MeshNet: Learning Hierarchical Flow Aggregation for Image-based Virtual Try-on

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

[1]Media and Data Science Research Lab, Adobe Experience Cloud
[2]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 sparse priors of the 3D model geometry which results in unnatural rendering of granular details. We propose MeshNet, 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 (GatedFlow) and dense structural priors at various stage of the network. MeshNet achieves state-of-the-art results as observed qualitatively, and on benchmark image quality measures (PSNR, SSIM, and FID scores). The paper presents detailed comparisons with other existing solutions, as well as ablation studies to gauge the effect of each of our contributions. Finally, we also present evidence of the generalized efficacy of GatedFlow through its effect on human pose transfer task.

1. Introduction

With recent socio-cultural events accelerating the shift towards contactless online commerce, there is an increasing interest in providing smart and intuitive experiences [20, 29, 3, 1, 6, 23] 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 [38, 16, 9], recent advances in generative networks have inspired researchers to pursue image-based virtual try-on [42, 20, 41, 14, 44], based solely on RGB images, by formulating the problem as that of conditional image synthesis.

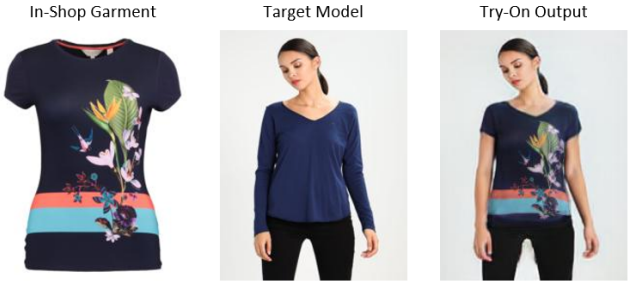

Figure 1. Image-based virtual try-on involves synthesizing a *try-on output* where the *target model* is wearing the *in-shop garment* while other characteristics of the model and garment are preserved. The above output is generated by our proposed method MeshNet

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. Successful try-on experiences depend upon synthesizing sharp, realistic outputs that preserve the textural and geometric integrity of both the garment and model. Yet, issues may arise from improper warping or incorrect texture fusion due to the non-

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

rigid nature of garments and sparse priors on the 3D model geometry, which results in unnatural rendering of granular clothing details. Alleviating these concerns is the focus of this work.

Several recent research efforts have been directed towards these challenges. [15, 41] proposed thin plate spline (TPS) based warping of in-shop garment where [15] uses image descriptors and [41] uses neural networks to learn the transformation parameters. [20, 42] seek to improve the stability of TPS warping via multi-stage cascaded parameter estimation in [20] and second order difference constraints in [42] as well as utilize an *apriori* estimate of the try-on output's target clothing segmentation to condition texture fusion. To enhance warping flexibility and accuracy with higher degrees of freedom, [14] predicts a dense per-pixel appearance flow (instead of TPS) to spatially deform the in-shop garment and also utilizes the conditional clothing segmentation prior when generating the try-on output. On one hand, the limited degrees of freedom in TPS (2x5x5 as in [41]) leads to inaccurate transformation estimations when large geometric change occur since each parameter defines the spatial deformation for large pixel-blocks. On the other hand, while dense appearance flow allows pixel-to-pixel matching between in-shop and warped garments, the high degrees of freedom often presents unappealing textural artefacts resulting from drastic deformation in the absence of proper regularization. Furthermore, while 2D segmentation estimates improve spatial coherence and bleeding through priors on the 2D contour, limited understanding of the 3D model geometry results in ambiguities in depth perception and body-part ordering. This results in unnatural try-on outputs with improper generation of necklines and handling of occlusion (part of the in-shop garment that should go behind the neck appears in front).

In this work, we seek to alleviate these issues with geometric and textural integrity through a combination of gated aggregation of hierarchical flow estimates and by providing dense structural priors at various stages. *First*, we introduce *GatedFlow* which regularizes per-pixel appearance flow by aggregating candidate flow estimates predicted across multiple granularities (pixel-block sizes). *Second*, to convey 3D geometry information of the body, we use a combination of UV projection maps and dense body-part segmentation (obtained via DensePose [13]) as constraints during warping and texture fusion. We aggregate these contributions to propose MeshNet, an end-to-end virtual try-on framework, to render try-on results with fewer artefacts. Finally, we also validate the generalizability of the proposed *GatedFlow* on the task of human pose transfer.

Our contributions can be summarized as:

- We propose MeshNet, an end-to-end try-on framework that utilize a combination of gated aggregation of hierarchical flow (*GatedFlow*) and dense geometric priors to render try-on results with fewer artefacts.
- We present extensive comparisons with existing state-of-the-art, along varying dimensions of output quality, and show significant improvement.
- We present detailed ablation studies (with quantitative and qualitative evidence) to analyse impact of different components and design choices in MeshNet.
- We validate the generalized efficacy of *GatedFlow* by: i) proposing and comparing multiple flow aggregation mechanisms for virtual try-on and, ii) adapting it to improve state-of-the-art methods for human pose transfer.

2. Related Work

Virtual Try-On While initial methods used 3D scanners for virtual fitting of clothing items [33, 30, 38, 46], recent progress in deep learning has motivated 2D image-based try-on as a scalable alternative. Such 2D methods can be categorized based on whether they preserve posture or not. [8] proposed multi-pose guided image-based virtual try-on. Analogous to MeshNet, most existing methods preserve pose and identity during try-on [15, 41, 42, 20, 14]. These methods use coarse human shape and pose maps as priors for generating the clothed images. VITON [15] uses a Thin-Plate Spline (TPS) based warping method to deform the in-shop garments and maps the warped garment onto the model image using an encoder-decoder refinement module. CP-VTON [41] adopts a similar framework but uses a neural network to regress the transformation parameters of TPS. SieveNet [20] improves over [41, 15] 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 [42] 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 [14] uses a per-pixel appearance flow (instead of TPS) predicted over multiple cascaded stages, and also utilizes the conditional layout constraint as in [20, 42]. 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 MeshNet, an end-to-end framework which seeks to preserve the geometric and textural integrity by a combination of gated aggregation of hierarchical (across pixel-block levels) flow estimates and dense structural priors at various stages of the network.

3D Human Representation While the optimal choice of 3D representations for neural networks remains an open problem, recent works in single image 3D reconstruction have explored voxel, point cloud, octree, surface and volumetric representations [39, 26, 43, 2, 45]. SMPL[26] pa-

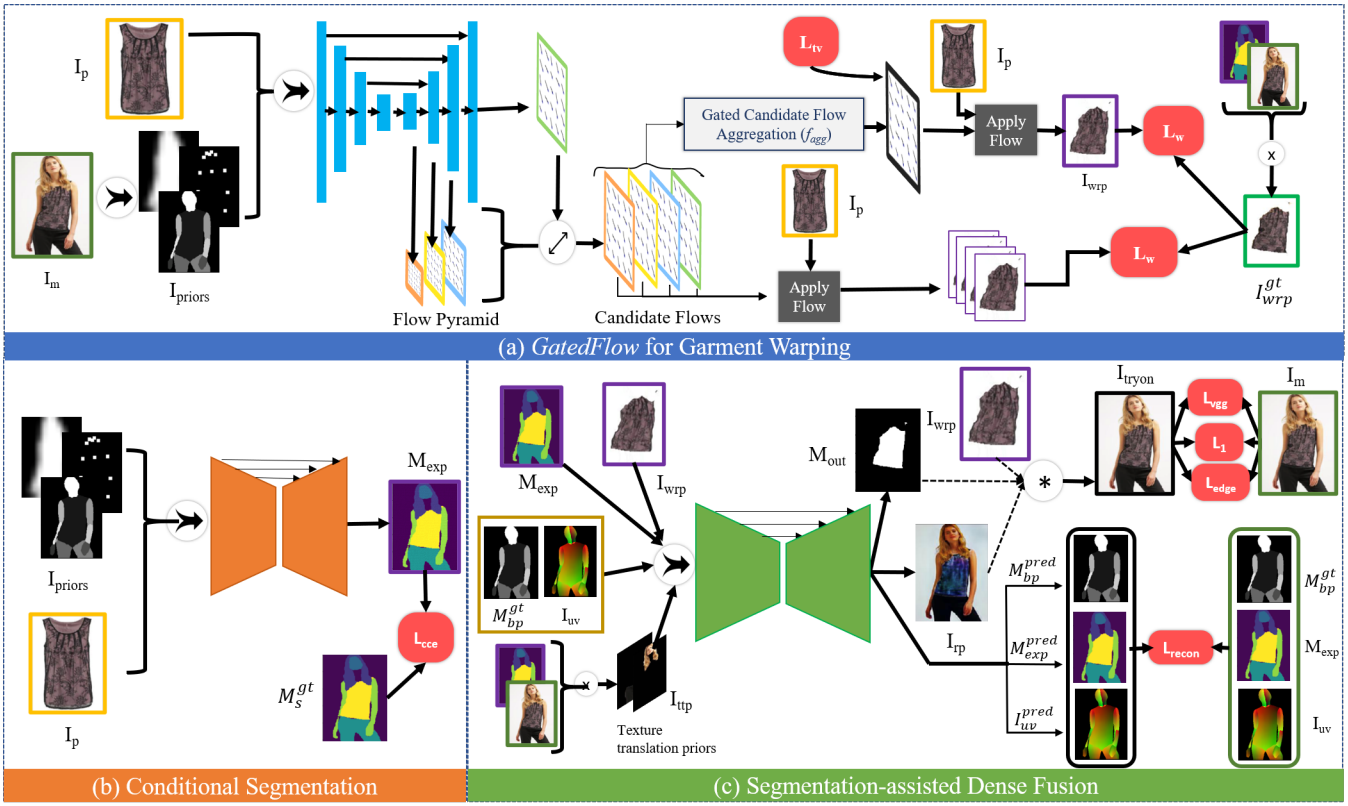

Figure 2. MeshNet comprises of three key modules: i) Garment Warping which deforms the garment I_p to align with model I_m and generates warped garment (I_{wrp}) ii) Conditional Segmentation which hallucinates the post try-on clothing segmentation of the model M_{exp} iii) Dense Fusion which combines the warped garment (I_{wrp}) and segmentation mask (M_{exp}) to generate the final try-on output (I_{tryon}). We introduce *GatedFlow* for garment warping to improve textural integrity of I_{tryon} by regularizing the high degrees of freedom in per-pixel flow estimation. We incorporate dense body shape constraints during warping and fusion to improve geometric integrity of the try-on output. Detailed analysis of MeshNet and each of these components is presented in Sections 5 and 6.

rameterizes the 3D mesh of a human body by 3D joint angles and a low dimensional linear shape space [21]. While volumetric representations utilize voxels [39, 19] for the reconstruction process, surface based representations [2, 43] use UV maps [11] to establish dense correspondence between pixels and human body surface and have achieved remarkable progress in recent times. In an attempt to preserve geometric integrity of the try-on output (specifically depth-ordering, pose, skin and neckline reconstruction) in our image-based setup, we incorporate dense geometric priors in the form of UV maps and body-part segmentation (obtained via pre-trained [13]) to enable virtual try-on with fewer artefacts. These design choices are motivated by the ability of [13] to handle complex poses even under heavy occlusion in a computationally inexpensive manner.

Human Pose Transfer Given the reference image of a model and a target pose, the task is to synthesize a new photo-realistic image of the model in the desired pose. [27] used a two stage image-to-image translation network to generate the target image with guidance of the source image and target pose. Recent works [36, 7, 4, 14, 24, 12] incorporate spatial deformation between the source and target to improve perceptual quality of the results by deforming pixels (or feature maps) of source to align with target. ClothFlow [14] predicted a dense appearance flow over multiple interconnected stages using a stacked network to warp source clothing pixels. Dense Intrinsic Flow (DIF) [24], a recent state-of-the-art, 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 efficacy of *GatedFlow* by adapting it for regression of 3D flows in [24]. We note subsequent work in human pose transfer [31] but highlight that [24] is highly competitive and ideal for our objective of validating the generalized efficacy of *GatedFlow*.

3. Methodology

As depicted in figure 2, MeshNet 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} with the target model wearing the garment. This transformation is composed of two key phases: i) **Garment Warping** which deforms the image I_p to align with pose of the model I_m and generates I_{wrp}, ii) **Texture Fusion** which composes the warped garment I_{wrp} with I_m to generate I_{tryon} over two steps: conditional segmentation and segmentation-assisted fusion.

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}. We propose *GatedFlow* which estimates per-pixel warp parameters by aggregating candidate flow estimates predicted across multiple granularities (pixel-blocks sizes).

Enriched Input Due to unavailability of ideal training triplets (where the same model wears two different garments) as discussed in [15], contemporary methods provide 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 and (18-channel) pose map used previously [42, 20, 14] with an additional dense (11-channel) body-part segmentation (M^{gt}_{bp}) of I_m to provide richer structural priors (I_{priors}). This subtle enhancement, as we delineate in section 6, cascades through the network and has significant impact on rendering try-on outputs with fewer artefacts.

GatedFlow predicts per-pixel appearance flow by aggregating candidate flow estimates across multiple granularities (pixel-blocks). The backbone network is a 12-layer Skip-Unet [32] where each of the six encoding (and decoding) layers sequentially downsample (and upsample) their incoming feature map by a factor of K(=2). Given an input RGB image of size (H, W), the last M decoding layers are used to predict appearance flow maps (f_l for $l \in \{0, 1, ..., M\}$) such that the l-th layer flow estimate corresponds to a K^{2l} size pixel-block. In this notation, $l = 0$ for last layer which predicts a per-pixel flow estimate. Each of the predicted flow maps are then interpolated (where K^{2l} pixels share flow estimates in l-th layer) to have identical height and width (H, W) generating a pyramid of M *candidate* appearance flow maps which conform to a coarse structural hierarchy. We aggregate candidate flows, to obtain a composite per-pixel appearance flow (f_{agg}), using a convolution gated recurrent-network (convGRU) [35]. 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. This has the benefit of preventing 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 (section 6) where we propose and contrast several alternative aggregation mechanisms with the dense per-pixel appearance flow. We note that additional details of the *ConvGRU* for aggregating flows are included in appendix due to limited space.

The composite appearance flow map f_{agg} is used to sample the warped image I_{wrp} and the warped binary garment mask M_{wrp} from the garment image I_p and mask M_p respectively. Additionally, the network also produces warped images and masks (I^l_{wrp}, M^l_{wrp}) for each of the l candidate flow maps by also sampling over I_p and M_p.

Losses Since the network is trained on paired data, the warped garments are subject to L1-norm loss L_1 and perceptual similarity loss L_{vgg} [37] with respect to garment regions of the model image (obtained by $I_m \odot M^{gt}_m$) where M^{gt}_m is a binary garment map obtained from the ground-truth segmentation mask (M^{gt}_s) for I_m. A total-variance loss L_{tv} is also applied on the flow maps to ensure smoothness in flow prediction. The combined warping loss is defined as L_{wrp} :

$$L_{wrp} = L_w(I_{wrp}, M_{wrp}, f_{agg}) + \sum_{l=0}^{l=M} L_w(I^l_{wrp}, M^l_{wrp}, f_l) \quad (1)$$

for,

$$L_w(I, M, f) = \beta_1 \|I \odot M, I_m \odot M^{gt}_m\|_1 + \beta_2 L_{vgg}(I \odot M, I_m \odot M^{gt}_m) + \beta_3 \|M, M^{gt}_m\|_1 + \beta_4 L_{tv}(f) \quad (2)$$

with $\beta_1, \beta_2, \beta_3, \beta_4$ as scalar hyperparameters.

Validation with Human Pose Transfer To validate the generalized efficacy of *GatedFlow*, we additionally utilize it to regress 3D flows for human pose transfer. DIF [24] is a recent state-of-the-art in pose transfer which generates photo-realistic images in the target pose by i) first regressing 3D appearance flow which map input to target pose and ii) then performing feature warping on the input using the flow estimates. We define *DIF-GatedFlow*, a variant of DIF [24], where we swap-in our proposed GatedFlow for 3D flow regression while retaining the feature warping module. We observe significant improvement in the generated target pose image and delineate results in section 6. Due to limited space, network details are included in the appendix.

3.2. Texture Fusion

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

a) Conditional Segmentation The inputs to this module are the in-shop garment (I_p) and dense clothing-agnostic prior (I_{priors}). We use I_{priors} to represent the target model since it encodes the target model geometry but is agnostic to the specific garment the model is wearing. This is important to prevent overfitting 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 [32] with six encoder and decoder layers and the output, M_{exp}, is the 7-channel clothing segmentation mask.

Losses The module is optimized on a weighted cross-entropy objective with respect to a ground-truth clothing segmentation mask (M_s^{gt}) obtained with the same pre-trained human parser as [20, 42, 14] . The weight for skin and background classes are increased (set to 3.0) to better handle bleeding and cases of 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:

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

We observe that using the dense cloth-agnostic prior, in contrast to the coarse supervision in [20, 42, 14], improves depth perception and handling of occlusion in M_{exp} which results in try-on outputs with fewer artefacts. We delineate this with evidence in section 6.

b) Segmentation-Assisted Dense Fusion The goal of this stage is to generate the final try-on output. The network architecture is also a Skip-UNet [32] 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 convey 3D geometry of model, we also input a dense prior (called *IUV prior*) 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 [37] (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.

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

where,

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

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, we optimize MeshNet 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 dataset collected by Han et al. [15]. 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 [25] which contains 52712 in-shop clothes images and 200000 cross-pose pairs of size 256x256. Following the setup in DIF [24], 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 [22]. We set $M = 3$ for *GatedFlow* 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 [22] and retain the configuration in [24] for the feature warping module. Additional hyperparameter details are in the appendix.

Evaluation Metrics For virtual try-on, we use SSIM [34], FID [17] and PSNR [18] 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 [34] and PNSR [18] to ensure consistency with baselines.

Baselines For virtual try-on, we compare performance with several recent state-of-the-art methods including CP-VTON [41], SieveNet [20], ClothFlow [14], VTNFP [44] and ACGPN [42]. For [41, 20, 42], we use author provided implementations and perform extensive qualitative and quantitative comparisons. For human pose transfer, the primary goal is to validate efficacy of *GatedFlow* which can be delineated by comparisons with the original Dense Intrinsic Flow [24]. For completeness, we also compare with other recent methods such as [28, 10, 27, 36, 14].

5. Results

In this section, we present quantitative (in Table 1) and qualitative results (Figure 3) for virtual try-on.

Method	SSIM ↑	PSNR ↑	FID ↓
VTNFP [44]†	0.803	-	-
ClothFlow [14]∗†	0.833	-	-
ACGPN [42]†	0.845	-	-
CP-VTON [41]	0.784	21.01	30.50
SieveNet [20]	0.837	23.52	26.67
DenseFlow	0.843 (0.835)	23.60	23.68
MeshNet	**0.885**	**25.46**	**15.17**

Table 1. MeshNet achieves significant improvement over existing baselines. † results may be inferred as indicative as they are transferred from corresponding papers. ∗ indicates warp-SSIM. We note (in brackets) the warp-SSIM for *DenseFlow*

Quantitative Results Table 1 compares performance of MeshNet against state-of-the-art baselines for virtual try-on. We report performance for TPS based baselines [41, 20] using author provided implementations. In comparison to [41, 20], MeshNet achieves drastically better SSIM of 0.885 (vs 0.837 or 0.845), PNSR of 25.46 (vs 23.52) and FID of 15.17 (vs 26.67). A key component of MeshNet is the flow based warping module *GatedFlow*. Hence, to validate the efficacy of the hierarchical estimation and aggregation in *GatedFlow*, we implement and compare with *DenseFlow* which uses a vanilla per-pixel flow estimation for warping (row 6). *DenseFlow* resembles the 1-stage variant of *ClothFlow* [14], and obtains similar SSIM (0.835 vs 0.833), for which the implementation was not available. We note that MeshNet with *GatedFlow* significantly outperforms *DenseFlow* as depicted in the table (row 6 vs 7). For completeness, we include the reported metrics for [14, 42, 44] which, while indicative, can corroborate the efficacy of MeshNet owing to the considerable improvement.

Qualitative Results Figure 3 illustrates qualitative comparison with SieveNet [20], CP-VTON [41] and ACGPN [42], 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* which is concerned with the accurate representation of the geometry of the target model, in-shop garment (and their interaction) in the try-on output. Specifically, we observe that MeshNet 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 granular *neckline reproduction & shoulder correction* in-coherence with garments structure (row 4, 5). We particularly note the improved neckline reproduction and depth-ordering in row 5 where all baselines are unable 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 MeshNet 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).

Apart from local considerations of geometry and texture, realistic try-on outputs may also represent correctly the dynamics of actual scene. This is often realized by the correct presentation of shadows and highlights in the generated image, especially along the boundaries of body parts. Row 11 demonstrates improvement along this dimension.

6. Discussion

In this section, we analyse the impact of different contributions of MeshNet and summarize results in Table 2.

6.1. Gated Aggregation of Hierarchical Flow

We posit that regularizing per-pixel appearance flow by aggregating hierarchical candidate flow estimates ($GatedFlow$) is beneficial for image warping. To corroborate this position, we: a) propose and contrast different flow aggregation schemes on virtual try-on and, b) validate efficacy of *GatedFlow* for 3D flow regression on pose transfer.

Configuration		Warp Garment (I_{wrp})		Try-On Output (I_{tryon})		
Garment Warping	*Texture Fusion*	*SSIM* ↑	*PSNR* ↑	*SSIM* ↑	*PSNR* ↑	*FID* ↓
DenseFlow	BaseFuse	0.835	20.54	0.843	23.60	23.48
ResidualFlow	BaseFuse	0.856	22.09	0.855	24.11	21.64
LSTMFlow	BaseFuse	0.862	22.56	0.860	24.33	18.89
GRUFlow	BaseFuse	0.871	23.14	0.865	24.47	18.89
GRUFlow (w/ I_{priors})	BaseFuse + L_{edge}	0.871	23.28	0.875	25.02	19.39
GRUFlow (w/ I_{priors})	BaseFuse + L_{edge} + L_{recon}	0.871	23.28	0.876	25.12	18.74
MeshNet (end-to-end)		**0.871**	**23.28**	**0.885**	**25.46**	**15.17**

Table 2. Ablation studies for various design choices for garment warping and texture fusion in MeshNet. *BaseFuse* is the texture fusion network trained without L_{edge} and L_{recon}. ResidualFlow, GRUFlow and LSTMFlow are variants of *GatedFlow* as noted in sec 6.1

.

Ablating Hierarchical Flow Aggregation We contrast the following variants of *GatedFlow*: **i) ResidualFlow** which performs residual gating [40] (summarized in appendix) on flow estimates of the last two decoding layers, **ii) LSTMFlow** which uses a convLSTM to aggregate flow predictions across $M(=3)$ scales and **iv) GRUFlow** which uses a convGRU to aggregate flow predictions across the $M(=3)$ scales and corresponds to the *GatedFlow* in MeshNet. We contrast these methods with **iv) DenseFlow** which directly uses the vanilla per-pixel flow estimate at the last decoding layer. Results in Table 2 (rows 1-4) show that all the three aggregation schemes significantly outperform *DenseFlow* on metrics for both the warped garment and try-on output. For instance, *GRUFlow* improves the warp garment SSIM (from 0.835 to 0.871) and PNSR (from 20.54 to 23.14) against *DenseFlow*. We note that this benefit also 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). *GRUFlow* is the best variant and is used for *GatedFlow* in MeshNet. Due to limited space, we substantiate these metric with qualitative results in the appendix.

***GatedFlow* in Human Pose Transfer** We swap-in *GatedFlow* (specifically *GRUFlow*) for flow regression in DIF [24] and formalise this variant as *DIF-GatedFlow*. 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 3 which indicates considerable improvement in SSIM (from 0.778 and 0.791) and PNSR (from 18.59 to 19.26). We note the significant gain over ClothFlow [14], which also uses flow regression, as a validation of the efficacy of *GatedFlow*.

Method	SSIM ↑	PSNR ↑
DPIG [28]	0.614	-
DSC [36]	0.756	-
PG2 [27]	0.762	-
ClothFlow [14]	0.771	-
VUnet [10]	0.786	-
DIF [24]	0.778	18.59
Ours	**0.791**	**19.26**

Table 3. Using *GatedFlow* for flow regression improves the quality of generated image in human pose transfer

6.2. Input Priors and Training Design

Dense Garment Agnostic Prior (I_{priors}) is proposed to provide richer structural priors for garment warping and conditional segmentation. Results in 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 4 vs 5 in Table 2).

IUV Priors composed of UV projection map (I_{uv}) and body-part segmentation (M^{gt}_{bp}) are used to convey 3D geometry of the target model during texture fusion with the network trained to additionally reconstruct these priors along the try-on output (I_{tryon}). Results in 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 5 vs 6 in Table 2).

Edge Loss (L_{edge}) based on sobel filters is used to better preserve high frequency details during texture fusion. Results in Table 2 show that this improves SSIM(from 0.865 to 0.875) and PSNR (from 24.47 to 25.02) of try-on output. We substantiate this with visual results in appendix which highlight improvement in texture, contrast and handling of bleeding in the generated images.

MeshNet is **finetuned end-to-end** with the warping and texture fusion modules jointly optimized. Results in Table 2 (row 6 vs 7) shows that this significantly improve SSIM (from 0.875 to 0.885), PSNR (from 25.10 to 25.46) and FID (from 19.09 to 15.17) of the try-on output.

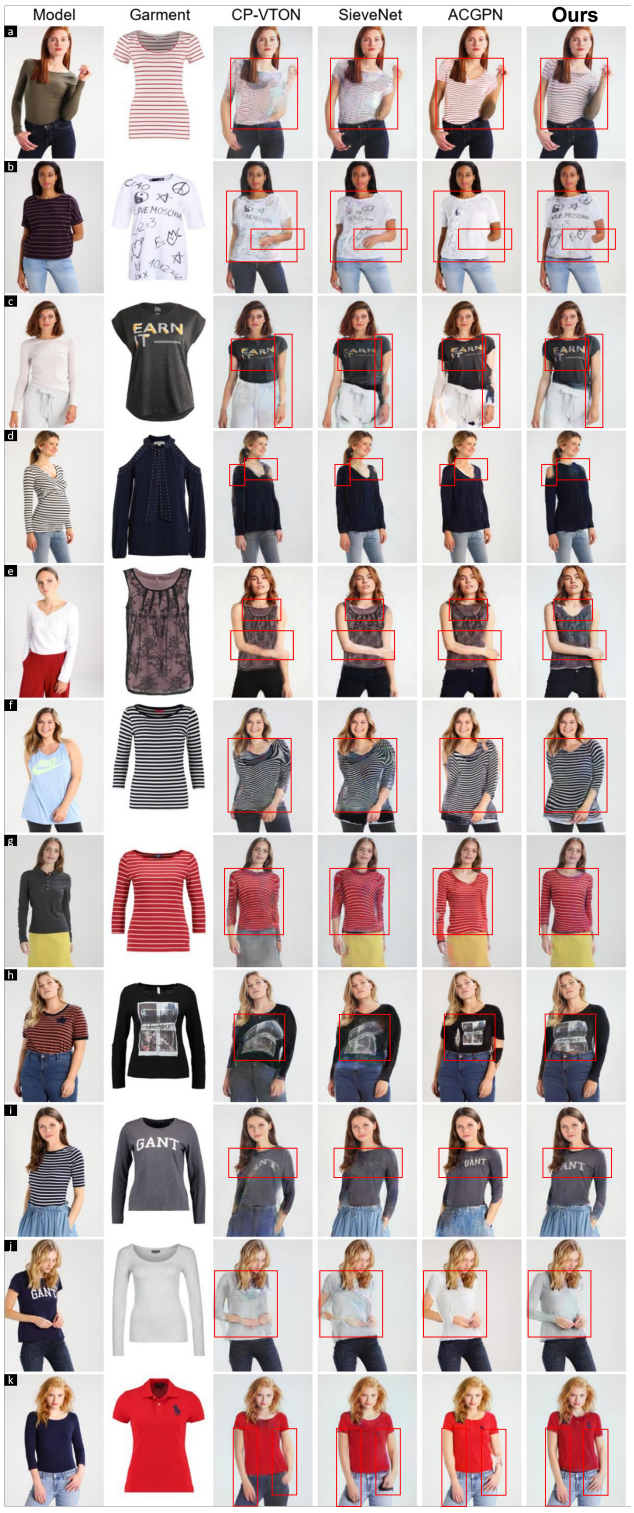

Figure 3. Qualitative Comparison of MeshNet with [41, 20, 42]. Rows **1-5** reflect improvements in preserving geometric integrity, and Rows **6-10**, texture integrity. Please note: a) Complex poses b) Depth ordering of body parts c) Skin generation (d, e). Neckline and shoulder correction (f, g) Pattern (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 for more results.

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

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

Figure 6. Employing the IUV priors during texture fusion improves the neckline (row 1), depth perception (row 2) and skin-generation (row 3) in the try-on output

7. Conclusion

We introduce MeshNet, an end-to-end try-on framework, which utilizes a combination of gated aggregation of hierarchical flow estimates ($GatedFlow$) and dense geometric

priors to render try-on outputs with fewer artefacts. We highlight effectiveness of MeshNet through comparisons with state-of-the-art and detailed ablation studies. Finally, we also validate the generalized efficacy of our proposed *GatedFlow* by: i) proposing and contrasting different flow aggregation mechanism for virtual try-on and, ii) adapting it to improve state-of-the-art work in 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.

[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.

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

[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.

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

[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.

[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.

[8] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9026–9035, 2019.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[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.