Supplementary Material: Learning 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

This supplementary material contains:

- Description of build the Body Shape Dataset
- Description of the Multi-layer Try-on Network system
- Detailed statistics of the training and test data splits of the Body-Diverse dataset.
- Ablation Studies on network structure and outfit encoding

1. Body Shape Dataset Construction

We first give the definitions of body shape. Let $\mathbf{\Omega} = \{\mathbf{T}_i\}_{i=1}^{N_T}$ be a set of human body models, where $\mathbf{T}_i$ represent a 3D body model. A body shape set $\mathbf{U}^k \subseteq \mathbf{\Omega}$ is defined as a subset of $\mathbf{\Omega}$ whose models share similar characteristics and can be categorized into the same body shape, where $k \in [1, K]$ and K is the number of examined body shapes. The detailed construction process includes the following steps:

Step 1: Generating SMPL Model. We generate 200,000 3D body models with diverse body sizes to ensure dataset variety by employing the Skinned Multi-Person Linear (SMPL) model [3]. SMPL is a learned model that accurately represents various human body sizes in different poses. The body model is generated according to shape parameters $\boldsymbol{\beta}$ and pose parameters $\boldsymbol{\theta}$. However, as we focus on variations in human body shape, we keep the pose parameters constant while generating body models. Thus, a one-to-one correspondence exists between the body model set and the shape parameter set. Let $\mathcal{F}_{\text{SMPL}}$ be the SMPL [3] model forward function, we denote the body model as $\mathbf{T}_i = \mathcal{F}_{\text{SMPL}}(\boldsymbol{\beta}_i)$.

Step 2. Measuring Anthropometric Data. We employ a body measurement tool [5] to acquire the anthropometric data containing 20 dimensions from the generated 3D model. Among these 20 measures, the *bust*, *waist*, *high hip* and *hip* circumferences are most important because we employ FFIT [7] to identify the body shape based on these four measures. In Figure 1, we visualize these circumferences, where the circle's size indicates the circumference's length. These circumferences are measured by locating body landmarks based on the regularities of cross-sectional body shapes. We modify the tool's localization criteria to maintain consistency with the body landmarks defined in FFIT. The obtained anthropometric data is denoted as $\boldsymbol{\omega} = \mathcal{F}_{\text{measure}}(\mathbf{T})$, where $\mathcal{F}_{\text{measure}}$ is the measuring process.

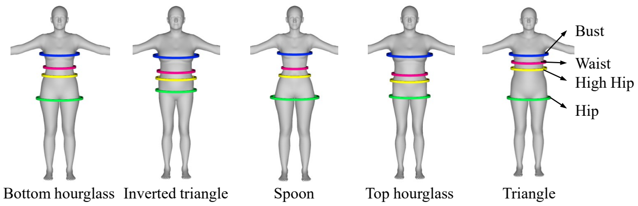

Figure 1. Body model examples of the proposed body shape dataset. Both *Bottom hourglass* and *top hourglass* have a well-defined waistline, but their difference lies in their hip-to-bust ratio; *triangle* and *inverted triangle* lack a well-defined waistline because they do not consider the bust-to-waist ratio; *Spoon* is characterized by a large gap between hip and bust circumference and a smaller bust-to-waist ratio than the *hourglass*.

Step 3. Cleaning Invalid Model. To ensure the generated body models are realistic, we eliminate invalid models that fall outside the standard range of human height-weight distribution [2, 4]. Consequently, 11.57% of the generated body models are retained. Then, we calculate the mean and variance of the remaining models' shape parameters $\boldsymbol{\beta}$ and use these distributions to generate a new set of 100,000 body models. This process improves the realism of the newly generated bodies, as the height and weight are more closely aligned with the normal distribution of humans.

Step 4. Annotating Body Shape. We employ the FFIT algorithm [7] to determine the body shape of each SMPL

*Corresponding author.

model. However, the classification results show that the distribution of body models across different body shapes is non-uniform. For instance, out of the 100,000 body models, only a small portion are identified as the *top hourglass* and *triangle*, with only 120 and 11 models, respectively. To address this imbalance issue, for each body shape, we compute its specific shape parameter distribution, which is then used to regenerate body models. This method effectively enhances the occurrence frequencies of these underrepresented body shapes by optimizing the shape parameters. Finally, we randomly select 4,000 valid human body models for each body shape to form the dataset. We display five body shapes in Figure 1 to visually illustrate the differences between different body shapes.

Step 5: Capturing Frontal View Image. It is worth noting that we capture the frontal view image for each body model. This is accomplished by rendering it in a virtual environment using an orthographic camera. The resultant image has a resolution of 1024×512 pixels and is saved in PNG format. It is imperative to underscore that these frontal view images hold paramount importance, as they serve as the foundation for extracting visual-level representations of human body shapes. We use the notation $\mathcal{F}_{\text{ortho}}$ to represent the orthographic projection process, and the resulting image is denoted as $\mathbf{I} = \mathcal{F}_{\text{ortho}}(\mathbf{T})$.

2. Multi-layer Try-on Network System

The M-VTON system comprises three stages: fashion key point detection and alignment, fashion segmentation, and try-on image synthesis.

Fashion Key Point Detection and Alignment. A trained detection model is first employed to identify the fashion-oriented key points from clothing images. Its backbone is a pre-trained pose estimation model [6], and we fine-tune it on the fashion key point dataset. It outputs the pixel locations of fashion key points as indicated by blue dots in Figure 2. By aligning items' key points with the corresponding key points of the mannequin, we can calculate each item's scaling ratio and pixel locations. These intermediate results will guide the generation of the try-on images in the synthesis stage.

Fashion Segmentation. A fashion segmentation model is also utilized to segment an item image into its front and back pieces. This process is critical for generating a realistic try-on appearance image. We employ the Object-Contextual Representations model [8] as the backbone and train it on a new fashion segmentation dataset with pixel-level annotations. As shown in Figure 2, the purple and green areas of the segmentation results stand for the front and back pieces, respectively. They are used in the synthesis stage.

Try-on Synthesis. Based on the previous results, we synthesize rescaled clothing pieces in a predefined wearing order and place them at corresponding pixel locations to generate the try-on image, which has a resolution of 1040×680 pixels. We use the symbol $\tilde{\mathbf{X}}$ to denote the generated try-on image.

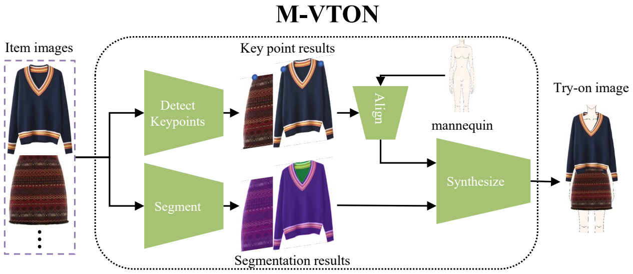

Figure 2. Workflow of M-VTON system comprising three stages: fashion key point detection and alignment, fashion segmentation, and try-on image synthesis.

Table 1. Statistics of the Joint Diverse-Body Dataset [1].

		Dress			Top		
	type	Bott.	Hour.	Rect.	Bott.	Hour.	Rect.
Train	body	18	23	10	18	28	11
	clothing	538	444	217	556	530	320
Test	body	4	5	2	4	6	3
	clothing	123	108	51	138	145	89

Table 2. Statistics of the Disjoint Diverse-Body Dataset [1].

		Dress			Top		
	type	Bott.	Hour.	Rect.	Bott.	Hour.	Rect.
Train	body	14	18	8	14	22	8
	clothing	423	323	95	396	423	308
Test	body	4	5	2	4	6	3
	clothing	50	58	28	51	98	23

3. Statistics of Body-Diverse Dataset

The details of the training-test partition for both the joint and disjoint versions are outlined in Table 1 and Table 2, respectively. We randomly split the item set and model set into training and test sets in a ratio of 8:2. Specifically, for dress (top) dataset, the training set comprises 711 (776) examples, while the test set encompasses 178 (195) examples. Following [1], we implement the propagation scheme to identify missing positive pairs among similar body shapes. For clarity, after applying the propagation process, it is exemplified that a total of 538 (423) dresses are compatible with 18 (14) *bottom hourglass* fashion models. This results in 173 (288) negative dresses for the joint (disjoint) dataset. As can be observed, the reduction in the number of trained models in the disjoint dataset has led to a corresponding decrease in the quantity of positive items.

Table 3. Comparison on variations of cross-modal attention.

Structure	mAP	CP	CR	CF1	OP	OR	OF1
dot-product	50.02	39.36	42.70	40.96	65.56	58.89	62.05
multi-layer	52.30	52.63	25.47	34.33	52.63	28.36	36.86
multi-head	58.71	54.27	60.24	57.10	68.05	79.65	73.39
cross-modal	**63.14**	**57.30**	**64.85**	**60.84**	**72.02**	**80.73**	**76.13**

Table 4. Comparison of encoding outfits *w/o-try-on*. The bold numbers indicate a larger value.

Method	Outfit Encoding	mAP	CP	CR	CF1	OP	OR	OF1
TDRG [9]	separate	49.97	37.54	56.61	45.14	62.78	78.44	69.74
	try-on	**54.66**	**50.80**	**63.60**	**56.48**	**65.42**	**78.85**	**71.51**
M3TR [10]	separate	53.90	52.99	57.07	54.95	64.34	78.13	70.57
	try-on	**61.37**	**55.92**	**61.19**	**58.44**	**69.37**	**79.65**	**74.15**
CSRA [11]	separate	57.38	55.92	54.59	55.24	69.31	72.81	71.02
	try-on	**61.38**	**56.63**	**61.18**	**58.82**	**71.82**	**76.79**	**74.22**

4. Ablation Studies on Network Structure and Outfit Encoding

Ablation Study on Network Structure. We test three variations of the cross-modal attention mechanism, and report the quantitative results in Table 3. Specifically, we replace the cross-modal attention with the *dot-product* attention, *i.e.*, the weight W is removed from Equation 7 in the main paper. The performance of the model is observed to decrease due to this operation. A possible reason may be attributed to the difference in input modalities, as the attention module in ViBA-Net receives inputs from different modalities, which is unsuitable for simple dot-product attention. We further present the results of adopting *multi-layer* and *multi-head* of cross-modal attention. As shown in the Table. However, they fail to achieve better results.

Comparing Outfit Encoding On Baselines. We compare the model performance of three baselines which are trained using try-on images and separate clothing images in Table 4. From the table, we can observe a consistent improvement in performance across all methods when using the try-on image to represent the outfit. Notably, M3TR achieves a +7.47, 3.49, and 3.58 improvements on mAP, CF1, and OF1 metrics, respectively. This is reasonable to infer that owing to the try-on image capturing the interdependent relationships between clothing items, which allows M3TR to learn the underlying contextual relationships better.

References

[1] 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. 2

[2] Aline Jelenkovic, Reijo Sund, Yoon-Mi Hur, Yoshie Yokoyama, Jacob v B Hjelmborg, Sören Möller, Chika Honda, Patrik KE Magnusson, Nancy L Pedersen, Syuichi Ooki, et al. Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. *Scientific reports*, 6(1):28496, 2016. 1

[3] 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. 1

[4] Wayne J Millar. Distribution of body weight and height: comparison of estimates based on self-reported and observed measures. *Journal of Epidemiology & Community Health*, 40(4):319–323, 1986. 1

[5] Akash Sengupta. 3d body measurements. 1

[6] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16072–16081, 2021. 2

[7] 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. 1

[8] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 2

[9] 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. 3

[10] 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. 3

[11] 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. 3