ZFlow: Gated Appearance Flow-based Virtual Try-on with 3D Priors

Ayush Chopra*†2, Rishabh Jain *‡3, Mayur Hemani[1], and Balaji Krishnamurthy[1]

[1]Media and Data Science Research Lab, Adobe
[2]Media Lab, Massachusetts Institute of Technology
[3]BITS Pilani

Abstract

Image-based virtual try-on involves synthesising perceptually convincing images of a model wearing a particular garment and has garnered significant research interest due to its immense practical applicability. Recent methods involve a two stage process: i) warping of the garment to align with the model ii) texture fusion of the warped garment and target model to generate the try-on output. Issues arise due to the non-rigid nature of garments and the lack of geometric information about the model or the garment. It often results in improper rendering of granular details. We propose ZFlow, an end-to-end framework, which seeks to alleviate these concerns regarding geometric and textural integrity (such as pose, depth-ordering, skin and neckline reproduction) through a combination of gated aggregation of hierarchical flow estimates termed Gated Appearance Flow, and dense structural priors at various stage of the network. ZFlow achieves state-of-the-art results as observed qualitatively, and on quantitative benchmarks of image quality (PSNR, SSIM, and FID). The paper presents extensive comparisons with other existing solutions including a detailed user study and ablation studies to gauge the effect of each of our contributions on multiple datasets.

1. Introduction

With recent socio-cultural events accelerating the shift towards online commerce, there is an increasing interest in providing smart and intuitive experiences [19, 27, 3, 1, 6, 22] that can compensate for the lack of in-store interaction. Virtual try-on is concerned with the visualization of clothes in a personalized setting and is of great importance to a plethora of real world applications. While attractive even before the renaissance of deep learning [37, 15, 8], recent advances in generative networks have inspired researchers to pursue image-based virtual try-on [41, 19, 40, 13, 43], based solely on RGB images, by formulating the problem as that of conditional image synthesis.

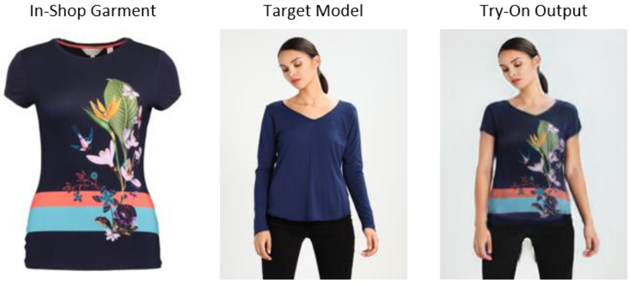

Figure 1. Image-based virtual try-on involves synthesizing a *try-on output* where the *target model* is wearing the *in-shop garment* while other characteristics of the model and garment are preserved. The above output is generated by our proposed method ZFlow

Given as input the images of an *isolated in-shop garment* and a *target model*, the objective of image-based virtual try-on is to synthesise a perceptually convincing new image (referred to as the *try-on output*) where the target model is wearing the in-shop garment (Figure 1). Recent methods employ a two step process consisting of: a) *warping* of in-shop garment to align with pose and body shape of the target model and, b) *texture fusion* of the warped garment and target model images to generate the try-on output. A successful try-on experience depends upon synthesizing a sharp, realistic image that preserves the textural and geometric integrity of both the garment and model. Issues arise from improper warping or incorrect texture fusion due to the non-rigid nature of garments and the lack of understanding of the 3D geometry of the garment and the model. This results in unconvincing rendering of granular clothing details. Alleviating these concerns is the focus of this work.

Recent research [14, 40, 19, 41] has been directed to-

*equal contribution
†work done while working at Adobe MDSR Lab
‡work done as part of Adobe MDSR internship

wards these challenges. [14, 40] proposed thin-plate spline (TPS) based warping of the garment image. [19, 41] improve the stability of TPS warping via multi-stage cascaded parameter estimation, and second order difference constraints respectively. However, TPS based warping leads to inaccurate transformation estimation when large geometric deformation is required, since each parameter defines the spatial deformation for a coarse block of pixels. To address this issue, [13] proposes to use dense, per-pixel appearance flow [45] prediction to spatially deform the garment image. But owing to the high degree of freedom and the absence of proper regularization, this method often causes drastic deformation during warping resulting in significant textural artefacts. To address both issues - the inability of TPS to handle large deformations, and over-warping with appearance flows - we introduce *Gated Appearance Flow* (GAF) which regularizes per-pixel appearance flow by aggregating candidate flow estimates predicted across multiple scales.

Next, for improving texture fusion, especially the issue of bleeding colors, [19, 41] propose to use an *apriori* estimate of target clothing segmentation for the try-on output as conditioning. However, this method results in ambiguities in depth perception and body-part ordering because of the absence of 3D geometric priors. This is prominently visible in the generation of necklines, and handling cases with occlusion. For example, part of the garment that should go behind the neck appears in the front. To encode the 3D geometry information, we combine UV projection maps with dense body-part segmentation (obtained via DensePose [12]) as priors during warping and texture fusion.

Our contributions can be summarized as follows:

- We propose ZFlow, an end-to-end try-on framework, that utilizes gated appearance flow estimates and dense geometric priors to render high quality try-on results.
- We present extensive quantitative and qualitative comparisons as well as a detailed user study to show significant improvement over state-of-the-art methods.
- We present ablation studies to analyse impact of different design choices in ZFlow. We further reinforce the efficacy of *GAF* by adapting it to improve state-of-the-art for human pose transfer.

2. Related Work

Virtual Try-On Progress in deep learning has motivated 2D image-based try-on as a scalable alternative to older methods ([32, 28, 37, 46]) that used 3D scanners for virtual fitting of clothing items. Most of these new 2D image-based methods [14, 40, 41, 19, 13] pose the problem as that of synthesizing a realistic image of a model from a reference image and an isolated garment image. VITON [14] uses a Thin-Plate Spline (TPS) based warping method to deform the garment images and maps the warped garment onto the model image using an encoder-decoder refinement module. CP-VTON [40] improves over [14] using a neural network to regress the transformation parameters of TPS. SieveNet [19] improves over [40, 14] by estimating TPS parameters over multiple interconnected stages and also proposes a conditional layout constraint to better handle pose variation, bleeding and occlusion during texture fusion. ACGPN [41] utilizes a similar layout constraint and also imposes a second-order constraint on TPS warping to preserve local patterns. However, these methods can only model limited geometric changes and often unnaturally deform clothing due to limited degrees of freedom in TPS transformation. ClothFlow [13] uses a per-pixel appearance flow [45] (instead of TPS) predicted over multiple cascaded stages, and also utilizes the conditional layout constraint as in [19, 41]. Appearance flow [45] is used to spatially deform a source scene to a target scene by computing a pixel-wise 2D deformation field. This is conceptually distinct from optical flow and we refer the reader to [30] for a discussion regarding the difference. The high degree of freedom in per-pixel flow estimation as well as the limited (3D) structural information often results in geometric misalignment and unnatural and bleeding textures. We propose ZFlow, an end-to-end framework which seeks to preserve the geometric and textural integrity by a combination of gated aggregation of hierarchical flow estimates across pixel-block levels (Gated Appearance Flow) and dense structural priors (Dense Geometric Priors) at various stages of the network.

3D Pose Representation The optimal choice of 3D representations for neural networks is an open problem. Recent works in single image 3D reconstruction have explored voxel, point cloud, octree, surface and volumetric representations [38, 25, 42, 2, 44, 20, 18]. Surface based representation methods [2, 42] use UV maps [10] to establish dense correspondence between pixels and human body surface. To preserve the geometric integrity (depth-ordering, pose, skin and neckline reconstruction) of the try-on output in our image-based setup, we use dense geometric priors in form of UV maps and body-part segmentation masks obtained from a pre-trained DensePose [12]. These priors helps in handling complex poses even under heavy occlusion.

Human Pose Transfer Given a reference image of a person and a target pose, the task is to synthesize an image of the model in the desired pose. [26] uses a two stage, guided image-to-image translation network to generate the target. Recent work [35, 7, 4, 13, 23, 11] incorporates spatial deformation from the source to the target for better perceptual quality. ClothFlow [13] predicts a dense appearance flow over multiple interconnected stages using a stacked network to warp source clothing pixels. Dense Intrinsic Flow (DIF) [23], introduced a flow regression module to map input and target skeleton poses with 3D appearance flow which it then uses to performs feature warping on the input image and generate a photo-realistic target image. We validate the ef-

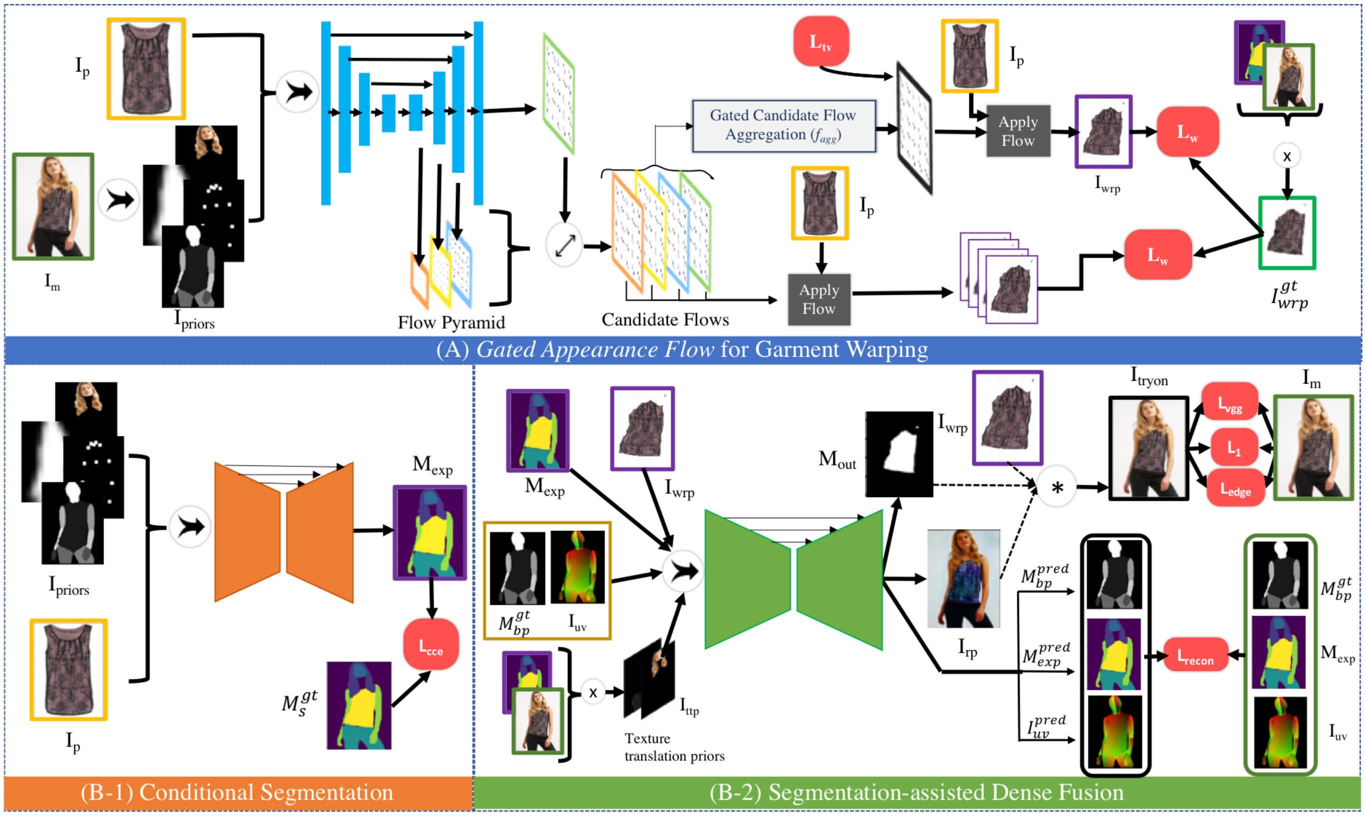

Figure 2. ZFlow comprises of two modules: A) Garment Warping to deform the garment I_p to align with model I_m and generates warped garment (I_{wrp}), and B) Texture Fusion which has two sub-steps - i. Conditional Segmentation to predict a post try-on clothing segmentation of the model M_{exp} ii. Dense Fusion which combines the warped garment (I_{wrp}) and segmentation mask (M_{exp}) to generate the final output (I_{tryon}). *Gated Appearance Flow* for garment warping improves textural integrity of I_{tryon} by regularizing the per-pixel flow estimation. Dense geometric priors I_{priors} improves geometric integrity of the try-on output.

ficacy of *Gated Appearance Flow* by adapting it for regression of 3D flows in [23]. We note subsequent work in pose transfer [29] but highlight that [23] is ideal for our objective of validating the efficacy of *GAF*.

3. Methodology

ZFlow takes as input images of a target model (I_m) and an isolated garment product (I_p) to generates the *try-on* output I_{tryon}, where the target model is wearing the garment. This transformation is composed of two key phases: (*A*) **Garment Warping** which deforms I_p to align with pose of the model in I_m and generates I_{wrp}, (*B*) **Texture Fusion** which composes the warped garment I_{wrp} with I_m to generate I_{tryon} over two steps: (*B-1*) conditional segmentation and (*B-2*) segmentation-assisted fusion (as in Figure 2).

3.1. Garment Warping

I_p is warped based on pose and shape of the target model I_m to produce a warped garment image I_{wrp}. For this, we propose *Gated Appearance Flow* which estimates per-pixel warp parameters by aggregating candidate estimates predicted across multiple scales (pixel-block sizes).

3.1.1 Enriched Input

Because training triplets where the same model wears two different garments are unavailable, contemporary methods use as input a clothing-agnostic prior of the target model (I_m) along with the garment I_p. We extend the conventional binary (1-channel) body shape, (18-channel) pose map and (3-channel) head region used previously [41, 19, 13] with an additional dense (11-channel) body-part segmentation (M_{bp}^{gt}) of I_m to provide richer structural priors (I_{priors}). This subtle enhancement, as we delineate in section 6, cascades through the network and results in significantly fewer artefacts in the output.

3.1.2 Gated Appearance Flow

This module predicts per-pixel appearance flow (pixel displacements) for warping the garment image by aggregating candidate flow estimates across multiple scales. The process comprises of first predicting the flow estimates and then aggregating them using a gating mechanism, along with losses that ensure smoothness (and regularity) of the flow predictions.

Multi-scale Appearance Flow Prediction The backbone network is a 12-layer Skip-Unet [31]. Given an input RGB image of size (H, W), the last K decoding layers are used to predict the candidate flow maps (f_l for $l \in \{0, ..., K\}$) such that a predicted map f_l is double the size of map f_{l-1}. All maps are then interpolated to have identical height and width (H, W) generating a pyramid of K *candidate* flow maps that correspond to a structural hierarchy.

Appearance Flow Aggregation The candidate flows are combined to obtain an aggregate per-pixel appearance flow (f_{agg}), using a convolution gated recurrent-network (ConvGRU) [34] (summarized in figure 2(A)). Intuitively, this is a per-pixel selection process that determine the aggregate flow by gating (allowing or dismissing) pixel flow estimates corresponding to different radial neighborhoods (for the multiple scales). This prevents over-warping of the garment image by regularizing the high degrees of freedom in dense per-pixel appearance flow. We corroborate this position with extensive ablation studies in section 6.1 where we propose and contrast several alternative flow aggregation mechanisms.

Garment Image Warping Next, the aggregate appearance flow map f_{agg} is used to warp the garment image I_p and mask M_p to obtain the warped image I_{wrp} and the warped binary garment mask M_{wrp} respectively. Additionally, the intermediate flow maps f_l for $l \epsilon \{0, .., K\}$ are also used to produce intermediate warped images and masks (I^l_{wrp}, M^l_{wrp}).

Losses Each of the warped images (final and intermediate) are subject to L1-loss L_1 and perceptual similarity loss L_{vgg} [36] with respect to garment regions of the model image. Each predicted warped mask is subject to a reconstruction loss with respect to M_m^{gt}. The predicted flow-maps are subjected to a total variation loss ($\beta_4 L_{tv}(f_l)$) to ensure spatial smoothness of flow-predictions. The combined warping loss is defined as L_{wrp} :

$$\begin{aligned} L_{wrp} &= L_w(I_{wrp}, M_{wrp}, f_{agg}) \\ &+ \sum_{l=0}^{l=K} L_w(I^l_{wrp}, M^l_{wrp}, f_l) \end{aligned} \tag{1}$$

for,

$$\begin{aligned} L_w(I, M, f) = \beta_1 \| I \odot M, I_m \odot M_m^{gt} \|_1 \\ + \beta_2 L_{vgg}(I \odot M, I_m \odot M_m^{gt}) \\ + \beta_3 \| M, M_m^{gt} \|_1 + \beta_4 L_{tv}(f) \end{aligned} \tag{2}$$

Validation with Human Pose Transfer For an extended validation of GAF's efficacy for estimating appearance flows, we use it to regress 3D flows for human pose transfer. The task involves producing an image of a person in a target pose from a reference image. We note that in contrast to virtual try-on where *GAF* is used for warping the *garment* based on the model pose, here it warps the target model pose itself. DIF [23] is a recent method for pose transfer that first regresses on a 3D appearance flow to map input to target pose and then performs feature warping on the input using the flow estimates. We swap-in our proposed GAF for 3D flow regression while retaining the feature warping module of DIF. We observe significant qualitative improvement in the generated image and discuss the results in section 6.

3.2. Texture Fusion

Once the warped garment (I_{wrp}) is obtained, the final try-on output is then generated over two steps (figure 2 B-1 and B-2): First, a conditional mask M_{exp} is predicted that corresponds to the clothing segmentation of the target model *after* garment change in try-on. Then, M_{exp} is combined with the warped garment (I_{wrp}) and the texture and geometry priors to produce the try-on output (I_{tryon}) .

3.2.1 Conditional Segmentation

The inputs to this module are the garment image (I_p) and the Dense Garment-Agnostic Representation (I_{priors}). The I_{priors} encodes the geometry of the target person and is agnostic to the specific garment the model is wearing. This is important to prevent over-fitting as the pipeline is trained on paired data where the input and output are the same images (and hence have the same segmentation mask). The network architecture is a Skip-UNet [31] with six encoder and decoder layers and the output, M_{exp}, is the 7-channel clothing segmentation mask.

Losses The module is trained with a weighted cross-entropy loss with respect to the ground-truth garment segmentation mask (M_s^{gt}) obtained with a pre-trained human parser (as used in [19, 41, 13]). The weight for skin and background classes are increased (3.0 in our experiments) for better handling of bleeding, and self-occlusion where the pose of the person results in certain parts of the garment or body to remain hidden from view. The loss is expressed as:

$$L_{cs} = -\frac{1}{n} \sum_{n} \sum_{i=0}^{6} w_i P_i^{gt} log(P_i^{pred}) \tag{3}$$
$$\text{where } w_i = [3, 1, 1, 1, 3, 1, 1] \text{ for } i \epsilon [0, 6]$$

We observe that using the Dense Garment-Agnostic Representation improves depth perception and handling of occlusion in M_{exp} which results in try-on outputs with fewer artefacts. We discuss this further in section 6.2.

3.2.2 Segmentation-Assisted Dense Fusion

This stage generates the final try-on output. The network architecture for this stage is also a Skip-UNet [31] with six encoder and decoder layers. The network inputs include outputs of the previous stages (I_{wrp} and M_{exp}) and texture translation prior ($I_{ttp} = I_m * M_{exp}$) representing the non-garment pixels of I_m. To include the 3D geometry of the model, we also input a dense prior (called *IUV Priors*) composed of UV map (I_{uv}) and body-part segmentation (M_{bp}^{gt}) of the target model. We note that M_{bp}^{gt} (*body-part* segmentation) is a function of the body geometry (agnostic of the specific garments) and differs from M_{exp} (or M_s^{gt}) (*clothing* segmentation) which is altered with changing garments (both are useful for try-on). The try-on output (I_{tryon}) is defined as:

$$I_{tryon} = M_{out} * I_{wrp} + (1 - M_{out}) * I_{rp} \tag{4}$$

where M_{out} and I_{rp} are generated by the network. M_{out} is a composite mask for the garment pixels in try-on output and I_{rp} is a *rendered person* comprising all target model pixels *except* the garment in the try-on output. To preserve structural and geometric integrity of the try-on output, we also constrain the network to reconstruct the input clothing segmentation (as M_{exp}^{pred}) and IUV (as $M_{bp}^{pred}, I_{uv}^{pred}$) priors which are unchanged during this step.

Losses I_{tryon} is subject to L_1, perceptual similarity [36] (L_{vgg}) and edge (L_{edge}) losses with respect to the model image I_m. L_{edge} is based on sobel filters (∇_x and ∇_y) and improves quality of the reproduced textures. Finally, M_{exp}^{pred}, M_{bp}^{pred} and I_{uv}^{pred} are subjected to reconstruction losses against their corresponding network inputs (M_{exp}, M_{bp}^{gt} and I_{uv} respectively). This reconstruction loss (L_{recon}) combines cross entropy (L_{cce}) for the categorical masks (M_{exp}^{pred}, M_{bp}^{pred}) and smooth L_1 for the I_{uv}^{pred} map.

$$\begin{aligned} L_{fus} = \lambda_1 * \|I_{tryon} - I_m\|_1 + \lambda_2 * L_{vgg}(I_{tryon}, I_m) \\ + \lambda_3 * L_{edge}(I_{tryon}, I_m) + \lambda_4 * L_{recon} \end{aligned} \tag{5}$$

where,

$$\begin{aligned} L_{recon} = L_{cce}(M_{exp}^{pred}, M_{exp}) + L_{cce}(M_{bp}^{pred}, M_{bp}^{gt}) \\ + \|I_{uv}^{pred} - I_{uv}\|_{smoothL1} \end{aligned} \tag{6}$$

We observe that conditioning texture fusion with these geometric priors via L_{recon} improves quality of try-on output via improved depth perception and structural coherence and explain this effect with evidence in section 6.

3.3. Training

Following a brief warm-up period of τ steps for the warping and texture fusion modules where they are trained separately, we optimize ZFlow end-to-end with the following loss function:

$$L_{total} = \alpha_1 * L_{wrp} + \alpha_2 * L_{cs} + \alpha_3 * L_{fus} \tag{7}$$

where $\alpha_1, \alpha_2, \alpha_3$ are scalar hyperparameters.

4. Experiments

In this section, we formalise the setup for our experiments with virtual try-on and human pose transfer.

Datasets For image-based virtual try-on, we use the VITON dataset [14] to ensure consistence with baseline methods. It contains 19000 images of front-facing female models and corresponding upper-clothing isolated garment images of size 256x192. There are 16253 cleaned pairs, which are split into train and test sets of 14221 and 2032 pairs. We also separate out 500 pairs from the train set into a validation set used exclusively for quantitative analysis. The images in the test set are rearranged into unpaired sets for qualitative evaluation. For human pose transfer, we use in-shop clothes benchmark from the Deep Fashion dataset [24] which contains 52712 in-shop clothes images and 200000 cross-pose pairs of size 256x256. Following the setup in DIF [23], we select 89262 pairs and 12000 pairs for train and test respectively.

Implementation Details All experiments are conducted using Pytorch on Tesla V100 GPUs. For virtual try-on, all modules are trained for 30 epochs with a batch size of 4 and learning rate of 1e-4 using Adam [21]. We set $K = 3$ for *Gated Appearance Flow* and the warm-up period τ is 5 epochs. For human pose transfer, we train the flow regression module for 40 epochs with learning rate=1e-4 using Adam [21] and retain the configuration in [23] for the feature warping module. Additional hyperparameter details are in the appendix.

Evaluation Metrics For virtual try-on, we use SSIM [33], FID [16] and PSNR [17] of the warp garment and try-on output. We avoid inception score (IS) following the considerations presented in [5]. For human pose transfer, we evaluate performance using SSIM [33] and PSNR [17] to ensure consistency with baselines. We note these metrics are chosen to ensure consistent comparison with prior work.

Baselines For virtual try-on, we compare performance with several recent state-of-the-art methods including CP-VTON [40], SieveNet [19], ClothFlow [13], VTNFP [43] and ACGPN [41]. For [40, 19, 41], we use author provided implementations and perform extensive qualitative and quantitative comparisons.

5. Results

We present quantitative (in Table 1) and qualitative results (Figure 3) along with a user study which highlight the superiority of ZFlow over strong baselines.

Method	SSIM ↑	PSNR ↑	FID ↓
VTNFP [43]†	0.803	-	-
ACGPN [41]†	0.845	-	-
CP-VTON [40]	0.784	21.01	30.50
SieveNet [19]	0.837	23.52	26.67
ClothFlow [13]	0.843	23.60	23.68
ZFlow	**0.885**	**25.46**	**15.17**

Table 1. ZFlow achieves significant improvement over existing baselines. † results may be inferred as indicative as they are transferred from corresponding papers.

Quantitative Results Table 1 compares performance of ZFlow against state-of-the-art baselines for virtual try-on. We report performance for TPS-based baselines [40, 19] using author provided implementations. In comparison to [40, 19], ZFlow achieves significantly better SSIM of 0.885, PSNR of 25.46 and FID of 15.17, compared to the next best values (SSIM=0.845, PSNR=23.60 and FID=23.68). We note that ZFlow with *GAF* significantly outperforms *ClothFlow* [13] which uses vanilla per-pixel appearance flow based warping for the garment image. Note that the official code for ClothFlow [13] was not available, we implement it as described and reproduce stated SSIM values.

Qualitative Results Figure 3 illustrates qualitative comparison with SieveNet [19], CP-VTON [40] and ACGPN [41], the baselines with available code implementations. We contrast the try-on outputs along varying dimensions of quality. These include factors that determine the realism of the generated image as a whole as well as the local geometry, colors and patterns.

Rows (1-5) demonstrate improvement in *geometric integrity* - the accurate representation of the geometry of the target model, the garment, and their interaction in the try-on output. Specifically, we observe that ZFlow improves the handling of *extreme pose* (row 1), *depth-ordering* of body parts, especially hands and neck region (row 2), *skin generation* for correct visibility of target garment and human skin (row 3) and *neckline reproduction & shoulder correction* in coherence with garments structure (row 4, 5). We highlight the improved neckline reproduction and depth-ordering in row 5 where none of the baselines are able to disambiguate front and back of the garment neckline.

Rows (6-10) demonstrate improvement in *texture integrity* which is concerned with accurate reproduction of patterns and colors of inshop garments in try-on output, and the handling of related artefacts. Specifically, we observe that ZFlow improves the reproduction of *pattern and texture* (stripes in row 6, 7), *print design* of the garment (graphic in row 8), *text* written on garment (row 9) and prevents color bleeding across part boundaries (row 10).

Shadows and highlights in the generated image, espe-

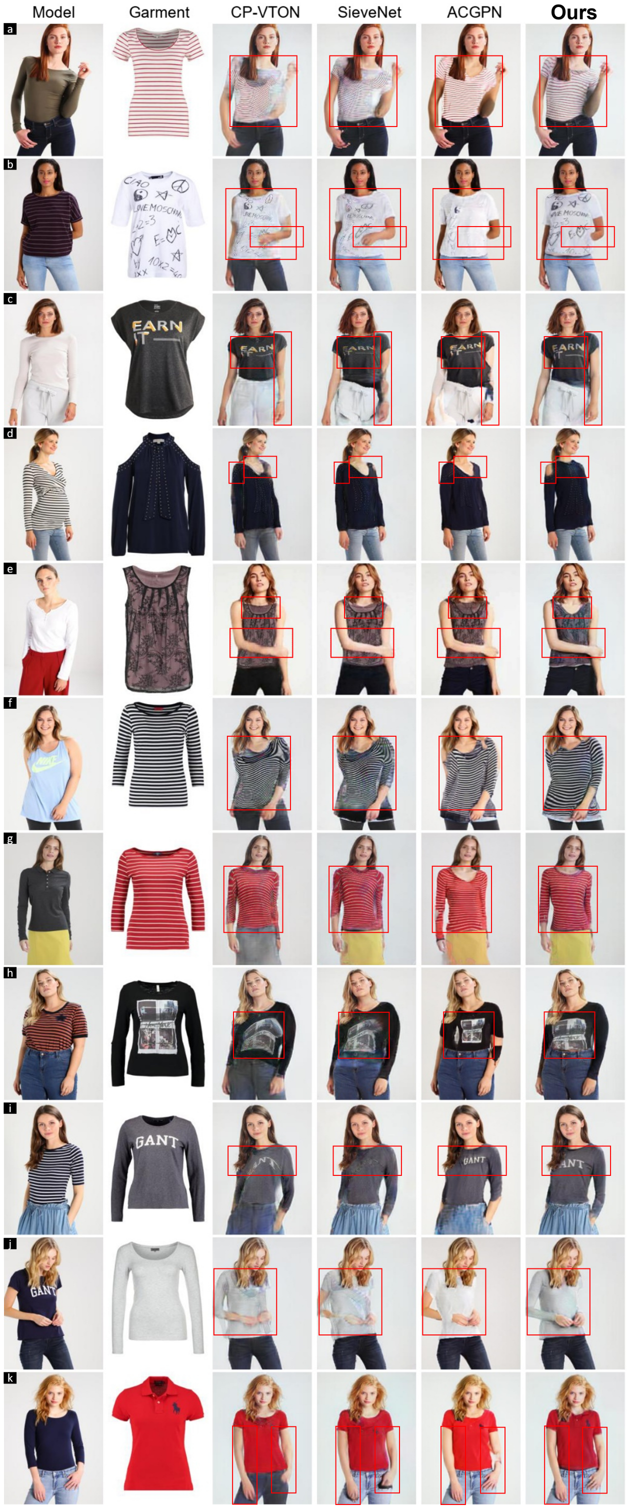

Figure 3. Qualitative Comparison of ZFlow with [40, 19, 41]. Rows **1-5** reflect improvements in preserving geometric integrity, and Rows **6-10**, texture integrity. Please note: (a) Complex poses (b) Depth ordering of body parts (c) Skin generation (d,e) Neckline and shoulder correction (f,g) Pattern (h) Texture, (i) Text, (j) Reduced bleeding across part boundaries. Row **11** (k) Realistic outline shadows for crisper image quality. (Best viewed with zoom). Please see appendix (pages 3-5) for more results.

Baseline	Prefer Baseline	Prefer ZFlow
CP-VTON [40]	8%	**92%**
SieveNet [19]	15%	**85%**
ACGPN [41]	29%	**71%**

Table 2. Survey results for gauging the human preference of ZFlow over competing baselines. The percentage indicates the ratio of images which are voted to be better than the compared method.

cially along the boundaries of body parts, are also important to correctly represent the dynamics of the actual scene. Row 11 demonstrates improvement along this dimension.

User Study We conduct a survey with 70 volunteers from 3 continents, 5 countries, 10 institutions across diverse age, gender and occupations. As in [41], we use *pairwise comparison* where each user is shown 100 distinct result *pairs* randomly sampled from 2000 test set results. Each pair consists of one ZFlow result and the other sampled from the results of (one of three) baselines ([41, 19, 40]). The in-shop garment and target model images are also shown for each result pair. Every volunteer is asked to select the best output of the two in each result pair in unlimited time. Results in Table 2 show overwhelmingly clear preference for ZFlow in ***all*** pairwise comparisons.

6. Ablation Studies

In this section, we analyse the impact of different contributions of ZFlow and summarize results in Table 3.

6.1. Gated Appearance Flow (GAF)

First, we demonstrate the impact of GAF for garment image warping by comparing it to an existing per-pixel appearance flow based warping technique proposed in ClothFlow [13]. Next, to justify our choice of using a ConvGRU layer for aggregating hierarchical candidate appearance flow estimates, we propose alternate flow-aggregation schemes and report comparison with ConvGRU.

ClothFlow and GAF Rows 1 and 2 of Table 3 compare the use of per-pixel appearance flow for garment image warping as described in [13] with the proposed gated aggregation of hierarchical flow estimates (GAF). GAF clearly outperforms the vanilla warping method corroborating our position that gated aggregation yields superior results both for the warping stage as well as for the try-on output.

Design choices for GAF In rows 3, 4 and 5 of Table 3, we compare following schemes for gated aggregation - *i) Using Residual Gating* to perform residual sum (operation from [39]) on flow estimates of the last two decoding layers. *ii) Using ConvLSTM* for the flow estimate aggregation over three layers (3 scales), and *iii) Using ConvGRU* for aggregating flow estimates. The results indicate clearly that using *ConvGRU* for gated aggregation produces the *best results of the three* and *hence is used in GAF*.

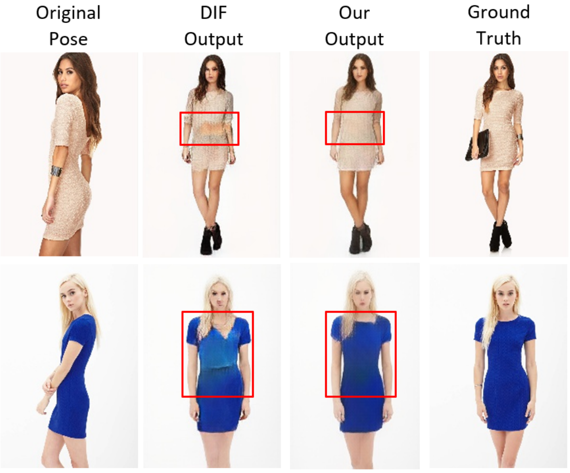

Figure 4. Using *GAF* for flow regression in pose transfer improves skin generation (row 1) and reduces bleeding (row 2).

Further, we note that all three aggregation schemes significantly outperform *ClothFlow* on metrics for both the warped garment and try-on output. For instance, *ConvGRU* improves the warp garment SSIM (from 0.835 to 0.871) and PSNR (from 20.54 to 23.14) against ClothFlow [13]. We note that this benefit translates to the try-on output where we observe consistent gains in SSIM (from 0.843 to 0.865), PSNR (from 23.60 to 24.47) and FID (from 23.48 to 18.89).

GAF in Human Pose Transfer As an additional test of the efficacy of the proposed appearance flow-aggregation, we adapt it for flow-regression for task of Human Pose Transfer, building upon baseline DIF [23]. This results in both qualitative (Figure 4) and quantitative (Table 4) improvements in the pose-transfer output. Figure 4 present evidence to show significantly improved skin generation (row 1), texture (row 2) and reduced bleeding (row 1, 2) in the generated image. We corroborate this with results in Table 4 which indicates considerable improvement in SSIM (from 0.778 and 0.791) and PNSR (from 18.59 to 19.26). We also note the significant gain over ClothFlow [13], which also uses flow regression, as a validation of the efficacy of GAF.

6.2. Input Priors, Losses and Training

Dense Garment-Agnostic Representation (I_{priors}) is proposed as structural priors for garment warping and conditional segmentation. Figure 5 shows that this improves depth perception, skin generation (row 1) and neckline reconstruction (row 2) in the try-on output. We note similar improvements during garment warping (qualitative in appendix) which is corroborated through increase in PSNR of the warp garment (row 5 vs 6 in Table 3).

Configuration		Warp Garment (I_{wrp})		Try-On Output (I_{tryon})		
Garment Warping	*Texture Fusion*	*SSIM* ↑	*PSNR* ↑	*SSIM* ↑	*PSNR* ↑	*FID* ↓
ClothFlow [13]	BaseFuse	0.835	20.54	0.843	23.60	23.68
GAF	BaseFuse	**0.871**	**23.14**	**0.865**	**24.47**	**18.89**
Various Gating Approaches for Flow Aggregation						
Residual Gating	BaseFuse	0.856	22.09	0.855	24.11	21.64
LSTM	BaseFuse	0.862	22.56	0.860	24.33	18.89
ConvGRU (GAF)	BaseFuse	0.871	23.14	0.865	24.47	18.89
Loss Functions						
GAF (w/ I_{priors})	BaseFuse + L_{edge}	0.871	23.28	0.875	25.02	19.39
GAF (w/ I_{priors})	BaseFuse + L_{edge} + L_{recon}	0.871	23.28	0.876	25.12	18.74
ZFlow (end-to-end training)		**0.871**	**23.28**	**0.885**	**25.46**	**15.17**

Table 3. Ablation studies for various design choices for garment warping and texture fusion in ZFlow. *BaseFuse* is the texture fusion network trained without L_{edge} and L_{recon}.

.

Method	SSIM ↑	PSNR ↑
DSC [35]	0.756	-
PG2 [26]	0.762	-
ClothFlow [13]	0.771	-
VUnet [9]	0.786	-
DIF [23]	0.778	18.59
Ours	**0.791**	**19.26**

Table 4. Using GAF for flow regression improves the quality of generated image in human pose transfer

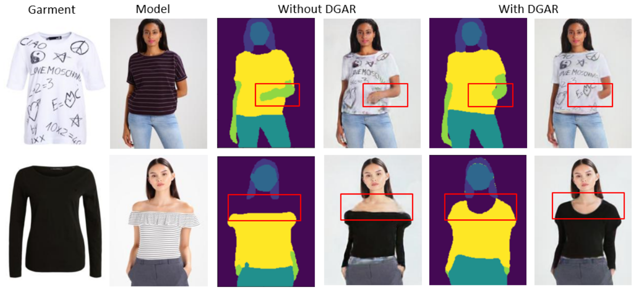

Figure 5. Using the dense garment-agnostic representation (DGAR) for conditional segmentation improves depth-perception (row 1), skin and neckline generation (row 2).

IUV Priors composed of UV projection map (I_{uv}) and body-part segmentation (M_{bp}^{gt}) are used to encode the 3D geometry of the target model during texture fusion. The $ZFlow$ network is trained to reconstruct these priors along with the try-on output (I_{tryon}). Figure 6 shows that conditioning on these IUV Priors via the reconstruction loss (L_{recon}) improves generation of neckline, skin (row 1) and depth perception (row 2) in the output. This is corroborated through improved PSNR (25.02 to 25.12) and FID (19.39 to 18.74) of the try-on output (row 6 vs 7 in Table 3).

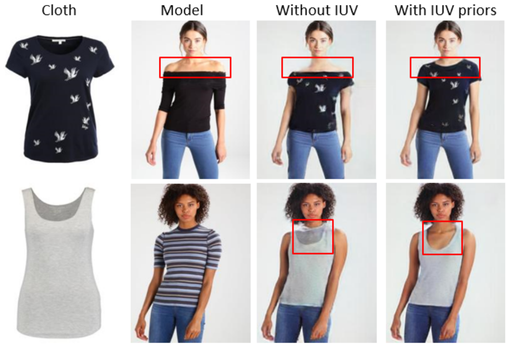

Figure 6. IUV priors during texture fusion improves the neckline (row 1), depth perception and skin generation (row 2)

Edge Loss (L_{edge}) based on Sobel filters is used to better preserve high frequency details during texture fusion. Table 3 show that this improves SSIM (from 0.865 to 0.875) and PSNR (from 24.47 to 25.02) of try-on output.

End-to-end Fine Tuning The end-to-end fine-tuning of the entire ZFlow network (including the warping and texture fusion modules) improves SSIM (0.876 to 0.885), PSNR (25.12 to 25.46) and FID (from 18.74 to 15.17) of the try-on output as indicated in Table 3 (row 7 vs 8).

7. Conclusion

We introduce ZFlow, an end-to-end try-on framework, which utilizes a combination of gated aggregation of hierarchical flow estimates (Gated Appearance Flow) and dense geometric priors (DGAR and IUV Priors) to reduce undesirable output artefacts. We highlight effectiveness of ZFlow through comparisons with state-of-the-art and detailed ablation studies. We also validate the efficacy of GAF as a general technique by applying it to human pose transfer.

References

[1] Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[3] Kumar Ayush. Context aware recommendations embedded in augmented viewpoint to retarget consumers in v-commerce. 1

[4] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 2

[5] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 5

[6] Ayush Chopra, Abhishek Sinha, Hiresh Gupta, Mausoom Sarkar, Kumar Ayush, and Balaji Krishnamurthy. Powering robust fashion retrieval with information rich feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[7] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems*, pages 474–484, 2018. 2

[8] Jun Ehara and Hideo Saito. Texture overlay for virtual clothing based on pca of silhouettes. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 139–142. IEEE, 2006. 1

[9] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 8

[10] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 2

[11] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided image generation. *arXiv preprint arXiv:1811.11459*, 2018. 2

[12] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[13] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[14] 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*, pages 7543–7552, 2018. 1, 2, 5

[15] Stefan Hauswiesner, Matthias Straka, and Gerhard Reitmayr. Virtual try-on through image-based rendering. *IEEE transactions on visualization and computer graphics*, 19(9):1552–1565, 2013. 1

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 5

[17] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

[18] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[19] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. Sievenet: A unified framework for robust image-based virtual try-on. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2182–2190, 2020. 1, 2, 3, 4, 5, 6, 7

[20] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] Yining Lang, Yuan He, Fan Yang, Jianfeng Dong, and Hui Xue. Which is plagiarism: Fashion image retrieval based on regional representation for design protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[23] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 2, 3, 4, 5, 7, 8

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[25] 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)*, 34(6):1–16, 2015. 2

[26] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017. 2, 8

[27] Kushagra Mahajan, Tarasha Khurana, Ayush Chopra, Isha Gupta, Chetan Arora, and Atul Rai. Pose aware fine-grained visual classification using pose experts. pages 2381–2385, 10 2018. 1

[28] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 2

[29] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[30] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 4, 5

[32] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014. 2

[33] Kalpana Seshadrinathan and Alan C Bovik. Unifying analysis of full reference image quality assessment. In *2008 15th IEEE International Conference on Image Processing*, pages 1200–1203. IEEE, 2008. 5

[34] Mennatullah Siam, Sepehr Valipour, Martin Jagersand, and Nilanjan Ray. Convolutional gated recurrent networks for video segmentation, 2016. 4

[35] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 2, 8

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4, 5

[37] Hiroshi Tanaka and Hideo Saito. Texture overlay onto flexible object with pca of silhouettes and k-means method for search into database. In *MVA*, pages 5–8, 2009. 1, 2

[38] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 2

[39] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019. 7

[40] 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)*, pages 589–604, 2018. 1, 2, 5, 6, 7

[41] 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*, pages 7850–7859, 2020. 1, 2, 3, 4, 5, 6, 7

[42] Pengfei Yao, Zheng Fang, Fan Wu, Yao Feng, and Jiwei Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*, 2019. 2

[43] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10511–10520, 2019. 1, 5, 6

[44] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019. 2

[45] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. *CoRR*, abs/1605.03557, 2016. 2

[46] Zhenglong Zhou, Bo Shu, Shaojie Zhuo, Xiaoming Deng, Ping Tan, and Stephen Lin. Image-based clothes animation for virtual fitting. In *SIGGRAPH Asia 2012 Technical Briefs*, pages 1–4. 2012. 2