neighborhood Consensus Network (NCNet) by leveraging the idea of semi-local constraint to eliminate ambiguous feature matches. Building upon the NCNet, Li et al. [4] further improved the alignment performance by introducing a self-similarity module.

2.2 Image-Based Virtual Try-On

Image-based virtual try-on methods, which do not rely on 3D scanning devices for data acquisition, have gained significant attention in the academic community. To achieve natural generation effects, it is necessary to warp the clothing based on the reference person's characteristics. Techniques such as VITON [5] and CP-VTON [6] introduced the Thin-Plate Spline (TPS) warping method to deform the clothing. Subsequent models have extended this approach, but traditional clothing warping modules struggle with accurate clothing alignment, leading to issues such as misalignment and ghosting in the generated clothing. Methods based on appearance flow estimation, as in some approaches [7, 8], simulate geometric transformations by estimating dense flows between the source and target clothing regions, enabling flexible clothing warping. However, methods that rely on dense flow often exhibit unstable warping results and may lead to inconsistencies

in rendering. In contrast, our proposed method performs dense semantic correspondence at multiple feature scales, capturing global dense semantic correlations, and filters out ambiguous and erroneous alignment results, resulting in more robust generated result.

3 Proposed Method

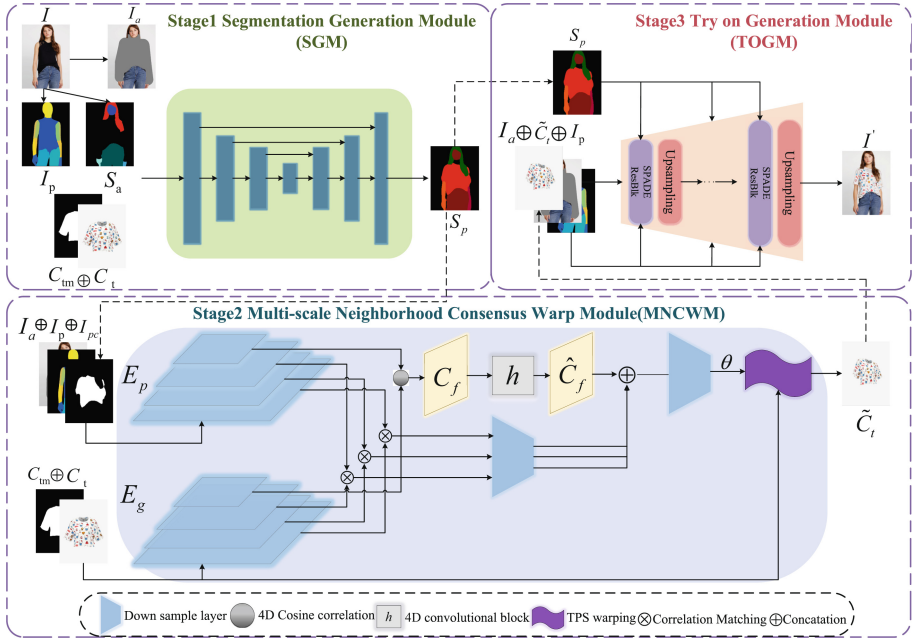


Fig. 1. Overview of MFAR-VTON

Figure 1 illustrates the overall architecture of MFAR-VTON. Given a person image $I \in \mathbb{R}^{H \times W \times 3}$ and a target clothing image $C_t \in \mathbb{R}^{H \times W \times 3}$ as inputs, it aims to seamlessly transfer C_t to the corresponding regions of the reference person I , while preserving the identity details of I unchanged. Our pipeline consists of three stages: (1) Segmentation Generation Module (SGM), which employs a U-net [9] to predict the post-try-on segmentation $S_p \in \mathbb{L}^{H \times W}$; (2) Multi-scale Neighborhood Consensus Warp Module (MNCWM), which performs clothing warping based on multi-scale neighborhood consensus; and (3) Try-on Generation Module (TOGM), which generates the final try-on result $I' \in \mathbb{R}^{H \times W \times 3}$. In the following sections we will describe the specific implementation details of the three-stage pipeline.

3.1 Segmentation Generation Module

As illustrated in Fig. 1, SGM takes the inputs of the cloth-agnostic human segmentation map $S_a \in \mathbb{L}^{H \times W}$, dense pose $I_p \in \mathbb{R}^{H \times W \times 3}$, target garment C_t , and cloth mask $C_{tm} \in$

$\mathbb{L}^{H \times W}$, and employs a U-net architecture to predict the post-try-on segmentation S_p . S_p determines whether the final generated content needs to be preserved or be generated. The total loss l of the SGM is defined as follows:

$$\mathcal{L}_{SSGM} = \lambda_1 \mathcal{L}_{CGAN} + \lambda_2 \mathcal{L}_C \quad (1)$$

where \mathcal{L}_C denotes the pixel-wise cross entropy loss, \mathcal{L}_{CGAN} denotes the conditional adversarial loss, λ_1 and λ_2 are hyperparameters. In this experiment, they are set to 10 and 1, respectively.

3.2 Multi-scale Neighborhood Consensus Warp Module

Multi-scale Neighborhood Consensus Clothing Registration

In this stage, the human representation (I_a, S_{pc}, I_p) and the clothing representation (C_t, C_{tm}) are used as inputs, where the $S_{pc} \in \mathbb{L}^{H \times W}$ denotes the clothing region in the predicted segmentation by SGM. Feature pyramids E_p and E_c are employed to extract multi-scale enhanced features $\{P_l\}_{l=1}^4$ and $\{G_l\}_{l=1}^4$, respectively. Subsequently, multi-scale similarity is computed between $\{P_l, G_l\}_{l=1}^3$. For the top-level features P_4, G_4 we compute the cosine similarity between their pixels exhaustively, resulting in a 4-D correlation tensor $C_{i,j,k,l}^f$:

$$C_{i,j,k,l}^f = \frac{\langle P_{i,j}^4, G_{k,l}^4 \rangle}{\|P_{i,j}^4\|_2 \|G_{k,l}^4\|_2} \quad (2)$$

where $\{i, k\} = 1, \dots, h$, $\{j, l\} = 1, \dots, w$ denote the feature indexing along the height and width directions of P_4 and G_4 . However, the correlation tensor C_f contains a significant amount of matching noise due to incorrect matches. Then, we employ a 4D convolutional filter to filter out the noise and retrieve reliable matching correspondences. Subsequently, we apply soft mutual nearest neighbor filtering to impose global constraints on the matches, obtaining $\bar{C}_{ijkl}^f = r_{ijkl}^{P_4} r_{ijkl}^{G_4} \hat{C}_{ijkl}^f$, where $r_{ijkl}^{P_4} = \hat{C}_{ijkl}^{P_4} / \max_{ab} \hat{C}_{abkl}$, $r_{ijkl}^{G_4} = \hat{C}_{ijkl}^{G_4} / \max_{cd} \hat{C}_{ijcd}$. Finally, the multi-scale correlation feature maps are added to \bar{C}_{ijkl}^f and used as the input to the regression layer to predict TPS transformation parameters $\theta \in \mathbb{R}^{2 \times 5 \times 5}$. θ are used to warp the target clothing, generating warped target cloth $\hat{C}_t \in \mathbb{R}^{H \times W \times 3}$.

Fabric Smoothing Constraint Based on Deformation Energy

The deformation function of TPS interpolation is as follows:

$$f(x, y) = a_0 + a_1 x + a_2 y + \sum_{i=1}^n \omega_i \phi(r_i) \quad (3)$$

where a_0, a_1, a_2 denote the affine transformation parameters, ω_i is the elastic component in elastic deformation, and $\phi(r_i)$ denotes the radial basis function. The cloth deformation energy, denoted as $E(f) = \iint R^2 \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right] dx dy$, is defined as a

measure of the deformation degree. To ensure the realism and natural smoothness of the distorted clothing, it is necessary to impose constraints on the distortion energy. When the radial basis function is chosen as the TPS $\phi(r) = r^2 \log r$, with a proportionality coefficient of α , $E(f)$ satisfies the following equation [10]:

$$E(f) = \alpha \omega^T K \omega$$

$$K = \begin{bmatrix} \phi(r_{11}) & \cdots & \phi(r_{1n}) \\ \vdots & \ddots & \vdots \\ \phi(r_{n1}) & \cdots & \phi(r_{nn}) \end{bmatrix}, w = [w_1, \cdots, w_n]^T \quad (4)$$

we convert it into an end-to-end trainable smooth distortion loss, thereby transforming the problem into $\arg \min_{\omega} (\alpha \omega^T K \omega)$. The total loss in the MNCWM is defined as follows, where β_1 , β_2 , and α are regularization parameters, in this experiment, they are set to 1, 0.1, and 0.005, respectively.

$$\mathcal{L}_{MNCWM} = \beta_1 \|\tilde{C}_t - I_c\|_{1,1} + \beta_2 \mathcal{L}_{vgg}(\tilde{C}_t, I_c) + \alpha \omega^T K \omega \quad (5)$$

3.3 Try-On Generator Module

In this stage, we use the predicted segmentation map S_p as guidance to fuse the warped clothing \tilde{C}_t , dense pose I_p , and clothing-agnostic representation I_a , generate the final try-on result I' . The try-on module consists of several SPADE residual blocks [11] and upsampling layers. Each scale's SPADE normalization layer predicts modulation parameters by S_p . Additionally, the multi-scale inputs (\tilde{C}_t, I_p, I_a) are concatenated with the activation before each residual block for optimization. We utilize the same losses as SPADE and pix2pixHD [12], and the total loss is defined as:

$$\mathcal{L}_{TOM} = \gamma_1 \mathcal{L}_{cGAN} + \gamma_2 \mathcal{L}_{vgg} + \gamma_3 \mathcal{L}_{FM} \quad (6)$$

where \mathcal{L}_{cGAN} represents the conditional adversarial loss, \mathcal{L}_{vgg} denotes the perceptual loss, and \mathcal{L}_{FM} represents the feature matching loss. γ_1 , γ_2 and γ_3 are the regularization parameters, which are set to 1, 10, and 10, respectively, in this experiment.

4 Experiments

4.1 Dataset

VITON-HD: Our experiments are conducted using the VITON-HD [13] datasets under a resolution of 192×256 . The VITON-HD dataset consists of front-view images of women and corresponding front-view clothing images. The dataset includes 11,647 pairs for the training set and 2,033 pairs for the testing set.

4.2 Implementation Details

We trained the SGM stages for 200,000 steps and MNCWM stages for 100,000 steps with a batch size of 8 with Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For the TOGM stage, we trained it for 100,000 steps with a batch size of 4 with Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.9$. The SGM and the MNCWM used the Adam optimizer with a same learning rate of 0.0002, while the discriminator had a learning rate of 0.0002 and 0.0001, respectively. The TOGM used the Adam optimizer with a learning rate of 0.0001, while the discriminator had a learning rate of 0.0004.

4.3 Qualitative Results



Fig. 2. Qualitative comparison on VITON-HD dataset.

Figure 2 shows a qualitative comparison between MNCWM and the latest baselines, HR-VTON [14], and FIFA [15]. On the left side of Fig. 2, it is evident that MFAR-VTON exhibits exceptional performance in the MNCWM module. In the first row, where the target clothing is similar in color to the background (Please note that we adjusted the contrast of the clothing image in the first row to distinguish it from the background, but the clothing image will not be changed during actual training.), the HR-VITON model mistakenly recognizes the clothing’s inner contour as the outer edge. FIFA can only generate blurry result. In contrast, our method can capture features at multiple scales, effectively filter out erroneous registration patterns, improve the model’s sensitivity to complex local features and shape variations, and thus improve the precision of clothing registration. In the second row, where the target clothing is not frontally placed, the compared models incorrectly identify the clothing alignment as short sleeves. MFAR-VTON can accurately aligns the target clothing and generates reasonable try-on results. On the right side of Fig. 2, it can be observed that the deformation energy loss has a more natural constraint on clothing texture deformation, which better preserves fine details of the clothing.

4.4 Quantitative Results

We compared the generated quality of the final try-on results of several state-of-the-art virtual try-on models using the Fréchet Inception Distance (FID) and Structural Similarity (SSIM) metrics. The comparison results on the VITON-HD dataset are presented in Table 1. From the results, it can be observed that our method achieves the best performance in terms of SSIM and FID compared to the advanced methods listed in the table.

Table 1. Quantitative comparisons on VITON-HD dataset

Model	SSIM \uparrow	FID \downarrow
CP-VTON + [1]	0.750	21.08
ACGPN [2]	0.845	16.64
DCTON [16]	0.830	14.82
HR-VITON [14]	0.864	9.38
MFAR-VTON (ours)	0.874	7.11

4.5 Ablation Study

We conducted clothing deformation on paired garments and images, and assessed the disparity between the deformation mask results and the ground truth mask using the Intersection over Union (IOU) metric. As shown in Table 2. It can be observed that MNCWM effectively improves the accuracy of clothing alignment.

Table 2. Ablation study of the MNCWM

Method	IOU \uparrow
w/o MNCWM	0.808
MFAR-VTON	0.822

To demonstrate the impact of the deformation energy constraint on clothing deformation, we conducted experiments by removing the deformation energy constraint from the network architecture and comparing the results with the complete MFAR-VTON model. The results are shown in the Fig. 3 With the deformation energy constraint, local clothing deformations are effectively constrained and smoothed, preserving the details of the target clothing. Without this constraint, exaggerated deformations lead to distorted clothing texture results.



Fig. 3. Ablation study of the deformation energy constraint

5 Conclusion

In this paper, we proposed a novel cloth-adaptive registration try-on architecture, MFAR-VTON, which effectively handles complex poses, accurately aligns clothing deformations, and naturally constrains complex textures of garments to generate photo-realistic virtual try-on results. The experimental results demonstrated that our method outperforms the state-of-the-art approaches both in qualitative and quantitative.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grants 92270117, 61973248.

References

1. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P.: Cp-vton+: clothing shape and texture preserving image-based virtual try-on. In: Proc. of the IEEE international conference on computer vision workshop (ICCVW). pp. 10–14 (2020)
2. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 7850–7859 (2020)
3. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Advances in neural information processing systems, pp. 1658–1669 (2018)
4. Li, S., Han, K., Costain, T. W., Howard-Jenkins, H., Prisacariu, V.: Correspondence networks with adaptive neighbourhood consensus. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 10196–10205 (2020)
5. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L. S.: Viton: An image-based virtual try-on network. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 7543–7552 (2018)
6. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proc. of the European conference on computer vision (ECCV), pp. 589–604 (2018)
7. Chopra, A., Jain, R., Hemani, M., Krishnamurthy, B.: Zflow: gated appearance flow-based virtual try-on with 3d priors. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 5433–5442 (2021)

8. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: a flow-based model for clothed person generation. In: Proc. of the IEEE international conference on computer vision (ICCV), pp. 10471–10480 (2019)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for bio-medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
10. Wahba, G.: Spline models for observational data. Society for industrial and applied mathematics (1990)
11. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 2337–2346 (2019)
12. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 8798–8807 (2018)
13. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 14131–14140 (2021)
14. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In: Proc. of the European conference on computer vision (ECCV), pp. 204–219 (2022)
15. Zunair, H., Gobeil, Y., Mercier, S., Hamza, A.: Fill in Fabrics: Body-Aware Self-Supervised Inpainting for Image-Based Virtual Try-On. In: BMVC (2022)
16. Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., Luo, P.: Disentangled cycle consistency for highly-realistic virtual try-on. In: Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 16928–16937 (2021)