Fashion-Specific Ambiguous Expression Interpretation with Partial Visual-Semantic Embedding

Ryotaro Shimizu
Waseda University, ZOZO Research

Takuma Nakamura
ZOZO Research

Masayuki Goto
Waseda University

Abstract

A novel technology named fashion intelligence system has been proposed to quantify ambiguous expressions unique to fashion, such as "casual," "adult-casual," and "office-casual," and to support users' understanding of fashion. However, the existing visual-semantic embedding (VSE) model, which is the basis of its system, does not support situations in which images are composed of multiple parts such as hair, tops, pants, skirts, and shoes. We propose partial VSE, which enables sensitive learning for each part of the fashion outfits. This enables five types of practical functionalities, particularly image-retrieval tasks in which changes are made only to the specified parts and image-reordering tasks that focus on the specified parts by the single model. Based on both the multiple unique qualitative and quantitative evaluation experiments, we show the effectiveness of the proposed model.

1. Introduction

When browsing fashion items online, users must interpret fashion images to resolve difficult questions that arise in their minds, without the shopkeeper's support. The questions are as follows: 1) "what would this outfit look like if it were more casual?" 2) "how office-casual is this outfit?" and 3) "what makes this outfit street?" It is difficult to answer these questions even for experts and particularly difficult for non-experts. This ambiguity may hinder the users from pursuing their deep interest in fashion, making it difficult for them to try new genres of clothing.

In response to this expectation, Shimizu et al. [26] proposed a "Fashion Intelligence System" to support the interpretation of these terms through various applications that can be provided by applying a visual-semantic embedding (VSE) model. This system helps with clarifying the relationships between the full-body outfit images and various expressions, including ambiguous expressions specific to the fashion domain. In addition, by enabling users to obtain the answers to their ambiguous and complex questions

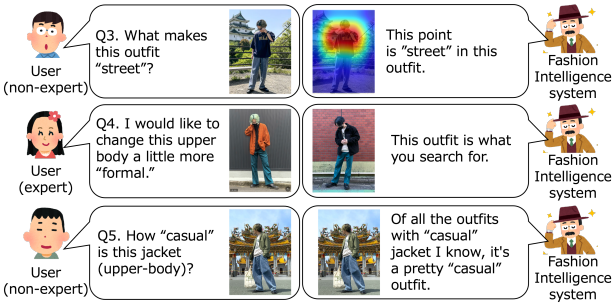

Figure 1. Image of fashion intelligence system

(in particular, the abovementioned questions 1–3), it is possible to reduce the ambiguity inherent in fashion and support the user in all fashion-related decisions such as what to wear and what items to purchase. However, the VSE in [26] has a simple model structure that maps the full-body outfit image to the projective space as a batch. The problem caused by this limitation is that the model cannot answer the questions which users truly want to be answered such as: 4) "what would the outfit look like if I changed the upper-body to make it a little more kawaii?" and 5) "how chic is this jacket?" Ideally, image-retrieval results should not show images in which changes are made to the entire body [14] because users consider how to dress based on the outfits they already have.

In this study, we propose a partial VSE (PVSE) model that enables the acquisition of an embedded representation corresponding to each part in a full-body outfit image while maintaining a simple model structure and a low computational complexity. The proposed model retains various previous practical application functions (answering questions 1–3) and enables image-retrieval tasks in which changes are made only to specified parts (answering question 4) and image reordering tasks, attentively focusing on the specified parts (answering question 5).

The main contributions of this study are as follows: 1) We propose a PVSE model that can map a full-body outfit image and rich tags into the same projective space and ob-

tain an embedded representation corresponding to each part included in the outfit. 2) Multifaceted unique evaluation experiments show that the proposed model contributes to more accurate mapping operations. 3) We show that the proposed model not only retains three practical application functions to support the user's fashion interpretation but also expands two tasks that attentively focus on a specific part despite its simple structure. The contribution of this study opens up a novel research field, and it is expected that more complex and various models to interpret fashion-specific ambiguous expressions will be proposed.

2. Related Research

VSE In the fashion domain, VSE has been used for text-based and individual clothing image retrieval [4, 31, 36] and learning outfit compatibility (individual outfit item matching) [11, 33]. Furthermore, the VSE included in [10] is a method for mapping fashion item images (not full-body outfit images) and specific words (not ambiguous words) contained in the item descriptions in the same projective space.

A VSE model that allows the mapping of full-body outfit images and ambiguous rich attributes was proposed as a starting point for research on fashion intelligence systems that support the interpretation of fashion-specific expressions [25, 26]. However, these models can only capture the entire atmosphere of a full-body outfit because of not including a mechanism to learn a full-body outfit separately.

Part-by-Part Learning There are studies that independently learn fashion images of each item included in a full-body outfit and recommend items based on the compatibility between different types of items [17, 29, 32, 40], studies that derive which combination of candidate (inside a wardrobe) items match [7, 13], and studies that search for other items that match a query item [3, 8, 20, 24]. While these tasks pertain to which combinations of items are fashionable, the interests of our study include: "what happens if this outfit becomes more casual?" and "how casual is this outfit?" Thus, the focus of this study is fundamentally different. Furthermore, the fact that each item is independently applied to a backbone model is also a major difference in our study, which focuses on learning full-body outfit images directly.

Studies in the field of person re-identification have used segmentation-based methods to extract the features for each part from a single full-body image [9, 21]. However, these studies are clearly different from our study in that they include an operation to mix the features of each part because there is no need to correspond the order of the parts to each dimension of the final required features. Furthermore, another method extracts a bounding box for each body part from a single image by detection and performs learning for each part [38, 39]. However, this detection-based (patch-based) method is less sensitive than the segmentation-based method, and the computational complexity necessary for applying each part to a large network remains a challenge.

Conversely, several approaches to fashion image generation have been studied based on shape features obtained by segmentation and capturing features for each part [6, 14, 16, 37]. Particularly, Hsiao et al. [14] performed excellent work based on the claim that "making minor changes to fashion is important to become fashionable." However, these approaches are based on image generation (i.e., items that do not strictly exist are generated) and do not focus on ambiguous expressions.

3. Methodology

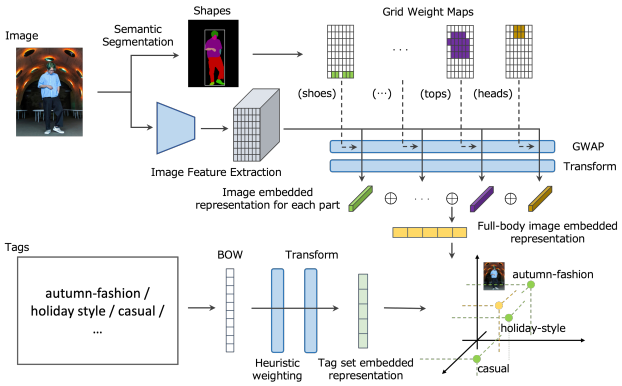

Figure 2. Structure of a prototype of our model proposal

Model Architecture The key feature of the proposed model is the inclusion of an architecture that considers a full-body outfit image as a collection of fashion items (parts) and acquires the embedded representation corresponding to each part. A simple but effective architecture is obtained by extending foreground-centered learning [26] and combining the grid weight map corresponding to each part obtained from the semantic segmentation model and the embedded image feature by global weighted average pooling [22]. Consequently, regardless of the number of parts that the full-body outfit is divided into, the number of times that the backbone model is applied to each image per epoch can be limited to only one time, thus avoiding an increase in computational complexity.

Part-by-Part Grid Weight Map Acquisition The segmentation included in this study calculates the probability of the fashion item appearing in each pixel. The part with the highest probability for each pixel is considered a part of that pixel. Here, the grid refers to the area in which the target image is divided vertically into I and horizontally into J. The grid weight map for the l-th part is defined as $G_l = \{g_{(1,1),l}, ..., g_{(i,j),l}, ..., g_{(I,J),l}\}$ when the number of all parts is defined as L. $N_{(i,j),l}$ is defined as the count of the l-th part of the pixels contained in the (i,j)-th grid and $g_{(i,j),l} = N_{(i,j),l} / \sum_{i=1}^{I} \sum_{j=1}^{J} N_{(i,j),l}$. In other words,

$g_{(i,j),l}$ is the percentage of pixels in the (i,j)-th grid out of all the pixels in the l-th part.

Parameter Optimization The dataset used in this study consists of a single full-body outfit image to which multiple tags are assigned. First, the image is embedded using the image features obtained from the backbone model and a grid weight map G_l. Furthermore, the embedded representation of the tag set assigned to an image is heuristically weighted to generate an embedded representation considering the bias in the frequency with which each tag is assigned to the entire dataset [26].

By optimizing the N-pair angular loss [35], which balances N-pair [28] and batch angular [35] losses, between the abovementioned image and tag set embedded representations, the full-body outfit image, and the attached tags are mapped into the same projective space. Here, N-pair angular loss is adopted based on the preliminary experiments comparing with triplet [34], N-pair, single angular [35], and batch angular losses. Consequently, the image and tag embedded representations in which each dimension is corresponding to each part are obtained.

4. Experimental Evaluation

Settings The total number of full-body outfit images in the experimental data accumulated in the fashion coordination posting application WEAR [41] was 15,740, and the number of unique tags attached to all the images was 1104. Additionally, all the participants reflected in the target images were female. The dimensions of the embedded representation were set as 128. Furthermore, we used Inception v3 [30] pre-trained on ImageNet [5] based on preliminary experiments comparing with several backbone models [12, 27] including ViT [18] and BEiT [1]. Moreover, the public pre-trained model [19] was used for segmentation.

Quantitative Evaluation: Each Component's Position The validity of the obtained embedded representation is verified by whether the image and tag that should be nearby are mapped together closely. For example, an image with a "casual" tag is quantitatively evaluated based on the idea that it should be mapped near a "casual" tag. Precision and normalized documented cumulative gain (NDCG) [15] were used as the accuracy measures (P@M and N@M, respectively). We tested the eight methods listed in Table 1 as a comparison model. Here, H in transformer VSE (TVSE)-H represents the number of heads in the MHA layer. Furthermore, to check the change in the accuracy of the proposed model relying on the division of the parts, we tested the three PVSE-L models. Furthermore, the experiment was repeated 30 times, and a t-test (significance level: 5%) was performed between the results of the proposed models and the model with the highest accuracy among the comparison models.

Table 1. Summary of model-type evaluation values (experiment 1)

	P@5	P@10	P@15	N@5	N@10	N@15
Random	0.091	0.085	0.086	0.085	0.085	0.086
VSE [10]	0.462	0.435	0.418	0.424	0.426	0.421
VSE+ [26]	0.483	0.451	0.428	0.441	0.443	0.432
GVSE [23]	0.276	0.276	0.278	0.250	0.262	0.268
DGVSE [25]	0.244	0.244	0.244	0.220	0.231	0.235
TVSE-4 [2]	0.195	0.199	0.198	0.176	0.188	0.191
TVSE-8 [2]	0.212	0.209	0.210	0.194	0.201	0.204
TVSE-16 [2]	0.197	0.198	0.201	0.176	0.186	0.191
PVSE-4	**0.833****	**0.771****	**0.714****	**0.760****	**0.759****	**0.733****
PVSE-8	0.799**	0.738**	0.694**	0.728**	0.726**	0.707**
PVSE-16	0.801**	0.738**	0.690**	0.732**	0.729**	0.706**

From Table 1, the proposed models are more accurate than the comparison models that include conventional VSE models, regardless of how the parts are divided. It can be concluded that the proposed model is more effective and sensitive compared to the conventional models, in which all the dimensions have semantic representations of all parts. Additionally, the rule of PVSE-4 that divides the parts into {head, upper-body, lower-body, shoes} shows the best accuracy. Therefore, it is important to set an appropriate number of parts and divide the information obtained from the full-body outfit image to the extent that important information is not lost in the mapping process.

Quantitative Evaluation: Attention to Appropriate Part It is possible to determine which regions in an image and tag are highly relevant by calculating the relevance score for each grid and tag. Here, we check whether the score is high for an appropriate region. Specifically, when the relevance scores are calculated between tags such as "t-shirt," "jeans," and "sneakers," and the images to which these tags are attached, the relevance scores should be higher for the regions containing "upper-body," "lower-body," and "shoes," respectively. Based on this idea, we compared the relevance scores calculated between a specific tag and each image attached to the tag, and the top five grids were obtained. True labels of the grids were based on the results of segmentation. This original evaluation method is unique to fashion data in checking the representations' quality in depth.

Table 2. Summary of model type evaluation values (experiment 2)

	Head		Upper-body		Lower-body		Shoes	
	P@5	N@5	P@5	N@5	P@5	N@5	P@5	N@5
Random	0.161	0.145	0.547	0.494	0.346	0.306	0.096	0.087
VSE	0.518	0.472	0.624	0.557	0.874	0.790	0.095	0.081
VSE(f)	0.690	0.627	0.850	0.770	0.854	0.774	0.134	0.121
GVSE	0.510	0.461	0.805	0.722	0.885	0.800	0.260	0.246
DGVSE	0.532	0.482	**0.874**	**0.786**	0.869	0.789	0.109	0.110
TVSE-4	0.030	0.022	0.340	0.294	0.160	0.137	0.078	0.071
TVSE-8	0.059	0.049	0.408	0.357	0.182	0.161	0.103	0.091
TVSE-16	0.055	0.048	0.435	0.382	0.194	0.169	0.114	0.102
PVSE-4	0.668	0.618	0.746	0.665	0.988	0.893	0.546	0.532
PVSE-8	0.643	0.589	0.611	0.558	**0.991**	**0.895**	**0.565**	**0.540**
PVSE-16	**0.779**	**0.742**	0.764	0.691	0.983	0.888	0.480	0.463

The results illustrated in Table 2 show that the proposed

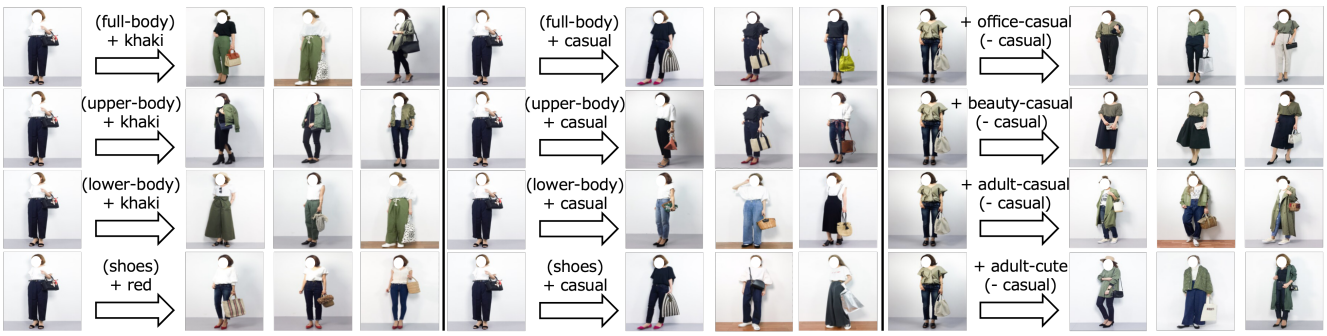

Figure 3. Example of image retrieval

Figure 4. Example of image reordering

model has better accuracy for most of the indices compared to the comparison model, including the conventional VSE models; only for the upper-body, the DGVSE indices are higher, while all the other parts have lower accuracy. However, the proposed model is universally accurate in all cases. This is the result of learning each part delicately, suggesting that the proposed model can learn in a way that satisfies the objective of this study.

Qualitative Evaluation: Image Retrieval & Reordering Examples of image retrieval obtained by image and tag operations are shown in Figure 3. First, we checked the validity of the results by observing search results with specific (color) tags (shown in the left column). Furthermore, it is possible to grasp the dressing method that makes a query attire casual using ambiguous tags (shown in the center column). For example, to change the full-body outfit, we can make the upper-body black and add a colorful item to make it more casual. Furthermore, to make a minor change to the lower-body and make the overall atmosphere casual, a change from navy skinny to jeans or loose skirts can be made. This retrieval application with specifying a part allows users to obtain the answer to the complex question 4, "what would the outfit look like if I changed the upper-body to make it a little more casual?"

The results of image reordering obtained by image and tag operations are shown in Figure 4. First, we checked the validity of the results by observing the sorting results by specific (color) tags. Additionally, it is possible to sort by ambiguous tags. For example, "beauty-casual" clothes with a thin silhouette are more "beauty-casual" than those with a loose silhouette. This image reordering application with specifying a part allows the user to obtain the answer to the complex question 5, "how beauty-casual is the upper-body of this outfit (this jacket)?" Furthermore, it can be understood that typical clothes for a wedding are dresses and those besides dresses are unusual. If a user wants to go with a unique outfit, it is preferable to choose an outfit with a low relevance score with the "wedding-party" tag; if user wants to go with a typical outfit, it is preferable to choose an outfit with a high relevance score. In this way, it is possible to rearrange images by specifying attention parts to meet the detailed needs of the user, to discover the typical full-body outfits indicated by ambiguous tags, and to discover suitable clothing for the situation.

5. Conclusions

In this study, we proposed a PVSE model that can obtain an embedded representation for each of the multiple parts included in the full-body outfit image and attached rich tags with little increase in the computational complexity. Although applications with attribute activation maps were not stated because of the space limitation, the proposed model retained the three functions feasible in the previous VSE models (questions 1-3) and added two new functions (questions 4-5). Additionally, we confirmed the effectiveness of the proposed model through multiple unique evaluation experiments. While the basic model of the current fashion intelligence system is still a simple structure, the contribution of this study opens up a novel research field, and it is expected that more complex and various models to interpret fashion-specific ambiguous expressions will be proposed.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 3

[2] Muhammet Bastan, Arnau Ramisa, and Mehmet Tek. T-VSE: Transformer-based visual semantic embedding. In *CVPR 2020 Workshop on Computer Vision for Fashion, Art, and Design*, 2020. 3

[3] 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 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2662–2670, 2019. 2

[4] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 136–152, 2020. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 3

[6] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8117–8125, 2020. 2

[7] Xue Dong, Xuemeng Song, Fuli Feng, Peiguang Jing, Xin-Shun Xu, and Liqiang Nie. Personalized capsule wardrobe creation with garment and user modeling. In *Proceedings of the ACM International Conference on Multimedia*, pages 302–310, 2019. 2

[8] Zunlei Feng, Zhenyun Yu, Yongcheng Jing, Sai Wu, Mingli Song, Yezhou Yang, and Junxiao Jiang. Interpretable partitioned embedding for intelligent multi-item fashion outfit composition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(2s), 2019. 2

[9] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3641–3650, 2019. 2

[10] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1472–1480, 2017. 2, 3

[11] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the ACM International Conference on Multimedia*, pages 1078–1086, 2017. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[13] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7161–7170, 2018. 2

[14] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5046–5055, 2019. 1, 2

[15] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000. 3

[16] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. ClothFormer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10799–10808, 2022. 2

[17] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian McAuley. Complete the look: Scene-based complementary product recommendation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10524–10533, 2019. 2

[18] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[19] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self Correction for Human Parsing. Retrieved from https://github.com/GoGoDuck912/Self-Correction-Human-Parsing. Accessed October 30, 2022, 2019. 3

[20] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, 2017. 2

[21] Yaoyu Li, Hantao Yao, Tianzhu Zhang, and Changsheng Xu. Part-based structured representation learning for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(4), 2020. 2

[22] Suo Qiu. Global weighted average pooling bridges pixel-level localization and image-level classification. *CoRR*, abs/1809.08264, 2018. 2

[23] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 207–211, 2016. 3

[24] Yuki Saito, Takuma Nakamura, Hirotaka Hachiya, and Kenji Fukumizu. Exchangeable deep neural networks for set-to-set matching and learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 626–646, 2020. 2

[25] Ryotaro Shimizu, Masanari Kimura, and Masayuki Goto. Fashion-specific attributes interpretation via dual

gaussian visual-semantic embedding. *arXiv preprint arxiv:2210.17417*, 2022. 2, 3

[26] Ryotaro Shimizu, Yuki Saito, Megumi Matsutani, and Masayuki Goto. Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags. *Expert Systems with Applications*, 213:119167, 2023. 1, 2, 3

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 3

[28] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1857–1865, 2016. 3

[29] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. Neural compatibility modeling with attentive knowledge distillation. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 5–14, 2018. 2

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 3

[31] Ivona Tautkute, Tomasz Trzcinski, Aleksander P. Skorupa, Lukasz Brocki, and Krzysztof Marasek. DeepStyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019. 2

[32] 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 405–421, 2018. 2

[33] Jianfeng Wang, Xiaochun Cheng, Ruomei Wang, and Shaohui Liu. Learning outfit compatibility with graph attention network and visual-semantic embedding. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 2

[34] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1386–1393, 2014. 3

[35] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2612–2620, 2017. 3

[36] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317, 2021. 2

[37] Han Xintong, Wu Zuxuan, Wu Zhe, Yu Ruichi, and S. Davis Larry. VITON: An image-based virtual try-on network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018. 2

[38] Xishan Zhang, Jia Jia, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. Trip outfits advisor: Location-oriented clothing recommendation. *IEEE Transactions on Multimedia*, 19(11):2533–2544, 2017. 2

[39] Zhichen Zhao, Huimin Ma, and Shaodi You. Single image action recognition using semantic body part actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3411–3419, 2017. 2

[40] Xingxing Zou, Kaicheng Pang, Wen Zhang, and Waikeung Wong. How good is aesthetic ability of a fashion model? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21200–21209, 2022. 2

[41] ZOZO, Inc. WEAR. Retrieved from https://wear.jp/. Accessed October 30, 2022, 2022. 3