ViBE: Dressing for Diverse Body Shapes

Wei-Lin Hsiao¹,²       Kristen Grauman¹,²
¹The University of Texas at Austin       ²Facebook AI Research

Abstract

Body shape plays an important role in determining what garments will best suit a given person, yet today’s clothing recommendation methods take a “one shape fits all” approach. These body-agnostic vision methods and datasets are a barrier to inclusion, ill-equipped to provide good suggestions for diverse body shapes. We introduce ViBE, a Visual Body-aware Embedding that captures clothing’s affinity with different body shapes. Given an image of a person, the proposed embedding identifies garments that will flatter her specific body shape. We show how to learn the embedding from an online catalog displaying fashion models of various shapes and sizes wearing the products, and we devise a method to explain the algorithm’s suggestions for well-fitting garments. We apply our approach to a dataset of diverse subjects, and demonstrate its strong advantages over status quo body-agnostic recommendation, both according to automated metrics and human opinion.

1. Introduction

Research in computer vision is poised to transform the world of consumer fashion. Exciting recent advances can link street photos to catalogs [47, 54], recommend garments to complete a look [25, 33, 40, 73, 76], discover styles and trends [3, 32, 57], and search based on subtle visual properties [22, 56]. All such directions promise to augment and accelerate the clothing shopping experience, providing consumers with personalized recommendations and putting a content-based index of products at their fingertips.

However, when it comes to body shape, state-of-the-art recommendation methods falsely assume a “one shape fits all” approach. Despite the fact that the same garment will flatter different bodies differently, existing methods neglect the significance of an individual’s body shape when estimating the relevance of a given garment or outfit. This limitation stems from two key factors. First, current large-scale datasets are heavily biased to a narrow set of body shapes—typically thin and tall, owing to the fashionista or celebrity photos from which they are drawn [26, 51, 55, 66]. This restricts everything learned downstream, including the extent of bodies considered for virtual try-on [26, 63, 79]. Second, prior methods to gauge clothing compatibility often learn from co-purchase patterns [25, 76, 77] or occasion-based rules [40, 53], divorced from any statistics on body shape.

Body-agnostic vision methods and datasets are thus a barrier to diversity and inclusion. Meanwhile, aspects of fit and cut are paramount to what continues to separate the shopping experience in the physical world from that of the virtual (online) world. It is well-known that a majority of today’s online shopping returns stem from problems with fit [61], and being unable to imagine how a garment would complement one’s body can prevent a shopper from making the purchase altogether.

To overcome this barrier, we propose ViBE, a Visual Body-aware Embedding that captures clothing’s affinity with different body shapes. The learned embedding maps a given body shape and its most complementary garments close together. To train the model, we explore a novel source of Web photo data containing fashion models of di-
verse body shapes. Each model appears in only a subset of all catalog items, and these pairings serve as implicit positive examples for body-garment compatibility.

Having learned these compatibilities, our approach can retrieve body-aware garment recommendations for a new body shape—a task we show is handled poorly by existing body-agnostic models, and is simply impossible for traditional recommendation systems facing a cold start. Furthermore, we show how to visualize what the embedding has learned, by highlighting what properties (sleeve length, fabric, cut, etc.) or localized regions (e.g., neck, waist, straps areas) in a garment are most suitable for a given body shape.

We demonstrate our approach on a new body-diverse dataset spanning thousands of garments. With both quantitative metrics and human subject evaluations, we show the clear advantage of modeling body shape’s interaction with clothing to provide accurate recommendations.

2. Related Work

Fashion styles and compatibility Early work on computer vision for fashion addresses recognition problems, like matching items seen on the street to a catalog [47, 54], searching for products [22, 46, 86], or parsing an outfit into garments [17, 51, 83, 87]. Beyond recognition, recent work explores models for compatibility that score garments for their mutual affinity [24, 33, 34, 36, 73, 76, 77]. Styles—meta-patterns in what people wear—can be learned from images, often with visual attributes [2, 32, 43, 57], and Web photos with timestamps and social media “likes” can help model the relative popularity of trends [50, 74]. Unlike our approach, none of the above models account for the influence of body shape on garment compatibility or style.

Fashion image datasets Celebrities [30, 51], fashionista social media influencers [43, 52, 74, 83, 84], and catalog models [18, 26, 55, 66] are all natural sources of data for computer vision datasets studying fashion. However, these sources inject bias into the body shapes (and other demographics) represented, which can be useful for some applications but limiting for others. Some recent dataset efforts leverage social media and photo sharing platforms like Instagram and Flickr which may access a more inclusive sample of people [42, 57], but their results do not address body shape. We explore a new rich online catalog dataset comprised of models of diverse body shape.

Virtual try on and clothing retargeting Virtual try-on entails visualizing a source garment on a target human subject, as if the person were actually wearing it. Current methods estimate garment draping on a 3D body scan [20, 45, 62, 68], retarget styles for people in 2D images or video [14, 17, 77, 85], or render a virtual try-on with sophisticated image generation methods [23, 29, 63, 79]. While existing methods display a garment on a person, they do not infer whether the garment flatters the body or not. Furthermore, in practice, vision-based results are limited to a narrow set of body shapes (typically tall and thin as in Fig. 1) due to the implicit bias of existing datasets discussed above.

Body and garment shape estimation Estimating people and clothing’s 3D geometry from 2D RGB images has a long history in graphics, broadly categorizable into body only [8, 38, 89], garment only [11, 37, 82, 88], joint [49, 59, 67], and simultaneous but separate estimations [44, 56, 85]. In this work, we integrate two body-based models to estimate a user’s body shape from images. However, different from any of the above, our approach goes beyond estimating body shape to learn the affinity between human body shape and well-fitting garments.

Sizing clothing While most prior work recommends clothing based on an individual’s purchase history [28, 35, 39, 77] or inferred style model [53, 40, 53], limited prior work explores product size recommendation [14, 21, 41, 58, 72]. Given a product and the purchase history of a user, these methods predict whether a given size will be too large, small, or just right. Rather than predict which size of a given garment is appropriate, our goal is to infer which garments will flatter the body shape of a given user. Moreover, unlike our approach, existing methods do not consider the visual content of the garments or person [14, 21, 58, 72]. While SizeNet [41] uses product images, the task is to predict whether the product will have fit issues in general, unconditioned on any person’s body.

Clothing preference based on body shape To our knowledge, the only prior work that considers body shape’s connection to clothing is the “Fashion Takes Shape” project, which studies the correlation between a subject’s weight and clothing categories typically worn (e.g., curvier people are more likely to wear jeans than shorts) [69], and the recommendation system of [30] that discovers which styles are dominant for which celebrity body types given their known body measurements. In contrast to either of these methods, our approach suggests specific garments conditioned on an individual’s body shape. Furthermore, whereas [69] is about observing in hindsight what a collection of people wore, our approach actively makes recommendations for novel bodies and garments. Unlike [30], our method handles data beyond high-fashion celebrities and uses the inferred body shape of a person as input.

3. Approach

While the reasons for selecting clothes are complex [80], fit in a garment is an important factor that contributes to the confidence and comfort of the wearer. Specifically, a garment that fits a wearer well flatters the wearer’s body. Fit is a frequent reason for whether to make an apparel purchase [6]. Searching for the right fit is time-consuming:
women may try on as many as 20 pairs of jeans before they find a pair that fits [64].

The ‘Female Figure Identification Technique (FFIT) System’ classifies the female body into 9 shapes—hourglass, rectangle, triangle, spoon, etc.—using the proportional relationships of dimensions for bust, waist, high hip, and hips [13]. No matter which body type a woman belongs to, researchers find that women participants tend to select clothes to create an hourglass look for themselves [19]. Clothing is used strategically to manage bodily appearance, so that perceived “problem areas/flaws” can be covered up, and assets are accentuated [16,19]. Fig. 2 shows examples from fashion blogs with different styling tips and recommended dresses for different body shapes.

Our goal is to discover such strategies, by learning a body-aware embedding that recommends clothing that complements a specific body and vice versa. We first introduce a dataset and supervision paradigm that allow for learning such an embedding (Sec. 3.1, Sec. 3.2). Then we present our model (Sec. 3.3) and the representation we use for clothing and body shape (Sec. 3.4). Finally, beyond recommending garments, we show how to visualize the strategies learned by our model (Sec. 3.5).

3.1. A Body-Diverse Dataset

An ideal dataset for learning body-garment compatibility should meet the following properties: (1) clothed people with diverse body shapes; (2) full body photos so the body shapes can be estimated; (3) some sort of rating of whether the garment flatters the person to serve as supervision. Datasets with 3D scans of people in clothing [4,5,7,65] meet (1) and (2), but are rather small and have limited clothing styles. Datasets of celebrities [30,51], fashionistas [43,74,84], and catalog models [26,55,66] satisfy (2) and (3), but they lack body shape diversity. Datasets from social media platforms [42,57] include more diverse body shapes (1), but are usually cluttered and show only the upper body, preventing body shape estimation.

To overcome the above limitations, we collect a dataset from an online shopping website called Birdsnest [2]. Birdsnest provides a wide range of sizes (8 to 18 in Australian measurements) in most styles. Fig. 3 shows an example catalog page. It contains the front and back views of the garment, the image of the fashion model wearing the item, her body measurements, and an attribute-like textual description of the item. Most importantly, each item is worn by a variety of models of different body shapes. We collect two categories of items, 958 dresses and 999 tops, spanning 68 fashion models in total. While our approach is not specific to women, since the site has only women’s clothing, our current study is focused accordingly. This data provides us with properties (1) and (2). We next explain how we obtain positive and negative examples from it, property (3).

3.2. Implicit Rating from Catalog Fashion Models

Fashion models wearing a specific catalog item can safely be assumed to have body shapes that are flattered by that garment. Thus, the catalog offers implicit positive body-garment pairings. How do we get negatives? An intuitive way would be to assume that all unobserved body-garment pairings from the dataset are negatives. However,
about 50% of the dresses are worn by only 1 or 2 distinct bodies (3% of the models), suggesting that many positive pairings are not observed.

Instead, we propose to propagate missing positives between similar body shapes. Our assumption is that if two body shapes are very similar, clothing that flatters one will likely flatter the other. To this end, we use k-means [78] clustering (on features defined in Sec. 3.4) to quantize the body shapes in our dataset into five types. Fig. 4 shows bodies sampled from each cluster. We propagate positive clothing pairs from each model observed wearing a garment to all other bodies of her type. Since most of the garments are worn by multiple models, and thus possibly multiple types, we define negative clothing for a type by pairing bodies in that type with clothing never worn by any body in that type.

With this label propagation, most dresses are worn by 2 distinct body types, which is about 40% of the bodies in the dataset, largely decreasing the probability of missing true positives. To validate our label propagation procedure with ground truth, we conduct a user study explicitly asking human judges on Mechanical Turk whether each pair of bodies in the same cluster could wear similar clothing, and whether pairs in different clusters could. Their answers agreed with the propagated labels 81% and 63% of the time for the two respective cases (see Supp. for details).

3.3. Training a Visual Body-Aware Embedding

Now having the dataset with all the desired properties, we introduce our Visual Body-aware Embedding, ViBE, that captures clothing’s affinity with body shapes. In an ideal embedding, nearest neighbors are always relevant instances, while irrelevant instances are separated by a large margin. This goal is achieved by correctly ranking all triplets, where each triplet consists of an anchor \( z_a \), a positive \( z_p \) that is relevant to \( z_a \), and a negative \( z_n \) that is not relevant to \( z_a \). The embedding should rank the positive closer to the anchor than the negative, \( D(z_a, z_p) < D(z_a, z_n) \) (with \( D(\cdot, \cdot) \) denoting Euclidean distance). A margin-based loss [81] optimizes for this ranking:

\[
L(z_a, z_p, z_n) := \left( D(z_a, z_p) - \alpha_p \right)_+ + \left( \alpha_n - D(z_a, z_n) \right)_+
\]

where \( \alpha_p, \alpha_n \) is the margin for positive and negative pairs respectively, and the subscript \( + \) denotes \( \max(0, \cdot) \). We constrain the embedding to live on the \( d \)-dimensional hypersphere for training stability, following [70].

In our joint embedding ViBE, we have two kinds of triplets, one between bodies and clothing, and one between bodies and bodies. So our final loss combines two instances of the margin-based loss:

\[
L = L_{body,cloth} + L_{body, body} \quad (1)
\]

Let \( f_{cloth}, f_{body} \) be the respective functions that map instances of clothing \( x_g \) and body shape \( x_b \) to points in ViBE. For the triplet in our body-clothing loss \( L_{body,cloth}, z_a \) is a mapped body instance \( f_{body}(x_b^a) \), \( z_p \) is a compatible clothing item \( f_{cloth}(x_g^p) \), and \( z_n \) is an incompatible clothing item \( f_{cloth}(x_g^n) \). This loss aims to map body shapes near their compatible clothing items.

We introduce the body-body loss \( L_{body, body} \) to facilitate training stability. Recall that each garment could be compatible with multiple bodies. By simply pulling these shared clothing items closer to all their compatible bodies, all clothing worn on those bodies would also become close to each other, making the embedding at risk of model collapse (see Fig. 5a, blue plot). Hence, we introduce an additional constraint on triplets of bodies: \( z_a \) is again a mapped body instance \( f_{body}(x_b^a) \), \( z_p \) is now a body \( f_{body}(x_b^p) \) that belongs to the same type (i.e., cluster) as \( x_b^a \), and \( z_n \) is a body \( f_{body}(x_b^n) \) from a different type. This body-body loss \( L_{body, body} \) explicitly distinguishes similar bodies from dissimilar ones. Fig. 5a plots the distribution of pairwise clothing distances with and without this additional constraint, showing that this second loss effectively alleviates the model collapse issue.

We stress that the quantization for body types (Sec. 3.2) is solely for propagating labels to form the training triplets. When learning and applying the embedding itself, we operate in a continuous space for the body representation. That is, a new image is mapped to individualized recommendations potentially unique to that image, not a batch of recommendations common to all bodies within a type.

3.4. Clothing and Body Shape Features

Having defined the embedding’s objective, now we describe the input features \( x_b \) and \( x_g \) for bodies and garments. For clothing, we have the front and back view images of the catalog item (without a body) and its textual description.
We use a ResNet-50 [27] pretrained on ImageNet [12] to extract visual features from the catalog images, which captures the overall color, pattern, and silhouette of the clothing. We mine the top frequent words in all descriptions for all catalog entries to build a vocabulary of attributes, and obtain an array of binary attributes for each garment, which captures localized and subtle properties such as specific necklines, sleeve cuts, and fabric.

For body shape, we have images of the fashion models and their measurements for height, bust, waist, and hips, the so-called vital statistics. We concatenate the vital statistics in a 4D array and standardize them. However, the girths and lengths of limbs, the shoulder width, and many other characteristics of the body shape are not captured by the vital statistics, but are visible in the fashion models’ images. Thus, we estimate a 3D human body model from each image to capture these fine-grained shape cues.

To obtain 3D shape estimates, we devise a hybrid approach built from two existing methods, outlined in Fig. 5b. Following the basic strategy of HMD [89], we estimate an initial 3D mesh, and then stage-wise update the 3D mesh by projecting it back to 2D and deforming it to fit the silhouette of the human in the RGB image. However, the initial 3D mesh that HMD is built on, i.e., HMR [38], only supports gender-neutral body shapes. Hence we use SMPLify [8], which does support female bodies, to create the initial mesh. We then deform the mesh with HMD.

Finally, rather than return the mesh itself—whose high-dimensionality presents an obstacle for data-efficient embedding learning—we optimize for a compact set of body shape model parameters that best fits the mesh. In particular, we fit SMPL [56] to the mesh and use its first 10 principal components as our final 3D body representation. These parameters are then used to estimate SMPLify parameters (w/o body-body loss) and other components as shown in Fig. 5b.

In summary, for clothing, we accompany mined attributes (64 and 100 attributes for dresses and tops respectively) with CNN features (2048-D); for body shape, we accompany estimated 3D parameters (10-D) with vital statistics (4-D). Each is first reduced into a lower dimensional space with learned projection functions ($h_{\text{attr}}, h_{\text{cnn}}, h_{\text{smpl}}, h_{\text{meas}}$). Then the reduced attribute and CNN features are concatenated as the representation $x_g$ for clothing, and the reduced SMPL and vital features are concatenated as the representation $x_b$ for body shape. Both are forwarded into the joint embedding (defined in Sec. 3.3) by $f_{\text{cloth}}$ and $f_{\text{body}}$ to measure their affinity. Fig. 6 overviews the entire procedure. See Supp. for architecture details.

3.5. Recommendations and Explanation

After learning our embedding, we make clothing recommendations for a new person by retrieving the garments closest to her body shape in this space. In addition, we propose an automatic approach to convey the underlying strategy learned by our model. The output should be general enough for users to apply to future clothing selections, in the spirit of the expert advice as shown in Fig. 2—e.g., the styling tip for apple body shape is to wear A-line dresses—but potentially even more tailored to the individual body.

To achieve this, we visualize the embedding’s learned decision with separate classifiers (cf. Fig. 10). We first map a subject’s body shape into the learned embedding, and take
To propagate positive labels, we cluster the body shapes to one for tops, and train separately on each for all models. and Sec. 3.2, we collect two sets of data, one for dresses and using the process described in Sec. 3.1 Experiment setup.

We use [60] to localize important regions (as heatmaps) that we can apply CNN visualization techniques [15, 60, 71] (we training a classifier over the CNN features of the clothing, so that uses the exact same features and models as our body-aware embedding functions). It activates the positive or negative prediction.

4. Experiments

We now evaluate our body-aware embedding with both quantitative evaluation and user studies.

Experiment setup. Using the process described in Sec. 3.1 and Sec. 3.2 we collect two sets of data, one for dresses and one for tops, and train separately on each for all models. To propagate positive labels, we cluster the body shapes to k = 5 types. We find the cluster corresponding to an average body type is the largest, while tall and curvy is the smallest. To prevent the largest cluster’s bodies from dominating the evaluation, we randomly hold out 20%, or at least two bodies, for each cluster to comprise the test set. For clothing, we randomly hold out 20% of positive clothing for each cluster. Tab. I summarizes the dataset breakdown.

Baselines. Since no prior work tackles this problem, we develop baselines based on problems most related to ours: user-item recommendation and garment compatibility modeling. Suggesting clothing to flatter a body shape can be treated as a recommendation problem, where people are users and clothing are items. We compare with two standard recommendation methods: (1) body-AGNOSTIC-CF: a vanilla collaborative filtering (CF) model that uses neither users’ nor items’ content; and (2) body-AWARE-CF: a hybrid CF model that uses the body features and clothing visual features as content (“side information” [10]). Both use a popular matrix completion [45] algorithm [29]. In addition, we compare to a (3) body-AWARE-EMBEDDING that uses the exact same features and models as our body-aware-EMBEDDING (ViBE), but—as done implicitly by current methods—is only trained on bodies of the same type, limiting body shape diversity.

Table 1: Dataset statistics: number of garments and fashion models for each clustered type.

<table>
<thead>
<tr>
<th>type</th>
<th>Dresses</th>
<th>Tops</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>body</td>
<td>clothing</td>
</tr>
<tr>
<td>Train</td>
<td>18 7 11 4 6</td>
<td>358 481 301 165 167</td>
</tr>
<tr>
<td>Test</td>
<td>5 2 3 2 2</td>
<td>149 126 76 42 34</td>
</tr>
</tbody>
</table>

The proposed body-body triplet loss is not valid for this baseline.

Figure 7: Recommendation accuracy measured by AUC over all person-garment pairs. Our body-aware embedding (ViBE) outperforms all methods in quantitative evaluation for each method to run the human evaluation.

4.1. Quantitative evaluation

We compare the methods on three different recommendation cases: i) person (“user”) seen but garment (“item”) unseen during training, ii) garment seen but person unseen, iii) neither person nor garment seen. These scenarios capture realistic use cases, where the system must make recommendations for new bodies and/or garments. We exhaust all pairings of test bodies and clothing, and report the mean AUC with standard deviation across 10 runs.

Fig. 7a and Fig. 7b show the results. Our model outperforms all methods by a clear margin. AGNOSTIC-CF performs the worst, as all three test cases involve cold-start problems, and it can only rely on the learned bias terms. Including the person’s body shape and clothing’s features in the CF method (AWARE-CF) significantly boosts its performance, demonstrating the importance of this content for clothing recommendation. In general, the embedding-based
methods perform better than the CF-based methods. This suggests that clothing-body affinity is modeled better by ranking than classification; an embedding can maintain the individual idiosyncrasies of the body shapes and garments.

All methods perform better on dresses than tops. This may be due to the fact that dresses cover a larger portion of the body, and thus could be inherently more selective about which bodies are suitable. In general, the more selective or body-specific a garment is, the more value a body-aware recommendation system can offer; the more body-versatile a garment is, the less impact an intelligent recommendation can have. To quantify this trend, we evaluate the embeddings’ accuracy for scenario (iii) as a function of the test garments’ versatility, as quantified by the number of distinct body types (clusters) that wear the garment. Fig. 8 shows the results. As we focus on the body-specific garments (right hand side of plots) our body-aware embedding’s gain over the body-agnostic baseline increases.

4.2. Example recommendations and explanations

Fig. 9 shows example recommendations for all methods on two heldout subjects: each row is a method, with most and least recommended garments. Being agnostic to body shape, AGNOSTIC-CF and AGNOSTIC-EMBEDDING make near identical recommendations for subjects with different body shapes: top recommended dresses are mostly body-versatile (captured as popularity by the bias term in CF based methods), while least recommended are either body-specific or less interesting, solid shift dresses. ViBE recommends knee-length, extended sleeves, or wrap dresses for curvy subjects, which flow naturally on her body, and recommends shorter dresses that fit or flare for the slender subjects, which could show off her legs.

Fig. 10 shows example explanations (cf. Sec. 3.5) for ViBE’s recommendations. For a petite subject, the most suitable attributes are waistbands and empire styles that create taller looks, and embroidery and ruffles that increase volume. For a curvier subject, the most suitable attributes are extended or 3/4 sleeves that cover the arms, v-necklines that create an extended slimmer appearance, and wrap or side-splits that define waists while revealing curves around upper-legs. The heatmaps showing important regions for why a dress is suitable for the subject closely correspond to these attributes. We also take the top 10 suitable dresses and their heatmaps to generate a weighted average dress to represent the gestalt shape of suitable dresses for this person.

4.3. Human judged ground truth evaluation

Having quantified results against the catalog ground truth, next we solicit human opinions. We recruit 329 subjects on Mechanical Turk to judge which dresses better flatter the body shape of the test subjects. See Supp. for all user study interfaces. We first ask subjects to judge each dress as either body-specific or body-versatile. Then we randomly sample 10 to 25 pairs of clothing items that are the same type (i.e., both body-specific or -versatile) for each of 14 test bodies, and for each one we ask 7 subjects to rank which dress is more suitable for the given body. We discard responses with low consensus (i.e., difference of votes is less than 2), which yields 306 total pairs.

Table 2 shows the results for all methods. The overall trend is consistent with the automatic evaluation in Fig-

![Image](image-url)
We thank our human subjects: Angel, Chelsea, Cindy, Layla, MongChi, Ping, Yenyen, and our anonymous friends and volunteers from Facebook. We also thank the authors of [75] for kindly sharing their collected SMPL parameters with us. UT Austin is supported in part by NSF IIS-1514118.

5. Conclusion

We explored clothing recommendations that complement an individual’s body shape. We identified a novel source of Web photo data containing fashion models of diverse body shapes, and developed a body-aware embedding to capture clothing’s affinity with different bodies. Through quantitative measurements and human judgments, we verified our model’s effectiveness over body-agnostic models, the status quo in the literature. In future work, we plan to incorporate our body-aware embedding to address fashion styling and compatibility tasks.

Acknowledgements: We thank our human subjects: Angel, Chelsea, Cindy, Layla, MongChi, Ping, Yenyen, and our anonymous friends and volunteers from Facebook. We also thank the authors of [75] for kindly sharing their collected SMPL parameters with us. UT Austin is supported in part by NSF IIS-1514118.

Since we do not have these subjects’ vital statistics, we train another version of our model that uses only SMPL and CNN features.
References


Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In ICCV, 2015.


