Fashion is Taking Shape:
Understanding Clothing Preference Based on Body Shape From Online Sources

Hosnieh Sattar[1]   Gerard Pons-Moll[1]   Mario Fritz[2]

[1]Max Planck Institute for Informatics, [2]CISPA Helmholtz Center i.G.

Saarland Informatics Campus, Saarbrücken, Germany,

{sattar,gpons}@mpi-inf.mpg.de, fritz@cispa.saarland

Abstract

To study the correlation between clothing garments and body shape, we collected a new dataset (Fashion Takes Shape), which includes images of female users with clothing category annotations. Despite the progress in body shape estimation from images, it turns out to be challenging to infer body shape from such diverse, real-world photos. Hence, we propose a novel and robust multi-photo approach to estimate body shapes of each user and build a conditional model of clothing categories given body-shape. We demonstrate that in real-world data, clothing categories and body-shapes are correlated and show that our multi-photo approach leads to a better predictive model for clothing categories compared to models based on single-view shape estimates or manually annotated body types. We see our method as the first step towards the large-scale understanding of clothing preferences from body shape.

1. Introduction

Fashion is a \$2.4 trillion industry[1] which plays a crucial role in the global economy. Many e-commerce companies such as Amazon or Zalando makes it possible for their users to buy clothing online. However, based on a recent study, [2] around 50% of bought items were returned by users. One major reason of return is "It doesnt fit" (52%). Fit goes beyond the mere size — certain items look good on certain body shapes and others do not. In contrast to in store shopping where one can try on clothing, in online shopping users are limited to a coarse set of numeric size ranges (e.g. 36, 38 and so on) to predict the fitness of the clothing item. Also, they only see the clothing worn by a professional model, which does not account for the diverse body shape of peo-
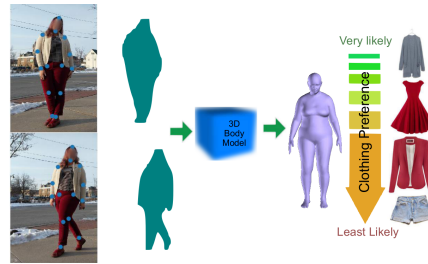

Figure 1: Our multi-photo approach uses 2D body joint and silhouette to estimate 3D body shape of the person in the photo. Our shape conditioned model of clothing categories uses the estimated shape to predict the best fitting clothing categories.

ple. A clothing item that looks very good on a professional model body could look very different on another person's body. Consequently, understanding how body shape correlates with people's clothing preferences could avoid such confusions and reduce the number of returns.

Due to importance of the fashion industry, the application of computer vision in fashion is rising rapidly. Especially, clothing recommendation [22, 36, 27, 14] is one of the hot topics in this field along with cloth parsing [42, 44, 11], recognition [9, 27, 17, 13, 2] and retrieval [40, 41, 23, 26]. Research in the domain of clothing recommendation studies the relation between clothing and categories, location, travel destination and weather. However, there is no study on the correlation between human body shape and clothing. This is probably due to the fact that there exists no dataset with clothing category annotations together with detailed shape annotations.

Therefore, our main idea is to leverage fashion photos of users including clothing category meta-data and, for every user, automatically estimating their body shape. Using this data we learn a conditional model of clothing given the inferred body shape.

Despite recent progress, the visual inference of body

---

shape in unconstrained images remains a very challenging problem in computer vision. People appear in different poses, wearing many different types of garments, and photos are taken from different camera viewpoints and focal lengths.

Our key observation is that users typically post several photos of themselves, while viewpoint and body pose varies across photos, the body shape *does not* (within a posted gallery of images). Hence, we propose a method that leverages multiple photos of the same person to estimate their body shape. Concretely, we first estimate body shape by fitting the SMPL body model [29] to each of the photos separately, and demonstrate that exhaustively searching for depth improves performance. Then, we reject photos that produce outlier shapes and optimize for a single shape that is consistent with each of the inlier photos. This results in a robust *multi-photo* method to estimate body shape from unconstrained photos on the internet.

We use image from the web [3] and collected a dataset (*Fashion Takes Shape*) which includes more than 18000 images with meta-data including clothing category, and a *manual shape annotation* indicating whether the person's shape is above average or average. The data comprises 181 different users. Using our multi-photo method, we estimated the shape of each user. This allowed us to study the relationship between clothing categories and body shape. In particular, we compute the conditional distribution of clothing category conditioned on body shape parameters.

To validate our conditional model, we compute the likelihood of the data (clothing categories worn by the user) under the model and compare it against multiple baselines, including a marginal model of clothing categories, a conditional model built using the manual shape annotations, and a conditional model using a state of the art single view shape estimation method [7].

Experiments demonstrate that our conditional model with multi-photo shape estimates always produces better data-likelihood scores than the baselines. Notably, our model using automatic multi-photo shape estimation even outperforms a model using a coarse manual shape annotations. This shows that we extract more fine-grained shape information than manual annotations. This is remarkable, considering the unavoidable errors that sometimes occur in automatic shape estimation from monocular images.

We see our method as the first step towards the large-scale understanding of clothing preferences from body shape. To stimulate further research in this direction, we will make the newly collected Fashion Takes Shape Dataset (FTS), and code available to the community. FTS includes clothing meta-data, 2D joint detection, semantic segmentation and our 3D shape-pose estimates.

---

³http://www.chictopia.com

2. Related Work

There is no previous work relating body shape to clothing preferences; here we review works that apply computer vision for fashion, and body shape estimation methods.

Fashion Understanding in Computer Vision. Recently, fashion image understanding has gained a lot of attention in computer vision community, due to large range of its human-centric applications such as clothing recommendation [22, 36, 27, 14, 18], retrieval [40, 41, 23, 26, 1], recognition [9, 27, 17, 13, 2], parsing [42, 44] and fashion landmark detection [27, 28, 39].

Whereas earlier work in this domain used handcrafted features (e.g. SIFT, HOG) to represent clothing [9, 22, 40, 26], newer approaches use deep learning [6] which outperforms prior work by a large margin. This is thanks to availability of large-scale fashion datasets [36, 27, 28, 13, 34] and blogs. Recent works in clothing recommendation leverage meta-data from fashion blogs. In particular, Han et al. [13] used the fashion collages to suggest outfits using multimodal user input. Zhang et al. [47] studied the correlation between clothing and geographical locations and introduced a method to automatically recommend location-oriented clothing. In another work, Simo-Serra et al. [36] used user votes of fashion outfits to obtain a measure of fashionability.

Although the relationship between location, users vote and fashion compatibilities is well investigated, there is no work which studies the relationship between human body shape and clothing. In this work, we introduce an automatic method to estimate 3D shape and a model that relates it to clothing preferences. We also introduce a new dataset to promote further research in this direction.

Virtual try-on: Another popular application of computer vision and computer graphics to fashion is virtual try-on, which boils down to a clothing re-targeting problem. Pons-Moll et al. [32] jointly capture body shape and 3D clothing geometry – which can be re-targeted to new bodies and poses. Other works by pass 3D inference; using simple proxies for the body, Han et al. [15] retarget clothing to people directly in image space. Using deep learning and leveraging SMPL [29], Lassner et al. [24] predicts images of people in clothing, and Zanfir et al. [45] transfers appearance between subjects. These works leverage a body model to re-target clothing but do not study the correlation between body shape and clothing categories.

3D Body Shape Estimation. Recovery of 3D human shape from a 2D image is a very challenging task which has been facilitated by the availability of 3D generative body models learned from thousands of scans of people [4, 33, 29]. Such models capture anthropometric constraints of the population and therefore reduce ambiguities.

Several works [35, 12, 16, 48, 10, 7, 19, 16, 48, 10] leverage these generative models to estimate 3D shape from single images using shading cues, silhouettes and appearance.

Recent model based approaches leverage deep learning based 2D detections [8] – by either fitting a model to them at test time [7, 3] or by using them to supervise bottom-up 3D shape predictors [30, 21, 38, 37]. Similar to [7], we fit the SMPL model to 2D joint detections, but, in order to obtain better shape estimates, we include a silhouette term in the objective like [3, 19]. In contrast to previous work, we leverage multiple web photos of the same person in different poses. In particular, we jointly optimize a single coherent static shape and pose for each of the images. This makes our multi-photo shape estimation approach robust to difficult poses and shape occluded by clothing. Other works have exploited temporal information in a sequence to estimate shape under clothing [5, 46, 43] in constrained settings – in contrast we leverage web photos without known camera parameters. Furthermore, we can not assume pose coherency over time [19] since our input are photos with varied poses. None of previous work leverage multiple unconstrained photos of a person to estimate body shape.

3. Robust Human Body Shape Estimation from Photo-Collections

Our goal is to relate clothing preferences to body shape automatically inferred from photo-collections. Here, we build on the SMPL [29] statistical body model that we fit to images. However, unconstrained online images make the problem very hard due to varying pose, clothing, illumination and depth ambiguities.

To address these challenges, we propose a robust multi-photo estimation method. In contrast to controlled multi-view settings where the person is captured simultaneously by multiple cameras, we devise a method to estimate shape leveraging multiple photos of the same person in different poses and camera viewpoints.

From a collection of photos, our method starts by fitting SMPL (Sec. 3.1) to each of the images. This part is similar to [7, 3] and not part of our contribution and we describe it in Sec. 3.2 for completeness. We demonstrate (Sec. 3.3) that keeping the height of the person fixed and initializing optimization at multiple depths significantly improves results and reduces scale ambiguities. Then, we reject photos that result in outlier shape estimates. Using the inlier photos, our multi-photo method (Sec. 3.4) jointly optimizes for multiple cameras, multiple poses, and a *single coherent shape*.

3.1. Body Model

SMPL [29] is a state of the art generative body model, which parameterizes the surface of the human body with shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$. The shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the

PCA coefficients of a shape space learned from thousands of registered 3D scans. The shape parameters encode changes in height, weight and body proportions. The body pose $\boldsymbol{\theta} \in \mathbb{R}^{3P}$, is defined by a skeleton rig with $P = 24$ joints. The joints $J(\boldsymbol{\beta})$ are a function of shape parameters. The SMPL function $M(\boldsymbol{\beta}, \boldsymbol{\theta})$ outputs the $N = 6890$ vertices of the human mesh transformed by pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$.

In order to "pose" the 3D joints, SMPL applies global rigid transformations $\mathbf{R}_{\boldsymbol{\theta}}$ on each joint i as $R_{\boldsymbol{\theta}}(J_i(\boldsymbol{\beta}))$.

3.2. Single View Fitting

We fit the SMPL model to 2D body joint detection $\mathbf{J}_{est}$ obtained using [20], and a foreground mask $\mathbf{S}$ computed using [31]. Concretely, we minimize an objective function with respect to pose, shape and camera translation $\mathbf{K} = [X, Y, Z]$:

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K}; \mathbf{J}_{\text{est}}, \mathbf{S}) = E_P(\boldsymbol{\beta}, \boldsymbol{\theta}) + E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{K}, \mathbf{J}_{\text{est}}) \\ + E_h(\boldsymbol{\beta}) + E_S(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K}; \mathbf{S}), \quad (1)$$

where $E_P(\boldsymbol{\beta}, \boldsymbol{\theta})$[4] are the four prior terms as described in [7], and the other terms are described next.

Joint-based data term: We minimize the re-projection error between SMPL 3D joints and detection:

$$E_J(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K}; \mathbf{J}_{\text{est}}) = \sum_{\text{joint} j} \omega_j \rho(\Pi_{\mathbf{K}}(R_{\theta}(J(\boldsymbol{\beta})_j)) - \mathbf{J}_{\text{est}, j})$$

(2)

where $\Pi_{\mathbf{K}}$ is the projection from 3D to 2D of the camera with parameters $\mathbf{K}$. ω_j are the confidence scores from CNN detection and ρ a Geman-McClure penalty function which is robust to noise.

Height term: Previous work [7, 25] jointly optimizes for depth (distance from the person to camera) and body shape. However, the overall size of the body and distance to the camera are ambiguous; a small person closer to the camera can produce a silhouette in the image just as big as a bigger person farther from the camera. Hence, we aim at estimating body shape up to a scale factor. To that end, we add an additional term that constrains the height of the person to remain very close to the mean height T_H of the SMPL template

$$E_h(\boldsymbol{\beta}) = ||M_h(0, \boldsymbol{\beta}) - T_H||_2^2,$$

where height $M_h(0, \boldsymbol{\beta})$, is computed on the optimized SMPL model before applying pose. This step is especially crucial for multi-photo optimization as it allows us to analyze shapes at the same scale.

---

[4]For details please look at the supplementary material

Silhouette term: To capture shape better, we minimize the miss-match between the model silhouette $\mathbf{I_s}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{K})$, and the distance transform of the CNN-segmented mask $\mathbf{S}$ [31]:

$$E_S(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K}; \mathbf{S}) = G(\lambda_1 \mathbf{I_s}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K})\mathbf{S} + \lambda_2(\mathbf{1} - \mathbf{I_s}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{K}))\overline{\mathbf{S}}), \quad (3)$$

where G is a Gaussian pyramid with 4 different levels, $\mathbf{K}$ is the camera parameter, and $\overline{\mathbf{S}}$ is the distance transform of the inverse segmentation, and λ is a weight balancing the terms.

3.3. Camera Optimization

Camera translation and body orientation are unknown. However, we assume that rough estimation of the focal length is known. We set the focal length as two times the width of the image. We initialize the depth Z via the ratio of similar triangles, defined by the shoulder to ankle length of the 3D model and the 2D joint detection. To refine the estimated depth, we minimize the re-projection error E_J of only torso, knee, and ankle joints (6 joints) with respect to camera translation and body root orientation. At this stage, $\boldsymbol{\beta}$ is held fixed to the template shape. We empirically found that a good depth initialization is crucial for good performance. Hence, we minimize the objective in 1 at 5 different depth initializations – we sample in the range of [-1,+1] meters from the initial depth estimate. We keep the shape estimate from the initialization that leads to a lower minimum after convergence. After obtaining the initial pose and shape parameter we refine the body shape model adding silhouette information.

3.4. Robust Multi-Photo Optimization

The accuracy of the single view method heavily depends on the image view-point, the pose, the segmentation and 2D joint detection quality. Therefore, we propose to jointly optimize one shape to fit several photos at once. Before optimizing, we reject photos that are likely to be outliers in order to make the optimization more robust. First, we compute the median shape from all the single view estimates and keep only the views whose shape is closest to the median. Using these inlier views we jointly optimize V poses $\boldsymbol{\theta}^i$, and a single shape $\boldsymbol{\beta}$. We minimize the re-projection error in all the photos at once:

$$E_{mp}(\boldsymbol{\beta}, \boldsymbol{\theta}^{\forall i}, \mathbf{K}^{\forall i}; \mathbf{J}_{\text{est}}^{\forall i}, \mathbf{S}^{\forall i}) = \sum_{i=1}^{K} E(\boldsymbol{\beta}, \boldsymbol{\theta}^i, \mathbf{K}^i; \mathbf{J}_{\text{est}}^i, \mathbf{S}^i)$$

where $E()$ is the single view objective function, Eq. 1 and K are number of views we kept after outlier rejection. Our multi-photo method leads to more accurate shape estimates as we show in the experiments.

4. Evaluation

To evaluate our method, we proposed two datasets: synthetic and real images. We used the synthetic dataset to perform an ablation study of our multi-photo body model. Using unconstrained real-world fashion images, we evaluate our clothing category model conditioned on the multi-photo shape estimates.

4.1. Synthetic Bodies

SMPL is a generative 3D body model which is parametrized with pose and shape. We observe that while the first shape parameter produces body shape variations due to scale, the second parameter produces shapes of varied weight and form. Hence, we generate 9 different bodies by sampling the second shape parameter from $\mathcal{N}(\mu, 1)$ and $\mu \in [-2 : 0.5 : 2]$. In Figure 2a, we show the 9 different body shapes and a representative rendered body silhouette with 2D joints which we use as input for prediction. We also generate 9 different views for each subject to evaluate our multi-photo shape estimation in a controlled setting (Figure 2b). In all experiments, we report the mean Euclidean error between the estimated shape parameters $\boldsymbol{\beta}$ and the ground truth shapes.

$\mu = -2 \quad \mu = -1.5 \quad \mu = -1 \quad \mu = -0.5 \quad \mu = 0 \quad \mu = 0.5 \quad \mu = 1 \quad \mu = 1.5 \quad \mu = 2 \qquad$ Input

(a) Variation in body shape

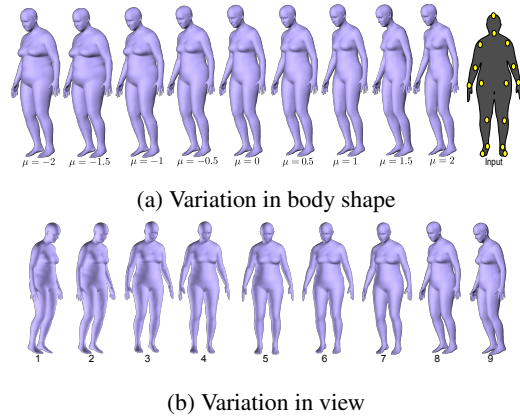1 2 3 4 5 6 7 8 9

(b) Variation in view

Figure 2: Using SMPL 3D body model we generate 9 subject each with 9 views. We study the effectiveness of our method on this dataset. As shown in (a) the input to our system is only image silhouette and a set of 2D body joints.
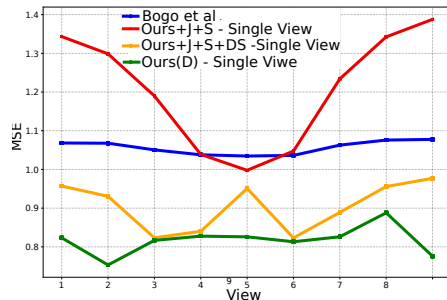

Figure 3: The plot shows the mean euclidean norm between estimated and the ground truth shape among all subjects for each view on synthetic data.

Single-View Shape Estimation: In this controlled, synthetic setting, we have tested our model in several conditions. We summarize the results of our single view method in the first column of Table 1 (mean shape error over all 9 subjects and 9 views), and plot the error w.r.t. viewpoints in Figure 3. Please note in Table 1 that SMPLify[7] can only use one photo, and therefore columns corresponding to multiple photos are marked as "na" (not available).

Overall, we see in Table 1 a reduction of shape estimation error from 1.05 by SMPLify [7] to 0.91 of our method by adding joint estimation (J), silhouette features (S) and depth selection (DS). The depth selection (DS) strategy yield the strongest improvement. We see an additional decrease to 0.84 by considering multiple photos ($k = 5$) despite having to estimate camera and body pose for each additional photo.

People with similar joint length could have different body mass. As SMPLify only uses 2D body joint as input, it is not able to estimate the body shape with high accuracy. Silhouette is used to capture a better body shape. However adding silhouette with a wrong depth data, decreases the accuracy of shape estimation drastically (Ours using 2D joint and silhouette (Ours+J+S), red curve in Figure 3 from 0.91 to 1.20. Hence, to study the impact of depth accuracy in shape estimation, we provide ground-truth depth (in Figure 3, green curve: *Ours(D)*) to our method. We observe that using ground truth depth information improves the error of Our+J+S from 1.20 to 0.86.

This argues for the importance of our introduced depth search procedure. Indeed, we find that our model with depth selection ("Ours+J+S+DS") yields a reduced error of 0.91 *without using any ground truth information.*

Multi-Photo Shape Estimation: Table 1 present also our result for multi-photo case. Real-world images exhibit noisy silhouettes and 2D joints, body occlusion, variation in camera parameters, viewpoints and poses. Consequently, we need a robust system that can use all the information to obtain an optimized shape. Since every single photo may not give us a very good shape estimation, we jointly optimize all photos together. However, for certain views, estimating the pose and depth is very difficult. Consequently, adding those views leads to worse performance. Hence, before optimization, we retain only the K views with shape estimates closest to the median shape estimate of all views. This effectively rejects outlier views. The results are summarized in Table 1. K is the number of photos we kept out of the total of 9 to perform optimization. Using only 2D joint data, we optimized the shape in multi-photo setting (Ours+J in Table 1). In the second step we add silhouette term to our method in multi-photo optimization (Ours+J+S). Both of these experiments shows decrease in accuracy of the the estimated shape compared to SMPLify. However, for our full method which uses up to $K = 5$ views, we observe a

	single-view	k=2	k=3	k=4	k=5	k=6	k=7
SMPLify [7]	1.05	na	na	na	na	na	na
Ours+J	1.05	1.81	1.80	1.80	1.80	1.80	1.82
Ours+J+S	1.20	1.80	1.80	1.80	1.80	1.80	1.77
Ours+J+S+DS	0.91	0.91	0.88	0.87	0.84	0.85	0.88
Ours(D)	0.86	0.84	0.79	0.77	0.80	0.83	0.92

Table 1: We present the results for Multi-Photo optimization on the synthetic data. The error is the L2 distance between ground-truth and estimated shape parameter. Ours(D) has ground truth camera translation(depth) data.

consistent decrease in error; beyond 5 the error increases, which supports the effectiveness of the proposed integration and outlier detection scheme. We improve over our single view estimate reducing the error *further* from 0.91 to 0.84 (for $K = 5$) – and even approach the oracle performance (0.86) where ground truth depth is given.
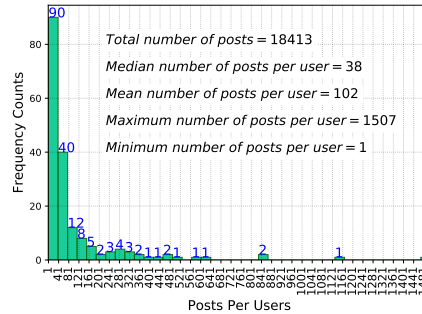

Figure 4: Histogram of posts in our dataset. A total number of posts from all our users are 18413. Each post has 1 or more images of a person with clothing.
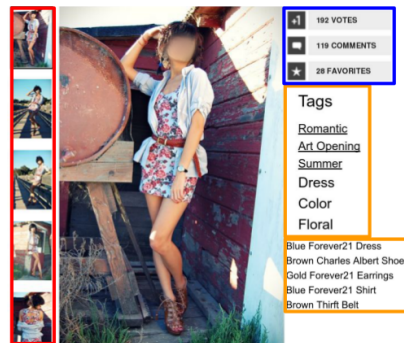

Figure 5: Chictopia's users upload images of themselves wearing different garments. Each post has 1 or more image of the person(red box). In addition to images, meta data such as "Tags"(orange box) and users opinion(blue box) are available.

4.2. Fashion Takes Shape Dataset

Not every clothing item matches every body shape. Hence our goal is to study the link between body shapes and clothing categories. In order to study these correlations, we collected data from 181 female users of "Chictopia"[5] (online fashion blog).

We look for two sets of users: in the first set, we collected data of users with average and below the average size, which we call group G_a; the second set contains data of above average (plus size) users referred to as G_p. In total, we have 141 users in group G_a and 40 in group G_p which constitute a diverse sample of real-world body shapes. Figure 4, shows the summary of our dataset. The total number of posts from all users is 18413 – each post can contains one or more images (usually between 2 to 4). The minimum number of posts per user is 1 and the maximum 1507. In average, we have 102 posts per user, and a median of 38 posts per user. Furthermore, each post contains data about clothing garments, other users opinions (Votes, likes) and comments. Figure 5 shows a post uploaded by a user.
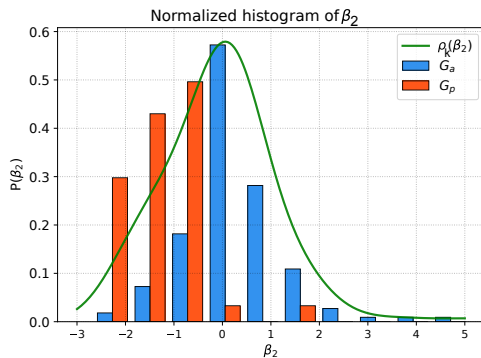

Figure 6: Shape distribution of our dataset. The thinner the person the higher values of β_2 they have. While the group G_p has lower values (negative).

4.3. Shape Representation

In order to build a model conditioned on shape, we first need a representation of the users' shapes. Physical body measurements can be considered as an option which is not possible when we only have access to images of the person. Hence, we use our multi-photo method to obtain a shape $\beta \in \mathbb{R}^{10}$ estimate of the person from multiple photos.

Since we do not have ground truth shape for the unconstrained photos, we have trained a binary Support Vector Machine (SVM) on the estimated shape parameters β for classification of the body type into G_a and G_p. The intuition is that if our shape estimations are correct, above average and below average shapes should be separable. Indeed, the

SVM obtains an accuracy (on a hold-out test set) of 87.17% showing that the shape parameter is at the very least informative of the two aforementioned body classes. Looking at the SVM weights, we recognize that the second entry of the β vector has the most contribution to the classifier. Actually, classifying the data by simple thresholding of the second dimension of the β vector results in an even higher accuracy of 88.79%. Hence, for following studies, we have used directly the second dimension of β. We illustrate the normalized histogram of this variable in Figure 6. The normalized histogram suggests that users in group G_p have negative values whereas group G_a have positive values. For later use, we estimate a probability density function (pdf) for this variable β_2 with a kernel density estimator – using a Gaussian kernel:

$$\rho_K(\beta_2) = \frac{1}{N} \sum_{i=1}^{N} K((\beta_2 - \beta_2^i)/h) \qquad (4)$$

The green line in Figure 6 illustrate the estimated pdf $\rho_K(\beta_2)$ of all users.

4.4. Correlation Between Shape and Clothing Categories

The type of clothing garment people wear is very closely correlated with their shape. For example, "Leggings" might look good on one body shape but may not look very good on other shapes. Hence, we introduce 3 models to study the correlation of the shape and clothing categories. Our basic model (**Model 1**) uses only data statistics with no information about users shape. In the second approach (**Model 2**), clothing is conditioned on binary shape categories G_a and G_p – which in fact requires manual labels. The final approach (**Model 3**) is facilitated by our automatic shape parameter estimation β_2.

We evaluate the quality of the model via the negative log likelihood of held-out data. A good model minimizes the negative log-likelihood. Hence a better model should have smaller value in negative log likelihood. The negative log likelihood is defined as:

$$LL = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{Z} \log P_M^{c_j^i} \qquad (5)$$

Where N is the number of users, Z is the number of clothing categories and $c_j^i \in C_i$ is a vector of user's clothing categories. M represents the model. In the following, we present the details of these three approaches. The log likelihood of each approach is reported in the Table 2.

Model 1: Prediction Using Probability of Clothing Categories: We established a basic model using the probability of the clothing categories $P_M = p(c)$. The clothing categories tag of the dataset is parsed for fourteen of the most
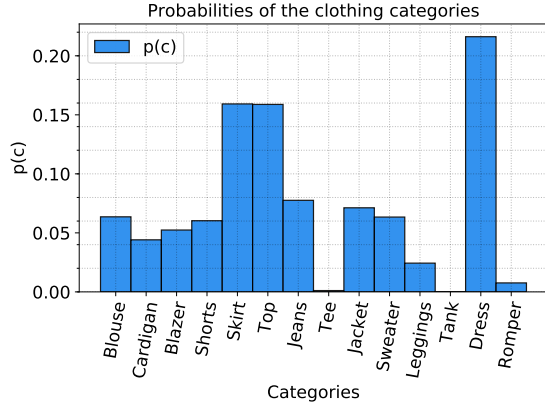
Figure 7: Probabilities of 14 different clothing garments of our dataset.

common clothing's categories. Category "Dress" has the highest amount of images (Figure 7) whereas "Tee and Tank" were not tagged very often.

Model 2: Prediction for Given G_a and G_p: This is based on the conditional probability of the clothing category given the annotated body type $P_M = p(c|G)$ where $G \in G_a, G_p$. From this measure we find out that several clothing categories are more likely for certain group (Figure 8). As an example, while "Cardigan" and "Jacket" have higher probabilities for the G_p group, users in G_a were more likely to wear "Short" and "Skirt".
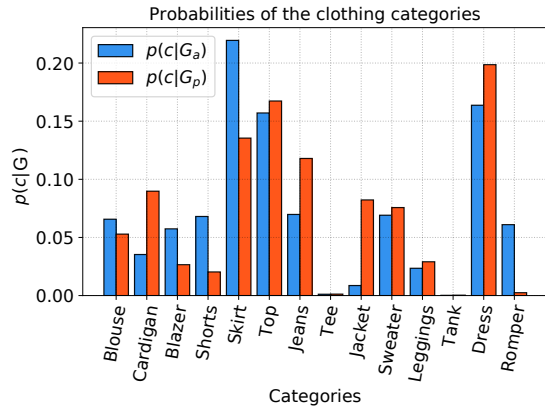

Figure 8: Given the Body type (G_a and G_p), we measured the probability of each clothing category in our dataset.

Model 3: Prediction for a Given Shape β_2: As shown in the second model, body types and clothing garments are correlated. However, categorizing people only into two or more categories is not desirable. First, it requires tedious and time consuming manual annotation of body type. Second,

the definition of the shape categories is very personal and fuzzy.

The estimated shape parameters of our model provides us with a continuous fine-grained representation of human body shape. Hence, we no further need to classify people in arbitrary shape groups. Using the shape parameter β_2 and statistics of our data, we are able to measure the conditional probability of shapes for a given clothing category $p(\beta_2|c)$. This probability is measured for wearing and not wearing a certain category. The result is shown in Figure 9 where for each category the Blue line represents wearing the category. Similarly to the previous model, one can see users with negative values of β_2 wearing "Cardigan", where the probabilities of wearing "Short" and "skirt" is skewed towards positive values of β_2. Furthermore, using the Bayes rule we can predict clothing condition on the body shape $P_M = p(c|\beta_2)$ as:

$$p(c|\beta_2) = \frac{p(\beta_2|c)p(c)}{P(\beta_2)} \qquad (6)$$

The green line in Figure 9 illustrate the $p(c|\beta_2)$.

Negative Log likelihood We quantify the quality of our prediction models by the negative log likelihood of held out data. As we are using negative of log likelihood, the model with smallest values is the best. The results, of each model on our dataset, is summarized in the Table 2. Also, we used the estimated shape parameters of Bogo et al [7] and ours for comparison. In addition to our method which optimized multi-photo, we only used the median shape among photos of a user as a baseline as well. We also include the likelihood under the prior as a reference. For better analysis, we split the users into 4 groups: The first group contains users which there is only a single photo ($I = 1$) of them in each post. In the second group, users always have 2 images($I = 2$) and third group contains users with 3 or more images of a clothing($I \geq 3$). Finally, we also show results with taking into account all users. Table 2, shows that our method obtained the smallest negative log-likelihood on the full dataset – in particular outperforming the model that conditions on the two discrete labeled shape classes, shape based on prior work SMPlify[7], as well as a naive multi-photo integration based on a median estimate. While the median estimate is comparable if only two views are available, we see significant gains for multiple viewpoint – that also show on the full dataset.

5. Qualitative Results on Shape Estimation

In Figure 10 we present example results obtained with our method and compare it to the result obtained with SMPLify [7]. SMPLify fits the body model based on 2D positions of body joints that often do not provide enough information regarding body girth. This leads to shape estimates
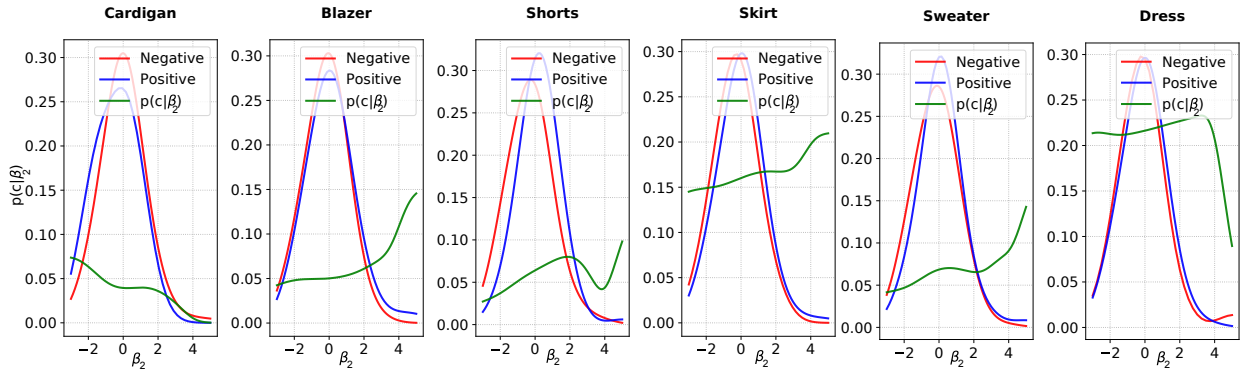
Figure 9: Probabilities of $p(\beta_2|c)$ for wearing (blue curves) and not wearing (red curves) of a clothing category on our dataset. Using Bayes rule we can estimate the probability of clothing given the shape $p(c|\beta_2)$(green curve). Negative values of β_2 corresponds to above average while average and below average users have positive values for β_2.

	Model 1 p(c)	Model 2 $p(c|G_a/G_p)$	$p(c|\beta_2)^1$	Medianβ_2	Model 3 $p(c|\beta_2)^2$
I = 1	12.81	12.80	13.63	-	**12.16**
I = 2	13.31	13.47	13.34	**13.09**	13.11
I ≥ 3	19.06	19.11	18.8	18.59	**17.85**
All	20.13	20.39	20.48	20.12	**19.81**

Table 2: We measured the negative Log-Likelihood of our different models on held out data. Numbers are comparable within the rows. Smaller is better. $p(c|\beta_2)^1$ uses estimated shape from SMPLify[7] and $p(c|\beta_2)^2$ uses our estimated shape.

that are rather close to the average body shape for above-average body sizes (rows 1 and 2 in Figure 10). SMLPify also occasionally fails to select the correct depth that results in body shape that is too tall and has bent knees (red box). The single-view variant of our approach improves over the result of SIMPLify for the first example in Figure 10. However it still fails to estimate the fine-grained pose details such as orientation of the feet (blue box). In the second example in Figure 10 the body segmentation includes a handbag resulting in a shape estimate with exaggerated girth by our single-view approach (yellow box). These mistakes are corrected by our multi-photo approach that is able to improve feet orientation in the first example (blue box) and body shape in the second example (yellow box).

6. Conclusion

In this paper we aimed to understand the connection between body shapes and clothing preferences by collecting and analyzing a large database of fashion photographs with annotated clothing categories. Our results demonstrate that clothing preferences and body shapes are correlated and that we can build predictive models for clothing categories based on the output of automatic shape estimation. To ob-

Figure 10: Shape estimation results on real data. Note that the shape estimates obtained with SMPLify [7] are rather close to the average body shape whereas our multi-photo approach is able to recover shape details more accurately both for above-average (rows 1 and 2) and average (rows 3 and 4) body types.

tain estimates of 3D shape we proposed a new approach that incorporates evidence from multiple photographs and body segmentation and is more accurate than popular recent SMPLify approach [7]. We are making our data and code available for research purposes.

References

[1] K. Ak, A. Kassim, J. Hwee Lim, and J. Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, 2018. 2

[2] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 1, 2

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *CVPR Spotlight*, 2018. 3

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, 2005. 2

[5] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *ECCV*, 2008. 3

[6] X. W. S. Y. Bo Zhao, Jiashi Feng. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, 2017. 2

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2, 3, 5, 7, 8

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[9] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 1, 2

[10] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In *ECCV*, 2010. 3

[11] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 1

[12] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 3

[13] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 1, 2

[14] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *ACM on Multimedia Conference*, 2017. 1, 2

[15] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. *arXiv preprint arXiv:1711.08447*, 2017. 2

[16] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel. Multilinear pose and body shape estismation of dressed subjects from image sets. In *CVPR*, 2010. 3

[17] W.-L. Hsiao and K. Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *ICCV*, 2017. 1, 2

[18] W.-L. Hsiao and K. Grauman. Creating capsule wardrobes from fashion images. In *CVPR*, 2018. 2

[19] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 3

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schieke. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 3

[21] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[22] M. H. Kiapour, Y. M. Hadi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 1, 2

[23] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 1, 2

[24] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *ICCV*, 2017. 2

[25] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 3

[26] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. *CVPR*, 2012. 1, 2

[27] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, June 2016. 1, 2

[28] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2

[29] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, oct 2015. 2, 3

[30] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018. 3

[31] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollr. Learning to refine object segments. In *ECCV*, 2016. 3, 4

[32] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017. 2

[33] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015. 2

[34] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-gen: The generative fashion dataset and challenge. In *arXiv preprint arXiv:1806.08317*, 2018. 2

[35] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*. 2008. 3

[36] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In *CVPR*, 2015. 1, 2

[37] V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, 2017. 3

[38] H. Tung, H. Wei, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3

[39] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 2

[40] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM International Conference on Multimedia*, 2011. 1, 2

[41] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 1, 2

[42] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1, 2

[43] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *ECCV*, Amsterdam, Netherlands, 2016. 3

[44] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2013. 1, 2

[45] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *CVPR*, 2018. 2

[46] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *CVPR*, 2017. 3

[47] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian. Trip outfits advisor: Location-oriented clothing recommendation. *IEEE Transactions on Multimedia*, 2017. 2

[48] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics*, 2010. 3