GT-MUST: Gated Try-on by Learning the Mannequin-Specific Transformation

Ning Wang[*]
Wuhan University
Wuhan, China
wang_ning@whu.edu.cn

Jing Zhang
The University of Sydney
Sydney, NSW, Australia
jing.zhang1@sydney.edu.au

Lefei Zhang[†]
Wuhan University
Hubei Luojia Laboratory
Wuhan, China
zhanglefei@whu.edu.cn

Dacheng Tao
JD Explore Academy
Beijing, China
dacheng.tao@gmail.com

ABSTRACT

Given the mannequin (i.e., reference person) and target garment, the virtual try-on (VTON) task aims at dressing the mannequin in the provided garment automatically, having attracted increasing attention in recent years. Previous works usually conduct the garment deformation under the guidance of "shape". However, "shape-only transformation" ignores the local structures and results in unnatural distortions. To address this issue, we propose a Gated Try-on method by learning the ManneqUin-Specific Transformation (GT-MUST). Technically, we implement GT-MUST as a three-stage deep neural model. First, GT-MUST learns the "mannequin-specific transformation" with a "take-off" mechanism, which recovers the warped clothes of the mannequin to its original in-shop state. Then, the learned "mannequin-specific transformation" is inverted and utilized to help generate the mannequin-specific warped state for a target garment. Finally, a special gate is employed to better combine the mannequin-specific warped garment with the mannequin. GT-MUST benefits from learning to solve a much easier "take-off" task to obtain the mannequin-specific information than the common "try-on" task, since flat in-shop garments usually have less variation in shape than those clothed on the body. Experiments on the fashion dataset demonstrate that GT-MUST outperforms the state-of-the-art virtual try-on methods. The code is available at https://github.com/wangning-001/GT-MUST.

CCS CONCEPTS

• **Computing methodologies → Computer vision tasks**.

KEYWORDS

virtual try-on, deep learning, inverse transformation, parse-based

ACM Reference Format:
Ning Wang, Jing Zhang, Lefei Zhang, and Dacheng Tao. 2022. GT-MUST: Gated Try-on by Learning the Mannequin-Specific Transformation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547775

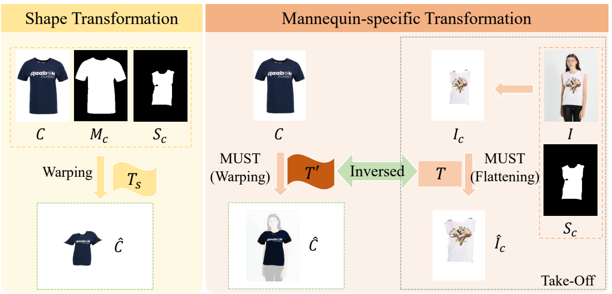

Figure 1: Left: The shape transformation simply conducts the garment transformation under shape constraint . Right: GT-MUST learns the mannequin-specific transformation by solving an easier "take-off" task first. The learned feature T provides the per-pixel tailored guidance for the target garment deformation, i.e., the inverse transformation T'.

1 INTRODUCTION

In recent years, online apparel and accessories shopping has become a new trend for the fast and convenient purchase journey in the context of rich and diverse garment resources [25, 42]. **Virtual try-on (VTON)**, which aims to put an image of a target garment on an image of a reference person, is regarded as a crucial task in online fashion shopping to narrow the gap between online and offline shopping experience. With the rapid development of image synthesis technology, 2D image-based try-on methods, like VITON [13], CP-VTON [36], ACGPN [38], and PF-AFN [9], have attracted increasing attention in recent years.

Current solutions for the image-based try-on task [7, 13, 38] follow a similar two-step pipeline: 1) generating the warped clothes image $\hat{C}$ according to the in-shop one C (i.e., target clothes); and 2) merging the warped clothes $\hat{C}$ with the mannequin I. Although the popular scheme can ensure an approximate fit between the

[*] This work was done during Ning Wang's internship at JD Explore Academy. She is affiliated with School of Computer Science, Wuhan University.
[†] Corresponding author. He is also affiliated with Hubei Key Laboratory of Multimedia and Network Communication Engineering, School of Computer Science, Wuhan University.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547775

target garment and the mannequin in the generated try-on image, they are somehow weak, due to the unreasonable deformation. Our human eyes are sensitive to the abnormal distortion as it contradicts the human visual perception of the natural world, thus modeling and simulating the deformation of clothes is very important in the VTON task. Some techniques are proposed in prior studies to retain the shape and structure of the target clothes, e.g., the second-order difference grid constraint in ACGPN [38] and the appearance online optimization strategy of O-VITON [30].

Nevertheless, the abnormal distortion may still exist. We argue that the main reason is only "shape transformation" is learned in the common try-on scheme, resulting in mismatches between mannequins and warped clothes. Specifically, the current methods [13, 29, 36, 38] always estimate the clothes transformation with the aid of "shape" like clothes mask. While being intuitive, these "shape transformation" constraints fail to consider the "mannequin-specific" deformation, which can take the characteristic of each mannequin into account for a per-pixel tailored transformation. Without specific knowledge about the target deformation, the shape-based warping process prefers to simply predict a transformation lying on the center of the learned transformation distribution, i.e., all the transformation-related details will be smoothed [43]. Therefore, these methods are weak in generating reasonable deformation and retention of clothes textures.

To overcome this problem, one important step is to obtain extra knowledge regarding the target deformation, i.e., the per-pixel tailored information that tells the network where the output's location should be. As the clothes' deformation manner is dependent on the posture and stature of the mannequin, we thus assume that the deformation in one garment, i.e., "mannequin-specific transformation" (MUST), can be transferred to the other given the same mannequin pose. The comparison of the common "shape transformation" and the proposed MUST is shown in Figure 1. The common "shape transformation" methods directly learn the warping for target garments. Although it looks straightforward, it is difficult to obtain accurate transformation as the guidance is not sufficient. We, on the contrary, design the novel MUST by learning an easier "take-off" task. As the mannequin should keep the same stature after "try-on", the per-pixel warping should be generally the same after "try-on" for any clothes. We can obtain the per-pixel tailored information MUST T with clothes flattening by learning an inverse task, i.e., "take-off". Then, the MUST T is employed as new per-pixel guidance for more accurate garment warping.

Based on the above motivation, in this paper, we propose a novel "mannequin-specific transformation" based gated try-on model named GT-MUST, which adopts a take-off and put-on strategy. GT-MUST contains three modules, namely, Inverse Learning Module (ILM), Mannequin-specific Warping Module (MWM), and Gated Try-on Module (GTM). Different from other try-on models, we first devise a take-off mechanism in ILM to learn the inverse clothes transformation by learning the warped state to flat state. A mannequin-specific transformation T is learned in this module accordingly. Then, under the guidance of the T, we can generate an inverse transformation T' in MWM that is able to construct mannequin-specific warped clothes for any in-shop clothes. Due to the style differences between the target garment and the garment on the mannequin, especially the different sleeve lengths, it is not feasible to attach the warped clothes generated by MWM to the mannequin directly. Thus, we leverage a special gate mechanism in the GTM to better retain clothes information and mannequin stature in the final output. The main difference between the proposed GT-MUST method and the other methods is that we employ the "take-off" task first to provide supervision for the inverse mannequin-specific transformation T' while the other methods usually use the "shape" transformation that lacks constraints on clothes content.

The main contributions of this paper are three-fold:

- We propose a novel idea, i.e., learning mannequin-specific transformation (MUST) by take-off first. The MUST can provide guidance for its inverse transformation, constructing mannequin-specific warped clothes for any in-shop clothes. It provides a new perspective for virtual try-on by learning to solve an inverse yet much easier "take-off first" task.
- We leverage a special gate mechanism to measure the differences between input and generated features, which determines the regions of garments to be retained and the features to be employed. The gate allows to preserve more clothes information and fit the shape of the mannequin better.
- We provide a lightweight and effective model. Experiments demonstrate that the proposed GT-MUST can generate photo-realistic try-on results and outperform the state-of-the-art methods both qualitatively and quantitatively.

2 RELATED WORK

VTON with paired data. The current VTON methods [3, 13, 21, 24, 28, 29, 36, 38, 39, 45] usually use paired data. CAGAN [21] first explores the VTON task based on the cycle-GAN [46] with paired data. However, in addition to the reference person image and the target in-shop clothing image, CAGAN also needs the in-shop clothing image worn by mannequin, which is hard to provide in practice. Later, VITON [13] eliminates the requirement of the in-shop garments worn by mannequins with the clothing-agnostic person representation. Similar to VITON, more and more VTON methods employ the clothing-agnostic person representation, including CP-VTON [36], VTNFP [39], CP-VTON+ [29], and ACGPN [38], and they share the common warp-and-refine pipeline. However, they only consider the shape transformation in the clothes warping step, resulting in distorted textures in the final output. To address this issue, some constraints are devised to improve garment warping, e.g., the gird interval consistency loss in LA-VITON [20] and second-order difference constraint loss in ACGPN [38]. ZFlow [3] also follows the warp-and-refine pipeline but leverages per-pixel content transformation to improve textural integrity for garment warping, instead of the shape transformation used by the above-mentioned methods.

Our method also uses pair data but solves the VTON task from a novel perspective i.e., learning a mannequin-specific content transformation by including a "take-off" stage before the common warp-and-refine pipeline. It is very different from the shape transformation and content transformation that always focus on the clothes warping and ignore the changes of person, although the try-on task is more dependent on the mannequin statures and postures.

VTON with unpaired data. Some researchers try to solve the VTON task with unpaired training data, which is more convenient

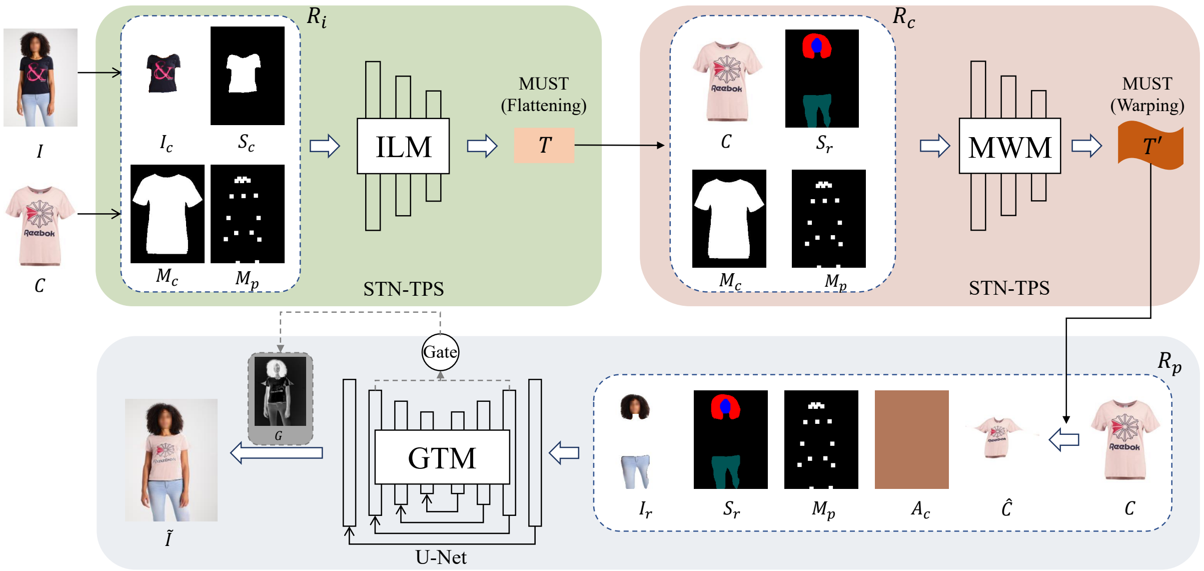

Figure 2: The overall architecture of GT-MUST. Mannequin-specific transformation T is learned by solving the inverse task of virtual try-on, i.e., take-off, in the ILM. Then T is utilized to get the inverse transformation T' for generating mannequin-specific warped clothes $\hat{C}$ in MWM. Finally, the virtual try-on result $\tilde{I}$ is generated in GTM with a garment-preserving gate G.

in practical applications. O-VITON [30] and TryOnGAN [26] propose to swap the latent codes of garments. While being effective, the feature-wise operation may result in distortions in delicate clothes structures, like logos and embroidery. WUTON [18] uses a teacher-student architecture to get rid of any requirements for information other than reference person and target garment information. Similarly, PF-AFN [9] adopts the "teacher-tutor-student" knowledge distillation scheme and employs the appearance flows to guide the "content" transformation, which can produce highly photo-realistic results without using human parsing [14] as model input during inference. Furthermore, He et al. [15] devise the style-based global appearance flows to improve local garment deformation.

Fashion synthesis. Some fashion synthesis methods can also be used for VTON. FiNet [12] employs the image inpainting to synthesize a clothing item for the masked region with natural texture that matches the surrounding garments. ClothFlow [11] proposes a pose-guided person generation method related to human pose estimation [37, 40], which can generate pose transfer and single-pose try-on. DiOr [5] utilizes a recurrent generation pipeline to enable different interactions of garments in several fashion editing tasks. wFlow [8] integrates 2D pixel flow and 3D vertex flow for dynamic try-on in the wild. Moreover, some methods incorporate the try-on task with pose transfer, like parsing generator [41], conditional parsing generator [7], the pose-guided parsing translator [4, 17].

3 METHOD

Given a mannequin image $I \in \mathbb{R}^{h\times w\times 3}$ and a target garment image $C \in \mathbb{R}^{h\times w\times 3}$, the proposed GT-MUST model aims to generate an image $\tilde{I} \in \mathbb{R}^{h\times w\times 3}$ depicting the reference person wearing the target clothes. As shown in Figure 2, GT-MUST is composed of three modules: Inverse Learning Module, Mannequin-specific Warping Module, and Gated Try-on Module. Note that, the training process is conducted with paired data, i.e., the target garment C in Figure 2 is the same clothes worn by the reference person during training.

3.1 Inverse Learning Module

ILM is designed to extract the MUST from mannequin, which has not been explored in previous works. Specifically, we adopt an inverse take-off strategy in this module to extract the mannequin-specific content transformation of the specific clothes warping manner according to the reference person's worn clothes.

Given a reference person image I and its corresponding parsing label S, we can extract the mask (region) S_c of the worn clothes by the reference person according to the clothes labels in S, i.e., $S_c = (S == L_c)$, where L_c is the label value that represents the upper garment. Then the clothes I_c worn by the reference person can be obtained by calculating the Hadamard product of reference person I and clothes mask S_c, i.e., $I_c = I \odot S_c$. Then, the clothes I_c from the reference person, the clothes region S_c, the clothes mask M_c of target garment, and the pose M_p of reference person are concatenated as inverse representation R_i and fed into the ILM to generate the mannequin-specific transformation T, as shown in Figure 2. The ILM is composed of a Spatial Transformation Network (STN) [19] with a TPS [1] transformation, which is widely adopted in the text recognition field [34]. Using the STN-TPS combination has been proven to be a successful strategy in other virtual try-on methods like ACGPN [38]. Specifically, denoting the ILM module as $F_{ILM}(\cdot)$, the mannequin-specific transformation T is calculated as follows:

$$T = F_{ILM}(R_i). \tag{1}$$

With the mannequin-specific transformation T, we can generate the in-shop clothes $\hat{I}_c$, i.e., $\hat{I}_c = F_{flow}(I_c, T)$, where the F_{flow} denotes grid sampling using the mannequin-specific transformation T. The generated $\hat{I}_c$ should be close to the given target garment C.

In this way, ILM learns to take the clothes I_c off from the reference person image I and recover it to the in-shop state C, i.e., implicitly learning this mannequin-specific transformation in the form of T. We will detail the training of ILM in Section 3.4.

3.2 Mannequin-specific Warping Module

After getting the mannequin-specific transformation T from the ILM, we devise the MWM to generate warped clothes of the target garment with the help of T. In line with the design of the ILM, we adopt the STN-TPS architecture to warp the target garment into a suitable shape. Technically, we first define a garment representation R_c, which consists of four components, namely the target garment C, the clothes mask M_c of target garment, the pose map M_p of reference person, and the retained parsing label region S_r of reference image I that contains head region and lower half of the body region in human parsing label. Meanwhile, T is treated as another input. As we only concern the transformation of the cloth region, we mask the T by M_c as $T_c = T \odot M_c + (1 - M_c)$ and use T_c as input instead. Then, the inverse mannequin-specific content transformation T' can be predicted by the MWM as follows:

$$T' = F_{MWM}(T_c, R_c). \tag{2}$$

T' can be used to transform the flat garment to the corresponding warped state, i.e., $\hat{C} = F_{flow}(C, T')$, $\hat{M_c} = F_{flow}(M_c, T')$, where F_{flow} denotes the bilinear warping.

Ideally, the generated $\hat{C}$ should be close to the desired warped clothes I_c. In addition, we also enforce two extra shape-based constraints to ensure that MWM predicts correct T' and generates reasonable $\hat{C}$. First, the matching clothes $\hat{C}_m = \hat{C} \odot S_c + (1 - S_c)$ in the generated clothes $\hat{C}$ based on the corresponding worn-out clothes region should be also close to I_c. Second, $\hat{M_c}$ should be close to the clothes mask S_c as well. We leverage these constraints to train MWM, which will be detailed in Section 3.4.

3.3 Gated Try-on Module

In the final step, we propose the GTM to fuse the warped clothes with the mannequin. We adopt the person representation R_p as input, which is composed of 1) the reserved information I_r of reference image I during the training process, i.e., $I_r = I \odot M_r$, 2) the retained parsing label region S_r, 3) the generated clothes $\hat{C}$ by MWM, 4) the given in-shop clothes C, 5) the simulated color A_c of arms region, which is generated based on the average arms color of the reference person, providing the arms information for the final try-on image $\tilde{I}$, and 6) the pose map M_p.

The GTM is based on the U-Net [31] with skip connections, which can preserve some useful information in the input, e.g., I_r is unchanged after VTON. Based on R_p, the GTM aims to generate the final person image that the reference person wears the target clothes. Denoting the GTM as $F_{GTM}(\cdot)$, this process can be formulated as follows:

$$\tilde{I} = F_{GTM}(R_p), \tag{3}$$

where the $\tilde{I}$ is the final output of our GT-MUST model.

To preserve more clothes information and fit the shape of the mannequin better, we leverage a gate mechanism F_{gate} to measure the differences between early and late features of U-Net. Passing the input R_p into U-Net, we can obtain the feature f_1 after the first layer and feature f_2 before the final layer. Before measuring difference, we adopt two convolution layers to process f_1 and f_2 respectively. Then the difference is sent into an additional convolution layer and a sigmoid layer to obtain the gate. It can be formulated as follows:

$$G = F_{gate}(R_p). \tag{4}$$

Then the try-on result $\tilde{I}$ is generated as follows:

$$\tilde{I} = f_3 \odot G + \hat{C} \odot (1 - G), \tag{5}$$

where f_3 is the feature of the last layer of U-Net.

The generated try-on image $\tilde{I}$ is desired to retain the identity of the reference person I and wear the target clothes C correctly. Since for a mannequin, there is only a target garment and a reference person image in the dataset, we, therefore, employ the corresponding clothes of reference person to implement supervised learning, i.e., the target garment C is the in-shop version of warped clothes I_c. In this way, the final try-on image $\tilde{I}$ should be the same as the reference person I.

3.4 Loss Function

Following prior works [9, 38], we leverage several losses to train the proposed GT-MUST model, including the MAE loss $\mathcal{L}_1$, perceptual loss $\mathcal{L}_{perc}$, style loss $\mathcal{L}_{sty}$ [22], and the second-order difference constraint $\mathcal{L}_{const}$ [38]. We also devise a gate loss $\mathcal{L}_{gate}$, which will be detailed later.

Loss functions for the ILM. The ILM aims to learn the mannequin-specific content transformation from the warped clothes of the reference person, i.e., transforming the warped clothes I_c back to its in-shop version $\hat{I}_c$. Therefore, we can define the training objective of the ILM as follows:

$$\begin{aligned} \mathcal{L}_{ILM} = \lambda_1 \mathcal{L}_1(C, \hat{I}_c) + \lambda_{sty} \mathcal{L}_{sty}(C, \hat{I}_c) \\ + \lambda_{perc} \mathcal{L}_{perc}(C, \hat{I}_c) + \mathcal{L}_{const}(\hat{I}_c). \end{aligned} \tag{6}$$

Specifically, denoting $\Phi_i(x)$ as the feature map extracted from the ith layer of VGG-19 [35], the perceptual loss $\mathcal{L}_{perc}$ is defined as,

$$\mathcal{L}_{perc}(C, \hat{I}_c) = \sum_{i=1}^{N} \mathbb{E} \parallel \Phi_i(C) - \Phi_i(\hat{I}_c) \parallel_2^2, \tag{7}$$

where N is the number of feature maps extracted from VGG-19 pre-trained on ImageNet [6], and $\mathbb{E}$ is an abbreviation for the averaging operation.

By vectoring the feature map $\Phi_i(x)$ with shape $C_i \times H_i \times W_i$ into a $\Psi_i(x)$ with shape $C_i \times H_i W_i$, the Gram matrix $\Theta_i(x)$ is calculated as $\Theta_i(x) = \Psi_i(x) \Psi_i(x)^T$. Then the style loss $\mathcal{L}_{sty}$ is defined as the squared Frobenius norm of the difference between Gram matrices of two images, i.e.,

$$\mathcal{L}_{sty}(C, \hat{I}_c) = \sum_{i=1}^{N} \mathbb{E} \parallel \Theta_i(C) - \Theta_i(\hat{I}_c) \parallel_F^2 . \tag{8}$$

Inspired by ACGPN [38], we further adopt the second-order difference constraint $\mathcal{L}_{const}$ to ensure the geometric matching of $\hat{I}_c$, i.e.,

$$\begin{aligned} \mathcal{L}_{const}(\hat{I}_c) = \sum_{p \in \mathrm{P}} max(\lambda_n \sum_{i=0,2} \mid \parallel pp_i \parallel_2 - \parallel pp_{i+1} \parallel_2 \mid \\ + \lambda_s \sum_{i=0,2} \mid S(p, p_i) - S(p, p_{i+1}) \mid - \delta, 0), \end{aligned} \tag{9}$$

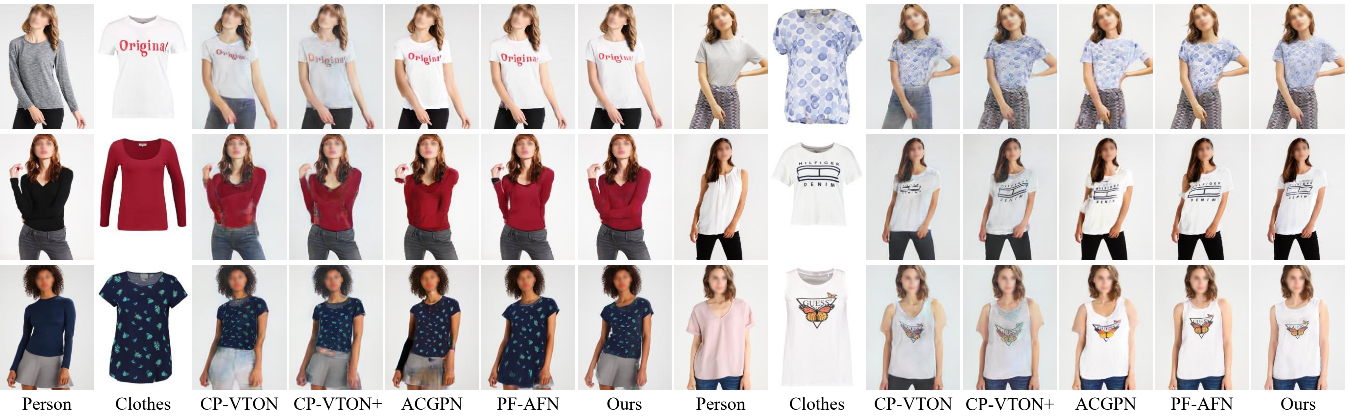

Figure 3: Qualitative comparisons on the VITON dataset. Compared with the state-of-the-art virtual try-on methods, including CP-VTON [36], CP-VTON+ [29], ACGPN [38], and PF-AFN [9], our GT-MUST generates more realistic try-on images.

where p denotes the sampled control point and $\mathbf{P}$ represents the sampled control points set for $\hat{I}_c$. p_0, p_1, p_2, p_3 are the top, bottom, left and right sampled control points of p, respectively. $S(p, p_i)$, $(i = 0, 1, 2, 3)$ represents the slope between two points p and p_i, which may be zero if the selected point p lie on the boundary. To avoid the divided-by-zero error, the slopes are calculated as follows,

$$\begin{aligned} &|S(p,p_i) - S(p,p_{i+1})\ | \\ &=|\ (y_i - y)(x_{i+1} - x) - (y_{i+1} - y)(x_i - x)\ |, \end{aligned} \tag{10}$$

where (x_i, y_i) is the location of p_i and $i \in \{0, 2\}$. λ_n and λ_s are used to balance the two terms.

Loss functions for the MWM. Similar to the ILM, the MWM also employs the $\mathcal{L}_1, \mathcal{L}_{sty}, and \mathcal{L}_{perc}$ losses during training. Besides, we also introduce a mask loss to preserve the shape of clothes. Finally, the training objective of the MWM is defined as follows:

$$\begin{aligned} \mathcal{L}_{MWM} = \mathcal{L}_1(S_c, \hat{M}_c) + &\sum_{\Omega\in\{\hat{C},\hat{C}_m\}} |\ \lambda_1\mathcal{L}_1(I_c, \Omega) \\ &+\lambda_{sty}\mathcal{L}_{sty}(I_c, \Omega) + \lambda_{perc}\mathcal{L}_{perc}(I_c, \Omega)\ |\ . \end{aligned} \tag{11}$$

Loss functions for the GTM. For the GTM, we adopt similar losses to the previous modules, including $\mathcal{L}_1, \mathcal{L}_{sty}, \mathcal{L}_{perc}$. Besides, as the U-Net tends to generate blurry textures for the final garment, we prefer to use the clothes information from the $\hat{C}$ via the gate mechanism. To this end, we also devise a gate loss $\mathcal{L}_{gate}$ to enforce a constraint on the magnitude of G, which is defined as follows,

$$\mathcal{L}_{gate}(G) = \mathrm{Mean}(G), \tag{12}$$

where $\mathrm{Mean}(\cdot)$ denotes the average operator. The total training objective of the GTM is defined as follows:

$$\begin{aligned} \mathcal{L}_{GTM} = \lambda_{gate}\mathcal{L}_{gate}(G) + &\sum_{\Lambda\in\{\tilde{I},I^{\dagger}\}} |\ \lambda_1\mathcal{L}_1(I, \Lambda) \\ &+\lambda_{sty}\mathcal{L}_{sty}(I, \Lambda) + \lambda_{perc}\mathcal{L}_{perc}(I, \Lambda)\ |, \end{aligned} \tag{13}$$

where $I^{\dagger}$ is a composed try-on image, i.e., $I^{\dagger} = (1 - M_r) \odot \tilde{I} + I_r$. M_r is the reserved region that contains the head region and lower half of the body region, i.e., $M_r = (S_r > 0)$. And I_r denotes the reserved information, i.e., $I_r = M_r \odot I$. Therefore, $I^{\dagger}$ replaces the clothes and arms information of I with generated information from corresponding regions of $\tilde{I}$. We employ it to enforce the consistency with the reference image.

4 EXPERIMENTS

In this section, we first introduce the dataset and implementation details. Then, we show qualitative and quantitative results of the proposed GT-MUST and state-of-the-art try-on methods. Finally, we present the ablation study and user study of GT-MUST. Note that, all calculations and evaluations are based on the original images. And the faces of the mannequins are blurred in the text for identity privacy protection.

4.1 Dataset

We evaluate GT-MUST on the VITON Zalando dataset [13], which is commonly used for benchmarking virtual try-on methods [9, 29, 36, 38]. The VITON dataset consists of 14,221 pairs of training images and 2,032 pairs of testing images. Each pair of images contains an in-shop garment image, a person image wearing the same garment, a binary mask of the in-shop garment, a human parsing label generated by LIP [10], and a human pose map by OpenPose [2]. The resolution of these images is 256 × 192. MPV is another dataset used for virtual try-on task that contains 35,687 persons and 13,524 clothes images at 256 × 192 resolution. 4,175 image pairs are split out as the test set in MPV. Different from VITON, MPV contains multiple views of mannequins, including half front, whole front, half back, and whole back views. We remove the back views images from MPV, since the clothes are only given from the front views.

4.2 Implementation Details

Training. Due to the lack of ground truth for arbitrary clothes try-on images, we use the in-shop version of the reference image as the target garment to conduct the supervised learning during the training process. The learning rate is initialized as $2e^{-4}$, and we use the Adam optimizer [23] with $\beta_1 = 0.9, \beta_2 = 0.999$ to optimize the model. All models are trained on Unbuntu 20.04 with AMD EPYC 7742 3.40GHz CPU and a 40G NVIDIA A100 GPU. The batch size is set to 16. The $\lambda_1, \lambda_{sty}, \lambda_{perc}$ are hyper-parameters, which are set

Methods	FID ↓	LPIPS ↓	User Study
CP-VTON [36]	24.43	0.126	11.44%/88.56%
CP-VTON+ [29]	16.31	0.117	11.74%/88.26%
ACGPN [38]	14.11	0.102	18.32%/81.68%
PF-AFN [9]	10.07	0.095	27.06%/72.94%
GT-MUST	**8.31**	**0.061**	reference

Table 1: Quantitative comparison of different methods on the VITON dataset. ↓ denotes the lower the better. For the user study results "a /b", a is the percentage where the compared method is considered better over our GT-MUST, and b is the percentage where our GT-MUST is considered better.

as 5, 120, and 0.05 empirically by referring to [22, 27]. The λ_n and λ_s have the same setting as [38].

Testing. In the testing process, we follow the same scheme as the training process but change the target garment to different clothes images in the test set. That is, the C and M_c in the three modules are changed accordingly.

4.3 Qualitative Results

Figure 3 shows the visual comparison on VITON dataset between the proposed GT-MUST and state-of-the-art virtual try-on methods, including CP-VTON [36], CP-VTON+ [29], ACGPN [38] and PF-AFN [9].

We can find that CP-VTON can generate a plausible shape for the target garment, but it tends to generate blurry textures. Moreover, the CP-VTON cannot retain the content of pants, since the body shape used in CP-VTON can only keep the shape of the pants without constraining the pants-related information explicitly. For example, in the row of CP-VTON's results in Figure 3, CP-VTON generates completely different pants from the reference person. As for CP-VTON+, it improves the ability to preserve pants while also suffering from obvious color and structure artifacts, e.g., the last row of CP-VTON+'s results. ACGPN can preserve more clothes characteristics than CP-VTON and CP-VTON+, e.g., the colors of ACGPN's results are much more natural than CP-VTON and CP-VTON+. Nevertheless, ACGPN is limited in generating mannequin-specific shapes. For example, the results of ACGPN in the last row always rely on the shape of the garment in the reference person rather than the mannequin. Especially for the right group, CP-VTON, CP-VTON+, and ACGPN generate unrealistic T-shirts on arms. These results demonstrate the weakness of only using shape transformation.

PF-AFN employs appearance flows to function the content transformation. As shown in Figure 3, PF-AFN generates much clearer garment patterns and reasonable mannequin shapes than CP-VTON, CP-VTON+, and ACGPN. Nevertheless, PF-AFN sometimes mismatches the clothes with the mannequin's arms and shape, e.g., the second row in the right group and the last row in the left group. As for the proposed GT-MUST, it can generate clear textures with fewer distortions and reasonable shapes that match the reference mannequin. Specifically, when the posture of the reference person is complex, e.g., the second row in the left group, GT-MUST can dress the red sweatshirt on the mannequin correctly by matching her posture, while the other methods fail to address this challenging case. These results confirm the superiority of the mannequin-specific transformation used in our GT-MUST.

Methods	Architecture	Size (M)	Speed (ms)
CP-VTON	U*1 + E*2 + TPS*1	40.41	0.0213
CP-VTON+	U*1 + E*2 + TPS*1	40.41	0.0122
ACGPN	U*4 + D*4 + STN/TPS*1	162.41	0.0363
PF-AFN	RU*2 + E*4 + AFEN*2	146.46	0.0812
GT-MUST	U*1 + STN/TPS*2	38.48	0.0099

Table 2: Model complexity analysis. U: U-Net. RU: Res-UNet. D: discriminator. E: feature extraction. AFEN denotes the progressive appearance flow estimation network devised in PF-AFN [9].

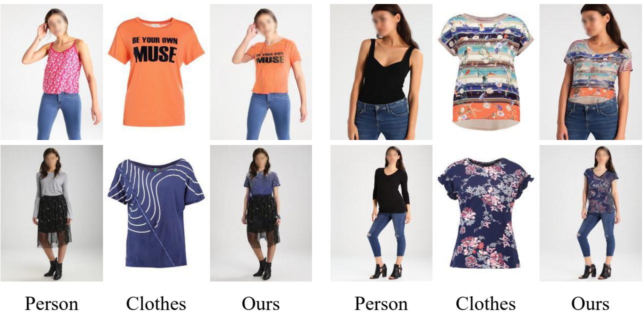

Figure 4: Results on the MPV dataset. The first row is the half front cases and the second row is the whole front cases.

Figure 4 presents the results on the MPV dataset. Given the target clothes, the proposed GT-MUST can not only generate realistic try-on results for half front views (top row) but also produce reasonable results for whole front views (bottom row). The shapes and patterns of target garments are well maintained on the mannequins. However, the proportion of the garment area in the whole front view images becomes smaller compared with that in the half front view images, resulting in errors in small areas like the necklines that may not receive sufficient guidance. For example, the neckline of the right case in the bottom row tends to mimic the shape of the mannequin's neckline rather than the shape of the target garment.

4.4 Quantitative Results

For quantitative evaluation, we adopt FID [16] and LPIPS [44] as metrics. LPIPS and FID are perceptual metrics, which calculate the distances of features between two images to measure the high-level similarity, i.e., perceptual similarity. Lower scores of FID and LPIPS indicate higher quality of the results. Since the most important goal of the VTON task is to generate visually reasonable and pleasing results, we do not employ the PSNR and SSIM as evaluation metrics, which typically cannot reflect the image-level perceptual quality. The Inception Score (IS) [33] is also not used since applying the IS to the models trained on datasets other than ImageNet will give misleading results [32].

Table 1 presents the quantitative results of CP-VTON [13], CP-VTON+ [29], ACGPN [38], PF-AFN [9], and the proposed GT-MUST on the VITON dataset. Compared with the other methods, the proposed GT-MUST achieves the best performance on both FID and LPIPS metrics. For example, it surpasses CP-VTON, CP-VTON+, ACGPN, and PF-AFN by 16.12, 8.00, 5.80, and 1.76 in terms of FID. It is also noteworthy that the quantitative evaluation results are generally consistent with the visual results in Figure 3. We also

Experiment Setting	FID ↓	LPIPS ↓
ILM	**72.78**	**0.360**
ILM w/o $\mathcal{L}_{const}$	113.16	0.361
MWM	**63.31**	**0.264**
MWM w/o T	69.29	0.272
MWM w/o $\hat{C}_m$	66.72	0.268
MWM w/ $\mathcal{L}_{const}$	72.18	0.295
GTM	**8.31**	**0.061**
GTM w/o T	8.95	0.063
GTM w/o $\mathcal{L}_{gate}$	9.22	0.071

Table 3: Ablation study of GT-MUST. Lower FID and LPIPS are better. Note that the evaluation of ILM is conducted by comparing the output $\hat{I}_c$ of ILM with C, while the evaluation of MWM is conducted by comparing the output $\hat{C}$ of MWM with I_c. The GTM results are measured by comparing the output $I^\dagger$ of GTM with I.

carry out a user study to evaluate the visual quality of the generated images of different methods, which will be discussed in Section 4.6.

4.5 Ablation Study

Efficiency analysis. Table 2 summarizes the network architectures and model sizes of four representative methods as well as the proposed GT-MUST. As can be seen, the models proposed in early works, e.g., CP-VTON and CP-VTON+, are much smaller than those in recent works, e.g., ACGPN and PF-AFN. CP-VTON and CP-VTON+ share the same architectures, which have only 40.41M of parameters. However, CP-VTON and CP-VTON+ generally perform worse than ACGPN and PF-AFN, which may take advantage of the larger model capacity. As for our GT-MUST, it is the smallest one among all these methods, which only has 38.48M of parameters, i.e., only 24% of ACGPN and 26% of PF-AFN. Nevertheless, it outperforms all the methods both quantitatively and qualitatively as shown in Figure 3 and Table 1, including the much larger ACGPN and PF-AFN models. The superior performance of GT-MUST is attributed to the efficient design of the proposed model, i.e., the ILM for learning effective mannequin-specific transformation and the GTM for generating visually consistent try-on images.

We also provide a comparison of the inference speed in Table 2, which refers to the average time required for processing one mannequin with one target garment. And all speeds are recorded with a V100 GPU. We can observe that, the proposed GT-MUST has the fastest inference speed on virtual try-on task, about 7 times faster than PF-AFN and 2.6 times faster than ACGPN. Our model is simple yet yields excellent results and is, in general, a relatively practical work to realize real-time virtual try-on.

Analysis of the mannequin-specific transformation. To demonstrate the effectiveness of the "take-off" strategy, we compare the performance of two modules, i.e.MWM and GTM, w/ and w/o the input of mannequin-specific transformation T predicted by ILM. As shown in Table 3, the MWM has better performance than the "MWM w/o T" by 5.98 in terms of FID, showing that the mannequin-specific transformation T is helpful to the generation of more realistic clothes. And the GTM also outperforms the corresponding ablation "GTM w/o T" by 0.65 in terms of FID. Please notice that, the FID results of MWM and GTM are quite different since the evaluations of MWM and GTM are based on clothes and mannequins respectively.

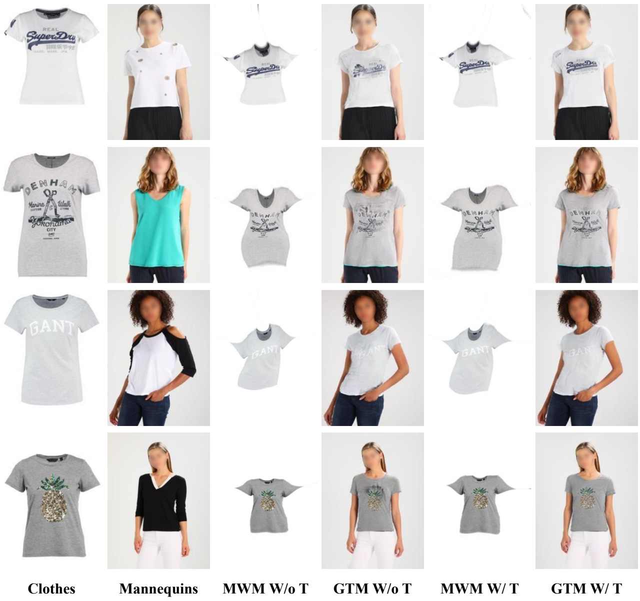

Figure 5: Qualitative comparison of the modules outputs in terms of w/ and w/o T.

Without specific knowledge about the target deformation, the previous works tend to predict a transformation T_s lying on the center of the learned transformation distribution, i.e., all the transformation-related details will be smoothed, leading to inaccurate transformation of the garment. This shape-based transformation T_s cannot provide right guidance for textures and patterns on garments. Therefore, the mannequin-specific transformation T is necessary and important as it provides additional knowledge to get the target per-pixel tailored transformation T'. According to the qualitative comparison results in Figure 5, we can find that T is good at yielding mannequin-fit garments warping. Both the qualitative and quantitative results confirm that the mannequin-specific transformation T helps to achieve better results than shape transformation.

Analysis of the gate mechanism. The gate G is utilized to dynamically choose regions of the warped clothes $\hat{C}$ that should be preserved. Typically, the black regions in G denote that the corresponding content of generated try-on $\tilde{I}$ will be copied from the warped clothes $\hat{C}$, while the white regions in G mean that the content of $\tilde{I}$ is generated by the U-Net in GTM. As shown in Figure 6, we can observe that the main part of clothes regions are black and part of the sleeves and body regions are white, which are reasonable for generating satisfactory try-on images by matching the target garment with the mannequin's posture, i.e., borrowing important information, like logos and embroideries, directly from $\hat{C}$ while keeping other body parts. The ablation of GTM w/o $\mathcal{L}_{gate}$ also demonstrates the importance of the gate mechanism as shown in Table 3.

We also investigate the influence of different values of λ_{gate} in Eq. (13), which is ranged in $\{0, 0.1, 0.2, ..., 1.1, 1.2\}$. Figure 7 plots the FID scores of the generated results by GT-MUST with different settings of λ_{gate}. As can be seen, with the increase of λ_{gate}, the

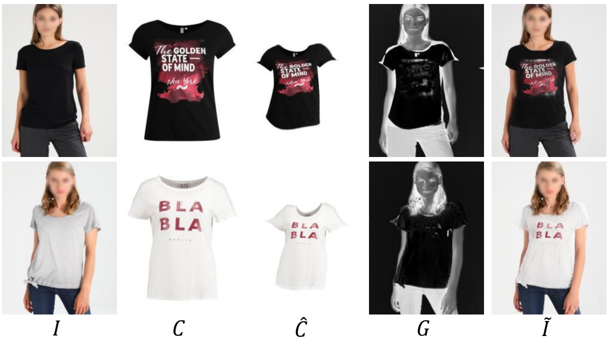

Figure 6: Illustration of gated try-on mechanism. From left to right are: (a) the reference person image I, (b) the target garment C, (c) the output $\hat{C}$ of MWM , (d) the generated gate G in GTM, and (e) the final try-on result $\tilde{I}$ in GTM.

performance first becomes better, reaches comparable low FID scores when λ_{gate} lies in [0.3,0.8], and then becomes worse. GT-MUST obtains the best FID when $\lambda_{gate} = 0.4$. It is noteworthy that when we remove $\mathcal{L}_{gate}$ from Eq. (13) (i.e., $\lambda_{gate} = 0$), GT-MUST suffers a significant performance loss, demonstrating the effectiveness of the proposed gate mechanism in generating high-quality try-on images.

Analysis of the matching clothes $\hat{C}_m$. Since STN/TPS in MWM only applies geometric warping to clothes, they cannot simulate the perfect shape of warped clothes, especially for the sleeves of clothes. Although a clothing mask loss is applied to constrain the shape, it cannot handle complex cases, such as crossed arms. To address this issue, we leverage the matching clothes $\hat{C}_m$ in Eq. (11) to constrain the output of STN/TPS at the correct position. Comparing the results of "MWM w/o $\hat{C}_m$" and "MWM" in Table 3, we can find that the matching clothes $\hat{C}_m$ facilitates the generation of more reasonable warped clothes in MWM.

Analysis of $\mathcal{L}_{const}$. As shown in Table 3, the geometric matching constraint $\mathcal{L}_{const}$ is useful in ILM but not suitable for GWM. It is because that $\mathcal{L}_{const}$ tends to restrict the warping degree of clothes and those in-shop garments usually have a flat state and similar shape while the warped clothes are in diverse postures. Therefore, $\mathcal{L}_{const}$ is only used in the ILM.

4.6 User Study

In Section 4.4, we employ FID and LPIPS as the objective metrics for the evaluation. Although they can measure the visual quality of generated images, they cannot comprehensively reflect whether the target clothes are naturally warped or whether the detailed texture of the garment is accurately preserved, which are also very important in the VTON task. To mitigate the issue, following the common practice in VTON works [9, 38], we further conduct a user study by recruiting 50 volunteers to compare 200 randomly selected image pairs by different methods in an A/B test manner. These test pairs are all assigned a different garment for each mannequin following CP-VTON+ [29]. For each image pair, the target garment and reference person images are also attached. Each volunteer is asked to choose a better image that is more photo-realistic as well as better in preserving garment content and matching the reference mannequin. The results are listed in the last column of Table 1, which again demonstrate the superiority of the proposed GT-MUST over the state-of-the-art methods.

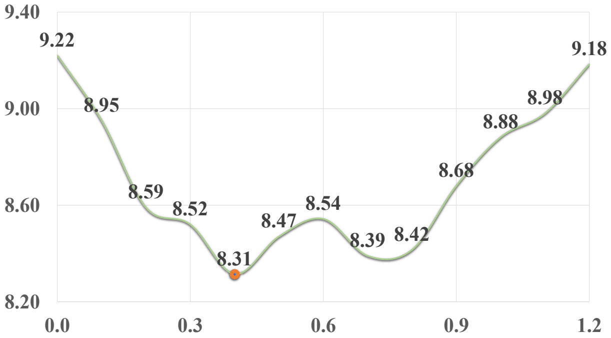

Figure 7: The FID for different sittings of λ_{gate}. The x-axis represents the different values of λ_{gate}, while the y-axis represents the corresponding score of FID.

4.7 Limitations and Discussion

There are two limitations of GT-MUST that could be addressed in the future study. First, GT-MUST is composed of three modules, which, as in previous works [9, 36], are not trained in an end-to-end manner, since the inaccurate predictions from the previous module may affect the subsequent one. How to develop an efficient end-to-end solution remains under-explored and deserves further research efforts. Second, each module of GT-MUST receives a bulk of images in different modalities as input and learns a mapping function implicitly from them, which lacks explainability about the role of each modality. Although the module-wise design can provide explicit intermediate outputs that can be used for troubleshooting, it deserves further study to devise explainable architectures for each module.

5 CONCLUSION

In this paper, we introduce a novel method named GT-MUST for the virtual try-on (VTON) task. It completes this task by learning mannequin-specific content transformation in the new "first take-off and then put-on" pipeline, showing its superiority over the widely used shape transformation in generating fine-detailed warped clothes and photo-realistic try-on results. Furthermore, we introduce a gate mechanism to handle the style differences between clothes, which can effectively match the target garments with different shapes to the reference mannequin in diverse postures. Experiments on the public dataset demonstrate that the proposed GT-MUST outperforms the state-of-the-art methods both quantitatively and qualitatively.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62076188, the Special Fund of Hubei Luojia Laboratory under Grant 220100014, and the Fundamental Research Funds for the Central Universities under Grant 2042021kf0196. The computations in this work had been supported by the supercomputing system in the Supercomputing Center of Wuhan University. Dr. Jing Zhang was supported by the ARC project FL-170100117.

REFERENCES

[1] Fred L. Bookstein. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI* 11, 6 (1989), 567–585.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. CVPR*. 7291–7299.

[3] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. 2021. ZFlow: Gated Appearance Flow-based Virtual Try-on with 3D Priors. In *Proc. ICCV*. 5433–5442.

[4] Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2021. Template-Free Try-on Image Synthesis via Semantic-guided Optimization. *IEEE TNNLS* (2021), 1–14.

[5] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. 2021. Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing. In *Proc. ICCV*. 3940–3945.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*. 248–255.

[7] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. 2019. Towards multi-pose guided virtual try-on network. In *Proc. ICCV*. 9026–9035.

[8] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K. Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. 2022. Dressing in the Wild by Watching Dance Videos. In *Proc. CVPR*. 3480–3489.

[9] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. 2021. Parser-Free Virtual Try-on via Distilling Appearance Flows. In *Proc. CVPR*. 8485–8493.

[10] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proc. CVPR*. 932–940.

[11] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. 2019. Clothflow: A flow-based model for clothed person generation. In *Proc. ICCV*. 10471–10480.

[12] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. 2019. Finet: Compatible and diverse fashion image inpainting. In *Proc. ICCV*. 4481–4491.

[13] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. Viton: An image-based virtual try-on network. In *Proc. CVPR*. 7543–7552.

[14] Haoyu He, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2020. Grapy-ML: Graph pyramid mutual learning for cross-dataset human parsing. In *Proc. AAAI*, Vol. 34. 10949–10956.

[15] Sen He, Yi-Zhe Song, and Tao Xiang. 2022. Style-Based Global Appearance Flow for Virtual Try-On. In *Proc. CVPR*. 3470–3479.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. NeurIPS*. 6629–6640.

[17] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2019. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proc. ACM MM*. 275–283.

[18] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. 2020. Do Not Mask What You Do Not Need to Mask: A Parser-Free Virtual Try-On. In *Proc. ECCV*. 619–635.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. In *Proc. NeurIPS*. 2017–2025.

[20] Hyug Jae Lee, Rokkyu Lee, Minseok Kang, Myounghoon Cho, and Gunhan Park. 2019. LA-VITON: a network for looking-attractive virtual try-on. In *Proc. ICCV Workshops*.

[21] Nikolay Jetchev and Urs Bergmann. 2017. The conditional analogy gan: Swapping fashion articles on people images. In *Proc. ICCV Workshops*. 2287–2292.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*. 694–711.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. ICLR*.

[24] Shizuma Kubo, Yusuke Iwasawa, Masahiro Suzuki, and Yutaka Matsuo. 2019. UVTON: UV mapping to consider the 3D structure of a human in image-based virtual try-on network. In *Proc. ICCV Workshops*. 3105–3108.

[25] Huyen TK Le, Andre L Carrel, and Harsh Shah. 2021. Impacts of online shopping on travel demand: a systematic review. *Transport Reviews* (2021), 1–23.

[26] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. 2021. Tryongan: Body-aware try-on via layered interpolation. *ACM TOG* 40, 4 (2021), 1–10.

[27] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. 2020. Recurrent feature reasoning for image inpainting. In *Proc. CVPR*. 7760–7768.

[28] Matiur Rahman Minar and Heejune Ahn. 2020. CloTH-VTON: Clothing Three-dimensional reconstruction for Hybrid image-based Virtual Try-ON. In *Proc. ACCV*. 154–172.

[29] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. 2020. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*.

[30] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. 2020. Image based virtual try-on network from unpaired data. In *Proc. CVPR*. 5184–5193.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*. 234–241.

[32] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. 2017. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987* (2017).

[33] Tim Salimans, Ian J Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Proc. NeurIPS*. 2226–2234.

[34] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust scene text recognition with automatic rectification. In *Proc. CVPR*. 4168–4176.

[35] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. ICLR*.

[36] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proc. ECCV*. 589–604.

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv preprint arXiv:2204.12484* (2022).

[38] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proc. CVPR*. 7850–7859.

[39] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. 2019. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proc. ICCV*. 10511–10520.

[40] Jing Zhang, Zhe Chen, and Dacheng Tao. 2021. Towards high performance human keypoint detection. *IJCV* 129, 9 (2021), 2639–2662.

[41] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. 2021. PISE: Person Image Synthesis and Editing with Decoupled GAN. In *Proc. CVPR*. 7982–7990.

[42] Jing Zhang and Dacheng Tao. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IoTJ* 8, 10 (2020), 7789–7817.

[43] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful Image Colorization. In *Proc. ECCV*. 649–666.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*. 586–595.

[45] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. 2021. M3D-VTON: A Monocular-to-3D Virtual Try-On Network. In *Proc. ICCV*. 13239–13249.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*. 2223–2232.