Learning Visual Body-shape-Aware Embeddings for Fashion Compatibility

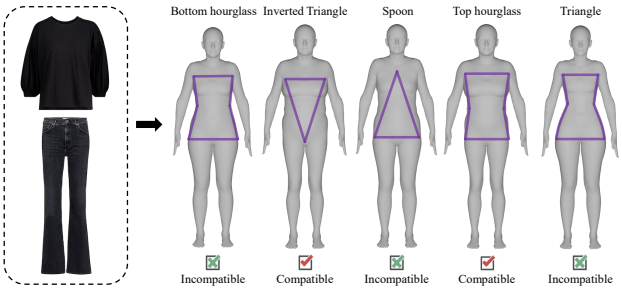
Kaicheng Pang, Xingxing Zou, Waikeung Wong*
{kaicpang.pang, aemika.zou}@connect.polyu.hk, calvinwong@aidlab.hk
School of Fashion and Textiles, The Hong Kong Polytechnic University
Laboratory for Artificial Intelligence in Design
Hong Kong SAR

Abstract

Body shape is a crucial factor in outfit recommendation. Previous studies that directly used body measurement data to investigate the relationship between body shape and outfit have achieved limited performance due to oversimplified body shape representations. This paper proposes a Visual Body-shape-Aware Network (ViBA-Net) to improve the fashion compatibility model's awareness of human body shape through visual-level information. Specifically, ViBA-Net consists of three modules: a body-shape embedding module, which extracts visual and anthropometric features of body shape from a newly introduced large-scale body shape dataset; an outfit embedding module, which learns the outfit representation based on visual features extracted from a try-on image and textual features extracted from fashion attributes; and a joint embedding module, which jointly models the relationship between the representations of body shape and outfit. ViBA-Net is designed to generate attribute-level explanations for the evaluation results based on the computed attention weights. The effectiveness of ViBA-Net is evaluated on two mainstream datasets through qualitative and quantitative analysis. Data and code are released[1].

1. Introduction

Fashion Recommendation Systems (FRSs) [2, 15] is not a new topic, but they still have great potential for economic benefits. Previous works have mainly focused on fashion compatibility learning (FCL) [6, 16, 17], which only considers the compatibility among fashion items. However, besides the outfit itself, consumers will be more concerned about how it looks when worn. Figure 1 demonstrates how fashion compatibility can vary depending on different body shapes. For instance, individuals with an *inverted triangle* body shape may find the outfit in Figure 1 (a) suitable, while those with a *triangle* body shape may not. Previous studies [12–14, 26] represent the body shape merely relying on body measurement data while overlooking the valuable visual features of body shape, which limits their ability to provide precise recommendations. To effectively incorporate accurate body shape information into FRSs, leveraging valuable information from body images is essential. Moreover, accurately representing outfits is also critical, as the scaling and spatial relationships between clothing items can impact how they fit and flatter different body shapes. Therefore, conventional outfit representation methods used in FCL, such as item-wise correlations [4, 31, 32] or graph neural networks [5, 28], are insufficient for modeling the relationships between body shape and an outfit. Lastly, providing a reasonable explanation for the evaluation is essential for personalized FRSs. However, previous studies [12, 21, 22] have not achieved this.



(a). Outfit Composition (b). Human body type

Figure 1. An example of the body-shape-aware fashion compatibility task. The outfit is compatible with the *inverted triangle* and *top hourglass* body shapes, but does not fit other body shapes.

To this end, this paper proposes a Visual Body-shape-Aware Network (ViBA-Net) to model the relationships between body shape and outfit. The ViBA-Net consists of three modules: Body-shape Embedding Module (BEM), Outfit Embedding Module (OEM), and Joint Embedding Module (JEM). The BEM combines visual and anthropometric features to obtain a general representation of the body shape. However, obtaining accurate visual features from body images requires a diverse dataset with explicit body shape annotations, which is currently unavailable.

*Corresponding author.
[1] https://github.com/BenjaminPang/ViBA-Net

Thus, we create a new dataset covering seven common body shapes; each contains 4,000 3D body models with varying but similar shapes. Every model within the dataset is accompanied by corresponding anthropometric data and frontal view images, which offer the visual features of the respective body shape. The OEM learns the outfit embedding by incorporating visual and textual features of the outfit. We propose to represent an outfit leveraging its try-on appearance instead of separate item images because the try-on image contains the scaling and spatial relationships among individual items. For the textual aspect, we exploit the fashion attributes information to enhance the outfit representation, where the attribute values are encoded into word embeddings. Finally, the JEM integrates representations of the body shape and outfit to compute the body-shape-aware embedding, which is then transformed by a linear function to obtain the final compatibility score. The core of the OEM and JEM is a cross-modal attention layer, allowing them to merge features from different modalities. The hierarchical design of ViBA-Net facilitates the propagation of cross-modal interactions between fashion attributes and body shapes through the computed attention maps, as visualized in Figure 6. We leverage these attention maps to generate the attribute-level explanations for the prediction results. All experiments are conducted on two mainstream fashion compatibility datasets, *i.e.*, Outfit for Your (O4U) [22] and Body-Diverse (BD) Dataset [14], that all include body shape annotations. Both qualitative and quantitative results show the advancement of the ViBA-Net. We summarize main contributions as follows:

- We propose ViBA-Net to obtain better body-shape-aware embeddings for fashion compatibility. We enhance the body-shape embedding by introducing visual features extracted from body images and representing the outfit using its try-on appearance.
- We introduce a new dataset with 28,000 body samples covering seven common body shapes, each with a 3D body model, anthropometric data, and a frontal view image. This dataset can also be useful for tasks such as virtual try-on and clothed human generation.
- We conduct experiments on the O4U and BD datasets, demonstrating the superiority of ViBA-Net over other state-of-the-art approaches.

2. Related Work

Body-shape-Aware Fashion Compatibility. With the development of FCL, researchers are increasingly aware of the importance of body shape to practical applications [12–14, 22, 26]. Hidayati *et al.* [13] represented body shapes of female celebrities using their body measurements. Sun *et al.* [29] proposed to use 3D features to represent female upper body shapes. Hsiao *et al.* [14] extracted body shape features using body measurements and SMPL [18] parameters through multiple MLPs. These approaches all neglect the visual features of human bodies. In this work, we propose to encode the body into a more comprehensive embedding incorporating anthropometric and visual features, which are extracted from body images.

Body Shape Classification. Most body-aware methods proposed to classify body shapes using clustering approaches, such as using the k-means in [14] and the affinity propagation in [12, 13]. In [26], authors separate body shapes into two groups according to users' sizes. However, research on classifying body shapes has been extensively investigated over the past two decades. Notably, Simmons [27] developed a well-known body shape classification system, the Female Figure Identification Technique (FFIT), which uses anthropometric data from 3D body scans for body shape classification. Subsequent research [7, 23, 33] improved the FFIT, which has become a widely accepted standard for body shape classification. So, in this work, we introduce a body shape dataset that classifies body shapes into seven well-known types using FFIT instead of clustering methods.

Fashion Outfit Representation. How to represent the outfit plays a crucial role in fashion recommendation. Early works addressing fashion compatibility Learning (FCL) problem [4,31,32] represented an outfit as pairwise relationships between fashion items and mapped fashion item embeddings into a unified space using category information. Beyond pairwise distance, some studies attempted to model high-order interactions among items [19, 22, 28]. These approaches have two limitations: 1. They omit the scaling and spatial relationships between individual clothing items when encoding the outfit; 2. Using only item category information is inadequate because adopting more specific fashion attribute information is useful. To this end, we propose to use try-on appearance images to represent outfits and exploit fashion attributes to enhance the model performance.

3. Body Shape Dataset

Previous studies [13, 23] have introduced a few body shape datasets. However, their number of body models is insufficient to represent body shapes. For example, Parker *et al.* [23] analyzed 1,679 3D body scans, but only 10 and 62 human bodies are categorized as *triangle* and *top hourglass* body shapes, respectively. Although Hidayati *et al.* [13] introduced a dataset consisting of 3,150 individual celebrities with their body measurement, no body shape labels are annotated. In light of this, we present a new dataset for the body shape representation. It features a diverse array of 28,000 individual models, spanning seven prevalent body shapes: *bottom hourglass, inverted triangle, spoon, top hourglass, triangle, hourglass*, and *rectangle*. The construction process involves five steps: 1. Randomly generating 200,000 3D body models using the SMPL method [18];

2. Measuring anthropometric data, including bust, waist, high hip, and hip circumferences from these models; 3. Removing unrealistic models and generating 100,000 more realistic bodies based on refined shape parameters; 4. classifying body shapes using the FFIT algorithm [33]; and 5. Capturing frontal view images for each model using an orthographic camera. The details of constructing this dataset are presented in Section 1 of the Supplementary Material.

4. Methodology

In this section, we elaborate on the details of the proposed ViBA-Net: 1. Clarify the task formulation; 2. Present the representations of body type, try-on image, and fashion attributes; 3. Describe the architecture of ViBA-Net.

4.1. Task Formulation

Following [22], we formulate this task as a multi-label classification task. Given a training set $\mathcal{T} = \{O^j, Y^j\}_{j=1}^N$ containing N outfits, we denote $O^j = \{\mathbf{X}^j, \mathbf{G}^j\}$ as the j-th outfit containing serveral individual clothing images $\mathbf{X}^j$ and structured fashion attributes $\mathbf{G}^j$. $Y^j = \{y_k^j | k = 1, \cdots, K\}$ refers to a set of ground truth labels for j-th outfit conditioned on K body shapes, where $y_k^j = 1$ indicates that outfit O^j is **incompatible** with k-th body shape. Our goal is to devise a learning function $\mathcal{F}$ to predict the compatibility score $\hat{y}_k^j$ between a query outfit O^j and k-th body shape:

$$\hat{y}_k^j = \mathcal{F}(\{\mathbf{X}^j, \mathbf{G}^j, \boldsymbol{\omega}^k, \mathbf{I}^k\} | \boldsymbol{\Theta}) \tag{1}$$

where $\boldsymbol{\omega}^k$ and $\mathbf{I}^k$ are the anthropometric data and front view image of *k-th* body shape. $\boldsymbol{\Theta}$ is the training parameters.

4.2. Body-shape Representation

We devise a Body-shape Embedding Module (BEM) to compute the embedding for the body shape by exploiting both visual and anthropometric features extracted from a representative body model, as illustrated in the top-left corner of Figure 2. To obtain the representative model for the k-th body shape, we first average the shape parameters of all body models belonging to the set $\mathbf{U}^k$, and then use the SMPL model [18] to generate the representative model according to the averaged parameters:

$$\bar{\mathbf{T}}^k = \mathcal{F}_{\text{SMPL}}(\bar{\boldsymbol{\beta}}^k) = \mathcal{F}_{\text{SMPL}}(\frac{1}{|\mathbf{U}^k|} \sum_{\mathbf{T}_i \in \mathbf{U}^k} \boldsymbol{\beta}_i) \tag{2}$$

where $\bar{\mathbf{T}}^k$ is the representative 3D model of k-th body shape, and $\bar{\boldsymbol{\beta}}^k \in \mathbb{R}^{1\times 10}$ is the averaged shape parameter vector. $|\mathbf{U}^k|$ means the size of set $\mathbf{U}^k$. Then, we use an orthographic camera to capture the corresponding frontal view image, denoted as $\bar{\mathbf{I}}^k = \mathcal{F}_{\text{ortho}}(\bar{\mathbf{T}}^k)$. We extract visual features of k-th body shape from $\bar{\mathbf{I}}^k$ by employing a ResNet-18 [11] model, which is trained on the body images of the proposed body shape dataset with a split ratio of 80%, 10%, and 10% for training, validation, and test. $\bar{\mathbf{v}}^k \in \mathbb{R}^{1\times 512}$ is the visual features, and $\mathcal{F}_{\text{body}}$ refers to the forward function of ResNet with the last linear layer discarded. The visual feature extraction process can be written as $\bar{\mathbf{v}}^k = \mathcal{F}_{\text{body}}(\bar{\mathbf{I}}^k)$.

We measure the representative model to acquire the anthropometric data, denoted as $\bar{\boldsymbol{\omega}}^k = \mathcal{F}_{\text{measure}}(\bar{\mathbf{T}}^k) \in \mathbb{R}^{1\times 20}$, where $\mathcal{F}_{\text{measure}}$ refers to the measuring process. Since body shape parameters contain information for characterize the body shape, we concatenate $\bar{\boldsymbol{\beta}}^k$ and $\bar{\boldsymbol{\omega}}^k$, and send it to a linear layer consisting of a linear transformation and a Rectified Linear Unit (ReLU) activation function. The resulting output is concatenated with $\bar{\mathbf{v}}^k$ to produce the body-shape embedding, denoted as $\bar{\mathbf{U}}^k \in \mathbb{R}^{1\times 1024}$. Formally, $\bar{\mathbf{U}}^k$ is calculated using the following equation:

$$\bar{\mathbf{U}}^k = \text{Concat}(\text{ReLU}(\text{Concat}(\bar{\boldsymbol{\beta}}^k, \bar{\boldsymbol{\omega}}^k)\mathbf{W}_B + \mathbf{b}_B), \bar{\mathbf{v}}^k) \tag{3}$$

where $\mathbf{W}_B \in \mathbb{R}^{30\times 512}$ and $\mathbf{b}_B \in \mathbb{R}^{1\times 512}$ are fully connected layer's weight matrix and bias vector, respectively. The resulting body shape features will be sent to the joint embedding module.

4.3. Try-on Image Representation

We leverage try-on images instead of individual clothing images to represent outfits. However, try-on images are not typically included in mainstream datasets for the FCL task, such as *Polyvore* [10], *Style4BodyShape* [13], and *O4U* [22] to name a few. To address this, a Multi-layer Virtual Try-On Network (M-VTON) system is utilized to synthesize separate item images while preserving clothing details as much as possible. Details can be found in Section 2 of the Supplementary Material. After obtaining the try-on image, we utilize a pre-trained ResNet model with its last pooling layer and linear layer discarded to extract its visual features. The motivation behind encoding it into multiple region-level features is that they can provide more accurate representations than a single global feature. Formally, the feature extraction process can be expressed as:

$$\mathbf{S} = \mathcal{F}_{\text{outfit}}(\tilde{\mathbf{X}}) = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}; \tag{4}$$

where $\mathbf{S}$ is the representation of try-on image containing 128 spatial features $\mathbf{x}_i \in \mathbb{R}^{512}$, and $\mathcal{F}_{\text{outfit}}$ refers to the forward function of the modified ResNet.

4.4. Fashion Attributes Representation

The clothing items are associated with a set of fashion attributes manually recognized from various attribute dimensions. For the sake of explanation, we show three fashion attributes in the bottom-left part of Figure 2. We utilize the union of all attributes associated with each item in an

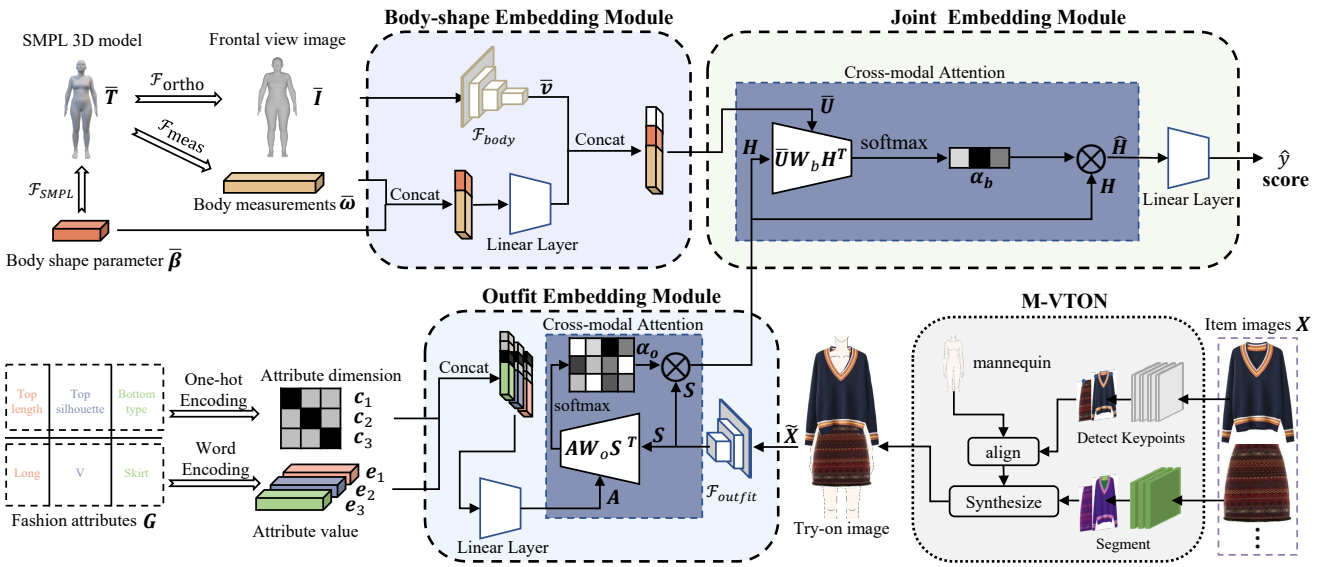


Figure 2. The proposed ViBA-Net consists of three modules. Body-shape Embedding Module represents the body shape using both body image features and anthropometric features. Outfit Embedding Module extracted outfit visual features from the try-on image using M-VTON. Finally, both body shape features and outfit features are sent to the Joint Embedding Module to learn body-shape-aware embeddings. The cross-modal attention mechanism employed in ViBA-Net computes attention weights to generate attribute-level explanations.

outfit to represent the fashion attributes of the entire outfit. For fashion attribute value, we use a pre-trained GloVe [24] model to encode its text into a word embedding, denoted as $\mathbf{e} \in \mathbb{R}^{d_{\text{text}}}$, where $d_{\text{text}} = 300$ is the dimensionality of the word embedding. For fashion attribute dimension, we encode it into a one-hot vector, denoted as $\mathbf{c} \in \mathbb{R}^{N_A}$, where $N_A = 15$ is the number of all fashion attributes used in this work. We then concatenate $\mathbf{c}$ and $\mathbf{e}$ to represent one fashion attribute and then apply a linear transformation to the concatenated vector. Suppose the j-th outfit possesses L^j fashion attributes, this outfit's attribute representation $\mathbf{A}^j \in \mathbb{R}^{L^j \times 512}$ is computed by:

$$\mathbf{A}^j = \{\text{ReLU}(\text{Concat}(\mathbf{c}_l, \mathbf{e}_l)\mathbf{W}_A + \mathbf{b}_A)\}_{l=1}^{L^j} \quad (5)$$

where $\mathbf{W}_A \in \mathbb{R}^{315 \times 512}$ and $\mathbf{b}_B \in \mathbb{R}^{512}$ is the weight matrix and bias vector of the linear transformation.

4.5. Body-type-Aware Network Architecture

We employ the cross-modal attention block [20] in both the Outfit Embedding Module (OEM)and Joint Embedding Module (JEM) of ViBA-Net to merge data representations from different modalities. This mechanism improves conventional attention mechanisms by introducing a learnable weight matrix in the score function, where two modalities are connected by calculating their compatibility scores. Specifically, it takes two inputs denoted as a *query* $\mathbf{Q} \in \mathbb{R}^{N_q \times d_q}$ and a *value* $\mathbf{V} \in \mathbb{R}^{N_v \times d_v}$, and the attention weights $\boldsymbol{\alpha} \in \mathbb{R}^{N_q \times N_v}$ is calculated as:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{Q}\mathbf{W}\mathbf{V}^T) \quad (6)$$

where $\mathbf{W} \in \mathbb{R}^{d_q \times d_v}$ is the learnable weight matrix, and the softmax operation is applied on the second dimension. According to the obtained attention distribution and value $\mathbf{V}$, the output of this block is computed by $\hat{\mathbf{V}} = \boldsymbol{\alpha}\mathbf{V}$, where $\hat{\mathbf{V}} \in \mathbb{R}^{N_q \times d_v}$ is the fused feature vectors. The OEM aims to acquire the outfit representation, denoted as $\mathbf{H}^j \in \mathbb{R}^{L^j \times 512}$, through integrating features of try-on image and fashion attributes using the cross-modal attention block:

$$\mathbf{H}^j = \boldsymbol{\alpha}_o \cdot \mathbf{S}^j = \text{softmax}(\mathbf{A}^j \mathbf{W}_o {\mathbf{S}^j}^T) \cdot \mathbf{S}^j \quad (7)$$

where $\mathbf{W}_o \in \mathbb{R}^{512 \times 512}$ is the learnable weight matrix and $\boldsymbol{\alpha}_o \in \mathbb{R}^{L^j \times 128}$ is the attention maps calculated in OEM. Then JEM learns the relationship between the k-th body shape features $\bar{\mathbf{U}}^k$ and the j-th outfit representation and outputs the compatibility vector between these two:

$$\hat{\mathbf{H}}_k^j = \boldsymbol{\alpha}_b \cdot \mathbf{H}^j = \text{softmax}(\bar{\mathbf{U}}^k \mathbf{W}_b {\mathbf{H}^j}^T) \cdot \mathbf{H}^j \quad (8)$$

where $\hat{\mathbf{H}}_k^j \in \mathbb{R}^{1 \times 512}$ is the body-shape-aware embedding, and $\mathbf{W}_b \in \mathbb{R}^{1024 \times 512}$ is the learnable weight matrix in the JEM. $\boldsymbol{\alpha}_b \in \mathbb{R}^{1 \times L^j}$ is the attention maps computed in JEM. We can observe that the second dimension of $\boldsymbol{\alpha}_b$ is the same as the number of the fashion attributes associated with the j-th outfit. Based on this characteristic of the ViBA-Net, we can obtain corresponding explanations based on the influence distribution of fashion attributes reflected in the attention maps computed in JEM. We visualize $\boldsymbol{\alpha}_b$ in Figure 6 to demonstrate the explainability possessed by ViBA-Net. Lastly, we compute the compatibility score by applying a linear transformation on $\hat{\mathbf{H}}_k^j$:

$$\hat{y}_k^j = \hat{\mathbf{H}}_k^j \cdot \mathbf{W}_s + b_s \quad (9)$$

where $\mathbf{W}_s \in \mathbb{R}^{512\times1}$ and $b_s \in \mathbb{R}$ are the linear transformation's weights and bias, respectively. Since the task is formulated as a multi-label classification task, we use the binary cross entropy loss to measure the difference between predicted scores $\hat{y}_k^j$ and target scores y_k^j.

5. Experiments

We conduct experiments on two fashion compatibility datasets to showcase the benefits of the proposed ViBA-Net model by addressing following research questions:

- **RQ1**: Is the ViBA-Net superior to the current state-of-the-art methods?
- **RQ2**: To what extent do the individual components of ViBA-Net influence the model's performance?
- **RQ3**: What can ViBA-Net generate for explainations?
- **RQ4**: How does ViBA-Net perform in the perceptual study?

5.1. Experimental Settings

Datasets We evaluate the proposed network on two public datasets: Outfit for You (O4U) [22] and Body-Diverse (BD) [14] datasets. **O4U** contains 15,748 compatible outfits and 82,017 clothing items. Each item is associated with a product image and several fashion attributes. On average, the top item contains 6.64 fashion attributes, while the bottom item contains 3.77 attributes. We use the public training, validation, and testing data split provided by O4U to ensure a fair comparison. The **BD dataset** comprises 889 dresses and 971 tops, spanning 57 individual fashion models. We classify these body models into three types (*Bottom hourglass, Hourglass, and Rectangle*) by aligning their body measurements with the models in our body shape dataset. We consider two scenarios for the dataset division: 1). The "easier" case involves seeing models from the test split during the training process denoted as the **Joint** version; 2). In the more "difficult" case, models from the test split are **not** included in the training process, as termed the **Disjoint** version. Please refer to Section 3 of the supplementary material for statistics details of the BD dataset.

Evaluation Metrics. For experiments on the O4U dataset, we employ a set of seven evaluation metrics to compare the performance of different models. This practice aligns with prior works such as [9], [22], and [21], which tackle multi-label classification problems. The metrics encompass Mean Average Precision (mAP), Average Per-Class Precision (CP), Recall (CR), F1 score (CF1), as well as Average Overall Precision (OP), Recall (OR), and F1 score (OF1). Notably, mAP, CF1, and OF1 hold greater significance due to their ability to provide a holistic evaluation of model performance. For experiments on the BD dataset, we evaluate performances using the Area Under Curve (AUC) metric.

Implementation Details. We adopt the SGD optimizer [25] with momentum factor equalling 0.9 and weight decay 5e-4. We gradually decrease the learning rate according to:

$$\mathrm{lr} = \mathrm{base\_lr} \times (1 - \mathrm{step\_num}/\mathrm{max\_step})^{0.9} \tag{10}$$

where the base learning rate is 0.1. The maximum steps and training batch size are set to 1,260 and 10, respectively. During training, we save the checkpoint model corresponding to the highest mAP performance achieved on the validation set and evaluate the saved model on the test set. We report the average evaluation results of five repeated experiments for all experiments.

5.2. Comparative Results (RQ1)

Baselines. We compare the ViBA-Net with seven baseline methods: (1) **StyleMe** [12], which extends AuxStyles [13] by using bidirectional symmetrical deep neural networks to learn a joint representation of outfits and body shapes. (2) **TDRG** [34], an effective multi-object recognition model that explores the structural and semantic aspect relations through Graph Convolutional Network. We use it to learn the joint relation of the try-on image. (3) **M3TR** [35], a multi-modal multi-label recognition model that incorporates global visual context and linguistic information through ternary relationship learning. We embed the body shape labels into the word embedding as the linguistic information and use try-on appearances as input images. (4) **CSRA** [36], which captures spatial regions of objects from different categories by effectively combining a simple spatial attention score with class-specific and class-agnostic features. We train CSRA using try-on images as input. (5) **FCN** [22], which employs a convolutional layer to embed the outfit based on fashion attribute features and utilize a GCN to learn multi-label classifiers based on word embeddings of body shapes. The compatibility scores are obtained by applying the learned classifiers to the outfit embedding. (6) **Mo *et al.*** [21], which learns the correlation between fashion images, fashion attributes, and physical attributes with two transformer encoders. (7) **ViBE** [14], which applies several MLPs to learn fashion clothing's affinity with body measurements. (8) **Body-aware CF** [1], which is a collaborative filtering-based method utilizing the fashion item and body measurement features.

Quantitative Results on O4U. We present the quantitative results on O4U dataset in Table 1. All baseline methods are trained on the training set of O4U. The random method means all predictions are given randomly. We observe that the proposed ViBA-Net achieves the best performances across all metrics. Specifically, it surpasses StyleMe by a clear margin (+14.06 on mAP). This may be because the bidirectional symmetrical deep neural networks utilized in StyleMe are limited in their ability to learn cross-modal relationships. Compared with the TDRG, M3TR, and CSRA

Table 1. Evaluation results on O4U dataset.

Methods	mAP	CP	CR	CF1	OP	OR	OF1
Random	45.01	44.27	23.04	30.31	44.91	21.93	29.47
StyleMe [12]	49.08	37.50	56.05	44.94	62.81	77.70	69.47
TDRG [34]	54.66	50.80	63.60	56.48	65.42	78.85	71.51
M3TR [35]	61.37	55.92	61.19	58.44	69.37	79.65	74.15
CSRA [36]	61.38	56.63	61.18	58.82	71.82	76.79	74.22
FCN [22]	62.34	56.96	62.41	59.55	71.42	78.14	74.62
Mo *et al.* [21]	62.38	55.24	62.10	58.47	67.17	79.34	72.75
ViBE [14]	62.18	55.63	64.43	59.71	70.79	79.25	74.78
ViBA-Net (Ours)	**63.14**	**57.30**	**64.85**	**60.84**	**72.02**	**80.73**	**76.13**

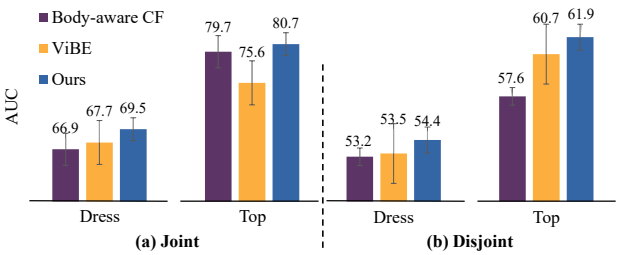


Figure 3. Evaluation results on Body-Diverse dataset. In the joint scenario, test models are seen during the training process In the disjoint version, training and test sets of models are completely separate. Our method notably outperforms other approaches across all scenarios and fashion categories, securing the highest AUC performance by a substantial margin.

methods, the ViBA-Net brings consistent +1.78∼8.5 mAP gains, +2.02∼4.36 CF1 gains, and +1.9∼4.6 OF1 gains over them. The reason may be that the ViBA-Net takes advantage of multi-modal features. ViBA-Net also outperforms the FCN, Mo *et al.* [21], and ViBE methods on all metrics. This may be attributed to the fact that these methods fail to learn body shape embeddings using visual features.

Quantitative Results on Body-Diverse Dataset. We report results on the Body-Diverse dataset in Figure 3. We compare ViBA-Net to the Body-aware CF and ViBE methods. The latter two methods rely solely on SMPL parameters and body measurements for representing body shape. Remarkably, our method consistently outperforms the others across all scenarios and fashion categories. Specifically, Figure 3 (a) shows the results on the "easier" test, and our method brings +2.6 and +1.8 AUC gains over CF and ViBE methods on the dress set, respectively. A substantial AUC improvement of +5.1 over ViBE is also observed on the top set. Figure 3 (b) shows the results on the disjoint test set. AUC performances of ViBA-Net are +0.9 and +1.2 higher than the ViBE method on the dress and top test sets, respectively. These consistent enhancements can be attributed to the incorporation of visual body features.

Notably, it can be observed in Figure 3 that all methods perform better on the joint dataset compared to the disjoint dataset, aligning with our expectations. Another observation is the evaluation results achieved in the top category

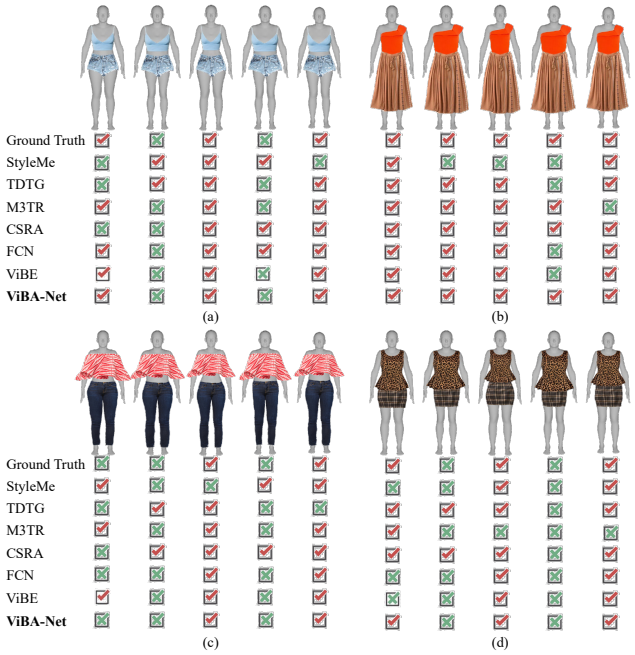


Figure 4. Qualitative comparison among different methods. The tick symbol indicates a match between the outfit and the body shape, while the cross symbol indicates a mismatch.

are superior to those in the dress category. This discrepancy may be because tops are predominantly associated with upper body parameters such as bust and waist rather than hip size. This specificity impacts model performance. In contrast, as full-body garments, dresses leverage data from all body dimensions, contributing to improved performance.

Qualitative Results. The quantitative results are presented in Figure 4. It is evident that among all baselines, the ViBA-Net consistently performs well with various outfit compositions. In Figure 4 (a), for example, the outfit consists of corset straps with hot pants, which might not be compatible with people having lower body segment obesity due to tight pants. However, the length of hot pants is short, exposing the legs, which can alleviate the feeling of envelopment, and thus, the outfit can still be compatible with body shapes such as *bottom hourglass, spoon*, and *triangle*. On the other hand, corset straps are heavy for people with *inverted triangle* or *top hourglass* body shapes, which also have larger breasts. Thus, matching hot pants with the same large exposure of skin is unsuitable. In contrast, as shown in Figure 4 (b), when the clothing is changed to a tank top and A-line long skirt, it can solve both problems. Similarly, in Figure 4 (c), the off-shoulder blouse is unsuitable for people with broad shoulders, and the tight jeans are not compatible with those with lower body segment obesity. Furthermore, for outfits with special silhouettes, such as the peplum top with an H-line short skirt in Figure 4 (d), the ViBA-Net can still accurately assess the compatibility between body shape and the outfit composition.

Table 2. Ablation results on representation learning. *backbone*: utilizing backbone (ResNet-18) as multi-label classifier. *w/o-body*: encoding the body shape into one-hot vector. *w/o-try-on*: encoding outfit using visual features from separate items. *w/o-attr*: removing fashion attribute data.

Methods	mAP	CP	CR	CF1	OP	OR	OF1
backbone	57.71	54.47	57.54	55.96	67.53	76.39	71.68
w/o-body	60.57	55.68	60.71	57.97	67.46	73.84	70.46
w/o-anth.	62.73	56.85	64.25	60.32	71.75	79.91	75.61
w/o-visual	62.61	56.92	64.85	60.63	71.83	80.33	75.84
w/o-try-on	61.72	56.29	62.77	59.35	71.43	78.99	75.02
w/o-attr	61.45	55.83	63.32	59.34	70.61	79.50	74.79
Full model	**63.14**	**57.30**	**64.85**	**60.84**	**72.02**	**80.73**	**76.13**

5.3. Ablation Study (RQ2)

We examine the effectiveness of components in the ViBA-Net by conducting several ablation studies.

Ablation Study on Representation Learning. We first demonstrate the effectiveness of the body shape and outfit representation applied in ViBA-Net, as shown in Table 2. Firstly, we investigate the overall contribution of ViBA-Net to the multi-label classification performance by comparing it with ViBA-Net's *backbone* model (ResNet-18). Our full network brings +5.43 mAP, +4.88 CF1, and +4.45 OF1 performance improvements. Furthermore, we proceed to evaluate the efficacy of our body-shape embedding approach by conducting experiments involving the removal of specific components: anthropometric features (*w/o-anth.*), visual features (*w/o-visual*), and a combination of both (*w/o-body*). Notably, consistent performance deterioration is observed across all three cases. This substantiates that anthropometric and visual features are pivotal in accurately representing body shapes. We also compare our try-on embedding method with a separate item embedding method (*w/o-try-on*). The result shows that ViBA-Net using the try-on embedding achieves higher scores (+1.42 mAP, +1.49 CF1, and +1.11 OF1) than the model using separate items, suggesting that our try-on embedding method captures more information from the try-on image compared with discrete items. Lastly, we investigate the impact of utilizing fashion attributes in our model. Results of *w/o-Attributes* demonstrate that using fashion attribute data can improve the model's overall performance, with the full model achieving increases of +1.69 mAP, +1.50 CF1, and +1.34 OF1, which suggest fashion attributes can provide valuable cues for personalized fashion recommendations.

More ablation studies on network structure and outfit encoding are discussed in the supplementary file Section 4.

Comparing Visual and Anthropometric Features.

Table 3 compares the performance of body shape classification methods. These results show that our visual-based classification approach (Ours) clearly outperforms other baselines. This could be because other baselines merely use anthropometric data to classify body shapes.

Table 3. Body shape classification accuracy comparing with available classifiers.

Available body shape classifiers				Ours
Lee *et al.* [33]	Francis [8]	Collings [3]	Hidayati *et al.* [12]	
28.63%	31.84%	37.87%	76.83%	97.60%

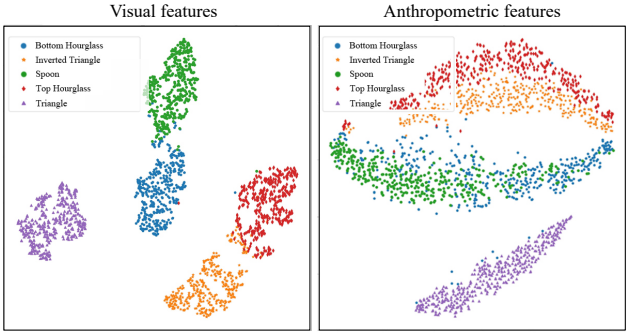


Figure 5. Visualization of different body features using t-SNE.

To further illustrate the difference between the visual and anthropometric features of the body shape, we visualize them in Figure 5 using t-SNE [30]. The visual features are extracted from the frontal view images, and the anthropometric features are measured from 3D models belonging to the testing set of the body shape dataset. We can observe that the five body shapes are separated more clearly from each other in the left part of Figure 5 compared with anthropometric features in right part. This suggests that the visual features contain more valuable information for characterizing the body shape. We also observe that the Euclidean distance between similar body shapes is closer. For instance, the distance between *inverted triangle* (orange star symbol) and *top hourglass* (red diamond symbol) is shorter than the distance between *inverted triangle* and *triangle* (purple triangle symbol). The main reason is that both *inverted triangle* and *top hourglass* body shapes have a wider upper body and a narrower lower body. In contrast, *triangle* body shape typically has larger hips. These results support the proposal that incorporating visual body features into the process of learning body-shape-aware embeddings is effective.

5.4. Explainability Analysis (RQ3)

We visualize the attention maps for three query outfits in Figure 6 to provide a visualization of the fashion attributes that the ViBA-Net focuses on when predicting compatibility. Each row entry of the attention map represents attention weights $\boldsymbol{\alpha}_b$ generated in the JEM, which indicates the significance of fashion attributes with respect to corresponding body shapes. In the first two examples (Figures 6 (a) and (b)), we present two outfits where the first query does not match the *bottom hourglass, spoon,* and *triangle* body shapes, while the second query is compatible with them. The attention maps indicate that ViBA-Net attends mostly to the *bottom silhouette* attribute dimension (last row), *i.e.*,

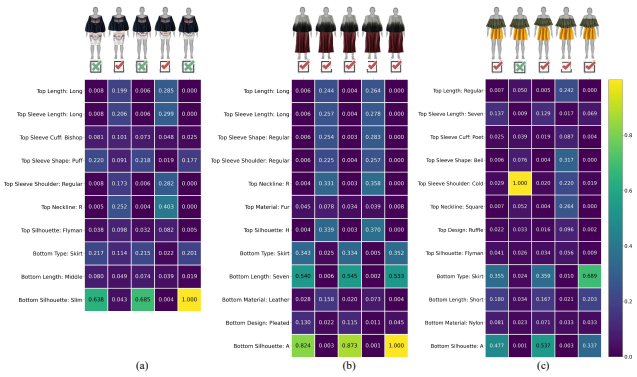


Figure 6. Visualization of attention maps computed in JEM. The vertical axis represents all the fashion attributes possessed by the query outfit. The horizontal axis represents five body shapes, namely, from left to right, *bottom hourglass, inverted triangle, spoon*, and *top hourglass*.

Slim and *A-line*, respectively. This may due to the fact that these three body shapes all possess a larger hip measurement, which is congruent with an *A-line* dress but not with a *slim* one. Additionally, Figures 6 (c) shows an outfit which is incompatible with *inverted triangle* body shape. ViBA-Net suggests that the main reason for this mismatch is the top item contains a *cold shoulder* design. From a fashion perspective, this inference is reasonable because tops with *cold shoulder* designs often fail to provide adequate support for the chest and upper body, which can be a concern for individuals with a larger bust resulting in an unflattering and uncomfortable fit.

Interestingly, the ViBA-Net has varied focuses on fashion attributes belonging to the bottom and top items of different body shapes. The network concentrates mainly on the bottom attributes for body shapes such as *bottom hourglass, spoon*, and *triangle*. Conversely, it pays more attention to the top attributes for the *inverted triangle* and *top hourglass*. This could be because the bottom attributes play a more critical role in determining compatibility for body shapes that tend to have a larger hip and thigh area. On the other hand, for body shapes that have broader shoulders and a smaller waist, the network focuses more on the top attributes to ensure a balanced overall look that accentuates the waistline.

5.5. Perceptual Study (RQ4)

Finally, we conduct a perceptual study to show the potentiality of the ViBA-Net in practical applications. Specifically, we invite ten experts working in the fashion industry to assess the results of all the compatibility models from the following two aspects, (1) Body-shape-Aware Compatibility score (OCs): whether the outfits are compatible with the body shape or not; (2) Explanation Confidence score (ECs): whether the explanation reasonable or not. The score range is $[0, 1]$, 0.1 per level, and the final score is the weighted average of all the scores given by those experts. The perceptual results are summarized in Table 4. It can be seen that the ViBA-Net enjoys the highest performance on Body-shape-Aware fashion compatibility while taking a unique advantage in explainability.

Table 4. Perceptual results of the compatibility models.

Methods	StyleMe [12]	TDRG [34]	M3TR [35]	CSRA [36]	FCN [22]	**ViBA-Net (Ours)**
OCs	49%	52%	51%	53%	59%	61%
ECs	-	-	-	-	-	67%

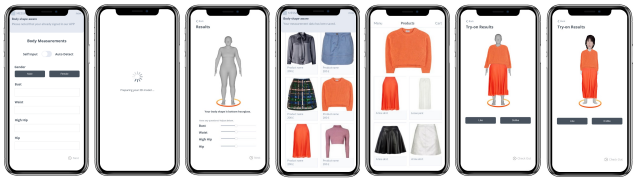


Figure 7. The pipeline of a prototype for applying ViBA-Net in a real application. Step (a): inputting the personal information; step (b): generating a 3D SMPL model according to the input measurements data; step (c): adjusting and confirming the body shape; step (d): browsing the fashion items; step (e): selecting one favour clothing item with corresponding outfit recommendations that consider the body shape; step (f): visualizing the outfit composition on the size of body shape; step (g): translating the SMPL model into a human image via generative model e.g., Midjourney.

In addition to the perceptual study, we also build the prototype for applying ViBA-Net in a real application to show the practicality of the proposed method. As shown in Figure 7, we present the main steps of the prototype for applying ViBA-Net in real applications. It can be seen that, with the awareness of body shape, customers can more easily and directly accept the recommended outfits. And connecting with the current cutting-edge techniques can generate more user-friendly and interesting results with huge economic potential, e.g., translating the SMPL model into a human image via generative models such as Midjourney or executing a call API of Large Language models such as ChatGPT to make the explanation more like a natural conversation.

6. Conclusion

Body shape is an essential consideration when recommending outfits to consumers in real-life applications. To this end, we propose ViBA-Net to learn better body-shape-aware embeddings for fashion compatibility and a new dataset containing varied information about body shape. Meanwhile, we also propose representing the outfit using its try-on appearance, which captures the scaling and spatial relationships between fashion items on the body. We conduct experiments on both the O4U and BD dataset to demonstrate the superiority of ViBA-Net compared to other state-of-the-art approaches.

Acknowledgements

This work is supported by Laboratory for Artificial Intelligence in Design (Project Code: RP 3-2) under InnoHK Research Clusters, Hong Kong SAR Government.

References

[1] Tianqi Chen, Weinan Zhang, Qiuxia Lu, Kailong Chen, Zhao Zheng, and Yong Yu. Svdfeature: a toolkit for feature-based collaborative filtering. *The Journal of Machine Learning Research*, 13(1):3619–3622, 2012. 5

[2] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2662–2670, 2019. 1

[3] Kat Collings. The foolproof way to find out your real body type. 7

[4] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12617–12626, 2019. 1, 2

[5] Zeyu Cui, Zekun Li, Shu Wu, Xiao-Yu Zhang, and Liang Wang. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *The world wide web conference*, pages 307–317, 2019. 1

[6] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. Disentangling features for fashion recommendation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1s):1–21, 2023. 1

[7] Priya Devarajan and Cynthia L Istook. Validation of female figure identification technique (ffit) for apparel software. *Journal of Textile and Apparel, Technology and Management*, 4(1):1–23, 2004. 2

[8] Cherene Francis. Body shape calculator. 7

[9] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing*, 30:5920–5932, 2021. 5

[10] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] Shintami Chusnul Hidayati, Ting Wei Goh, Ji-Sheng Gary Chan, Cheng-Chun Hsu, John See, Lai-Kuan Wong, Kai-Lung Hua, Yu Tsao, and Wen-Huang Cheng. Dress with style: Learning style from joint deep embedding of clothing styles and body shapes. *IEEE Transactions on Multimedia*, 23:365–377, 2020. 1, 2, 5, 6, 7, 8

[13] Shintami Chusnul Hidayati, Cheng-Chun Hsu, Yu-Ting Chang, Kai-Lung Hua, Jianlong Fu, and Wen-Huang Cheng. What dress fits me best? fashion recommendation on the clothing style for personal body shape. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 438–446, 2018. 1, 2, 3, 5

[14] Wei-Lin Hsiao and Kristen Grauman. Vibe: Dressing for diverse body shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5, 6

[15] Hyunwoo Hwangbo, Yang Sok Kim, and Kyung Jin Cha. Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, 28:94–101, 2018. 1

[16] Pang Kaicheng, Zou Xingxing, and Wai Keung Wong. Modeling fashion compatibility with explanation by using bidirectional lstm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3894–3898, June 2021. 1

[17] Yen-Liang Lin, Son Tran, and Larry S Davis. Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3311–3319, 2020. 1

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[19] Zhi Lu, Yang Hu, Yan Chen, and Bing Zeng. Personalized outfit recommendation with learnable anchors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12722–12731, 2021. 2

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 4

[21] Dongmei Mo, Xingxing Zou, Kaicheng Pang, and Wai Keung Wong. Towards private stylists via personalized compatibility learning. *Expert Systems with Applications*, 219:119632, 2023. 1, 5, 6

[22] Kaicheng Pang, Xingxing Zou, and Waikeung Wong. Dress well via fashion cognitive learning. In *British Machine Vision Conference (BMVC)*, November 2022. 1, 2, 3, 5, 6, 8

[23] Christopher J Parker, Steven George Hayes, Kathryn Brownbridge, and Simeon Gill. Assessing the female figure identification technique's reliability as a body shape classification system. *Ergonomics*, 64(8):1035–1051, 2021. 2

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[25] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 5

[26] Hosnieh Sattar, Gerard Pons-Moll, and Mario Fritz. Fashion is taking shape: Understanding clothing preference based on body shape from online sources. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 968–977. IEEE, 2019. 1, 2

[27] Karla Kristin Peavy Simmons. *Body shape analysis using three-dimensional body scanning technology*. North Carolina State University, 2002. 2

[28] Tianyu Su, Xuemeng Song, Na Zheng, Weili Guan, Yan Li, and Liqiang Nie. Complementary factorization towards outfit compatibility modeling. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4073–4081, 2021. 1, 2

[29] Jie Sun, Qianyun Cai, Tao Li, Lei Du, and Fengyuan Zou. Body shape classification and block optimization based on space vector length. *International Journal of Clothing Science and Technology*, 31(1):115–129, 2019. 2

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[31] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision (ECCV)*, pages 390–405, 2018. 1, 2

[32] Xuewen Yang, Dongliang Xie, Xin Wang, Jiangbo Yuan, Wanying Ding, and Pengyun Yan. Learning tuple compatibility for conditional outfit recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2636–2644, 2020. 1, 2

[33] Jeong Yim Lee, Cynthia L Istook, Yun Ja Nam, and Sun Mi Park. Comparison of body shape between usa and korean women. *International Journal of Clothing Science and Technology*, 19(5):374–391, 2007. 2, 3, 7

[34] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 163–172, 2021. 5, 6, 8

[35] Jiawei Zhao, Yifan Zhao, and Jia Li. M3tr: Multi-modal multi-label recognition with transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 469–477, 2021. 5, 6, 8

[36] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021. 5, 6, 8