FashionAI: A Hierarchical Dataset for Fashion Understanding

Xingxing Zou[‡*] Xiangheng Kong[†] Waikeung Wong[‡†] Congde Wang[†]
Yuguang Liu[†] Yang Cao[†]
[‡]The Hong Kong Polytechnic University, HongKong SAR
[†]Alibaba Group, Hangzhou, China
aemika.zou@connect.polyu.hk
{yongheng.kxh, yingxian, yuguang.lyg, yinming.cy}@alibaba-inc.com
calvin.wong@polyu.edu.hk

Abstract

Fine-grained attribute recognition is critical for fashion understanding, yet is missing in existing professional and comprehensive fashion datasets. In this paper, we present a large scale attribute dataset with manual annotation in high quality. To this end, complex fashion knowledge is disassembled into mutually exclusive concepts and form a hierarchical structure to describe the cognitive process. Such well-structured knowledge is reflected by dataset in terms of its clear definition and precise annotation. The problems which are common in the process of annotation, including structured noise, occlusion, uncertain problems, and attribute inconsistency, are well addressed instead of merely discarding those bad data. Further, we propose an iterative process of building a dataset with practical usefulness. With 24 key points, 245 labels that cover 6 categories of women's clothing, and a total of 41 subcategories, the creation of our dataset drew upon a large amount of crowd staff engagement. Extensive experiments quantitatively and qualitatively demonstrate its effectiveness.

1. Introduction

The fashion industry has attracted many attentions with its huge economic potential and practical value. Many pieces of research in this field have recently progressed from recognition-based clothing retrieval tasks[18, 25, 16, 26, 1, 34] to understanding-based tasks[14, 29, 6, 12, 5, 8, 11, 32], while the latter ones mean that model can not only recognize the attributes of fashion items but can further understand the meaning or expression of the combination of those attributes. Outfit recommendation[5, 8, 11], for example, is a kind of fashion understanding task that requires the model to learn the compatibility of fashion items. The fashion semantic of those items are consist of design attribute. Fashion compatibility learning is to learn the matching relation of a series of design attribute in fact. In view of this, systematical and comprehensive design attribute recognition is the foundation of fashion understanding tasks.

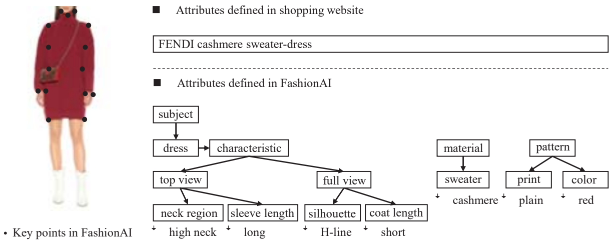


Figure 1. The difference of the attributes and key points between FashionAI and other datasets. For simplicity, we omit more attribute dimensions and values in the tree structure.

However, existing fashion attribute datasets[7, 26, 4, 18, 16] *etc*. designed for fashion retrieval task is not suitable enough for the desired understanding task because of the following limitations (noted that we use DeepFashion [26] as examples below as it is the mainstream fashion recognition dataset with the largest scale images and most diverse attributes currently).

Confusion of concept: Fashion semantic is composed by the expression behind each design attribute of an apparel, *e.g*. PeterPan collar refers to lovely. Thus, it is hard to understand high level semantically (*e.g*. style recognition) if the concept of attribute at the lower level (*e.g*. shirt cuff) is not independent. Meanwhile, as shown in Figure 1, the apparel is described as cashmere sweater-dress on the shopping website. Such mixed concepts, like "Sweater-dress", "Sweater", and "dress", together as the same classification

[*] Most of this work was performed at Alibaba Group, the first two authors equally contribute

[†] Corresponding author.

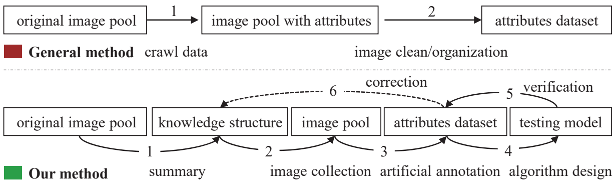


Figure 2. Comparison of the available process of building a dataset with ours

target would bring confusion into trained model.

Incompletion of attribute: To better understanding fashion, complete knowledge at attribute recognition stage is important. A problem that unclear concepts would easily bring is the missing basic design attributes. Particularly, we find 49 attributes in the region of collar over 216 attributes of Part type in DeepFashion [26]. Even the number of attributes looks comprehensive enough, however, it still lacks many common attributes such as rib collar. There are many duplicated concepts (*e.g*. mock, mock neck, and mock-neck) and some attributes among the rest attributes has tiny difference (*e.g*. print v-neck, fitted v-neck).

Mistake of annotation: Data-driven technology is based on data (image and label) thus accuracy of annotation is dominant in the performance of supervised training model. Annotation accuracy of the existing attribute datasets still has certain space for improvement. For instance we randomly choose “A-line” of Shape type in Deepfashion[26]. There are 59 out of 1,000 attributes belongs to “A-line” and 3,301 images in total. However, only 2,569 images (77.8%) of them are correctly labelled.

In light of this, we present FashionAI dataset with both attributes and key points for fashion understanding tasks. Specifically, we address above limitations by conducting the domain knowledge of fashion. The complex knowledge is disassembled mutually exclusive and reconstructed into a hierarchical structure. For the annotation accuracy, we give each attribute a clear definition. Meanwhile, experts not only train our crowd staffs for annotation but also spot check the data to ensure its quality.

Meanwhile, for practical usefulness, all images are intensionally sampled from hundreds of billions of clothing data in various seasons and categories to ensure the diversity of data. Also, as shown in Figure 2, our dataset is constructed in an iterative process, where the subsequent results would have impact on the previous steps. A dynamic correction process is therefore executed. Below are the details of establishment of a FashionAI dataset:

Step 1: With the assistant of fashion experts, knowledge of necessary attributes about apparel is established.

Step 2: According to the defined hierarchical structure, each attribute is utilized to collect corresponding images from online websites.

Step 3: The collected data will be annotated in line with pre-determined standards and regulations. Experts with fashion knowledge will check the annotated images ensuring a high quality of our dataset.

Step 4: An algorithm is designed to generate a model to avoid structure noise from having any effect on the trained model, and to verify the effect of annotated data.

Step 5: Images from real application are added in each iteration process to ensure that the trained model can obtain the consistent performance on both the builded dataset and the real applications.

Step 6: The knowledge structure is revised accordingly.

Our contributions are **three-fold:** (1) We organize the complex and huge fashion knowledge into a logical tree-structure and prove its advance compare with the unclear concepts and single layer structure. (2) We propose a new iterative framework to build a dataset with practical usefulness which attempt to offer a reference for building a professional recognition dataset in any other field with practical values. (3) We launch a large scale dataset with 357k images in high quality for fashion understanding. The diversity of collected data is ensured and practical usefulness is taken into consideration. Meanwhile, all attributes and key points are annotated under the supervisor of clear definitions and the annotation accuracy is higher than 95% by spot check of experts. The common problems in the process of annotation are well addressed to ensure its practical usefulness in the real e-commercial scenarios.

2. Related Work

Clothing Parsing Dataset. There are many researches focusing on clothing parsing[36, 35, 23, 33, 38]. Yamaguchi *et al*. presented Fashionista Dataset with 685 fully parsed images for clothing parsing task[36]. Its ground truth gave a total of 56 clothing labels covering 53 different clothing items such as boots, jacket, and jeans *et al*. Then, they further expanded the Fashionista dataset to form the Paper doll dataset[35]. Color, clothing item, or occasion were further taken into consideration for style retrieval. Additionally, Liu *et al*. proposed the Colorful-Fashion dataset (CFPD) consisting of 2,682 images annotated with pixel-level color-category labels [23]. Yang *et al*. constructed CCP with 2098 high-resolution fashion images[38]. Unlike these datasets constructed for clothing parsing task, our dataset is designed for fashion understanding.

Fashion Analysis Dataset. As mentioned before, there are many researches focus on fashion understanding task recently. For outfit recommendation[22, 30, 11], for example, Han *et al*. presented Polyvore Dataset with 21,889 outfits. The corresponding descriptions, *e.g*. off-white rose-embroidered sweatshirt, of those outfits were adopted as in-

Table 1. Comparison between FashionAI and the existing datasets for fashion attributes recognition

	WBID [18]	DDAD [4]	DARND [16]	DeepFashion [26]	FashionGen [28]	**FashionAI (ours)**
#images	78,958	341,201	453,983	800,000	293,008	357,000
#categories	11	15	20	50	169	6
#dimensions	4	2	8	6	-	68
#attribute values	62	67	179	1,050	169	245
#key points	-	-	-	294[10]	-	24
hierarchical	-	-	-	-	-	yes

put knowledge[11]. In terms of style analysis[27, 21, 31, 2, 19], the published datasets were focusing on different style analysis, *e.g.* five styles including hipster, bohemian, goth, preppy, and pinup in[19]. Additionally, for apparel generation[12, 39], the proposed datasets were designed for generating new fashion items. Unlike these works, we construct FashionAI to fine-grained recognize fashion items.

Fine-grained Recognition Dataset. In the context of fashion recognition, there were many useful datasets that are already published for academic use. Mainstream datasets for fashion attributes recognition are summarized in Table 1. To our knowledge, the source of building current fashion datasets[25, 9, 24, 20, 17, 37, 3, 28, 26] was all collected from the websites, and the original attributes and attribute system were used as knowledge structure directly (except a small-scale dataset named CCP[38] which consists of 2,098 fashion images).

Deepfashion obtained by 800,000 images with 1,050 attributes has been one of the most popular dataset for fashion related researches[26]. However, since it was designed for fashion retrieval, the attributes defined in Deepfashion were marketing-orient which was not systematical and comprehensive enough for fashion understanding task. Recently, Rostamzadeh *et al.* introduced Fashion-Gen with 293,008 fashion images parried with totally 169 fashion categories[28] for text-to-image and attributes-to-image synthesis task. The attributes were described in text format. In contract with these datasets, FashionAI is built from the perspective of design for fashion understanding task. The design regions (defined as attribute dimensions, *e.g.* sleeve length, sleeve cuff, or collar design *etc.*) and their belonging designs (defined as attribute values, *e.g.* cap sleeves) are summarized in a hieratical structure.

3. FashionAI Dataset

We introduce FashionAI, a high quality fashion dataset, to the academic society. It covers 6 categories of women's clothing, a total of 41 sub-categories on the website and has diverse data including different seasons (*e.g.* winter, summer), views (*e.g.* front, side), and types (*e.g.* products images, street images). The attribute system of FashionAI that includes design attributes and key points are presented in Figure 3.

3.1. FashionAI Structure

From the perspective of fashion design, we contribute the knowledge structure with a top-down mechanism. This structure with logical internal connections can satisfy the requirement of fashion profession and machine learning simultaneously. As shown in Figure 3, the complex fashion semantic is dissembled into mutually exclusive design attributes.

Professional knowledge: All women's wear items are divided into six categories, "blouse", "pants", "skirt", "dress", "jumpsuit", and "outwear", all located at the top of whole framework. Design attributes of apparel are divided into three parts, namely characteristic, material and pattern, on the first level. Characteristic refers to the design feature of apparel. Material is a kind of fabric used to make up the apparel items, such as "cotton". Finally, pattern is referred to color and graphic design. Meanwhile, the key points of each subject are defined from the perspective of garment making.

Hierarchical structure: As shown in Figure 3, instead of a single-layer one which used previously[7, 18, 26], FashionAI is hierarchical. All roots and leaf nodes in the attribute tree are named as attribute dimensions and attribute values, respectively. The total number of annotations are not the sum of all attribute values, but the product of the number of attribute value in each attribute dimension. For example, as shown in Figure 4, the attribute dimension of "sleeve style" is further divided into 4 sub-dimensions, including: shape, cuff, shoulder, and design. However, even there are just 5 attribute dimensions with 23 attribute values, 960 different designs of sleeves could still be presented. It is obvious that this hierarchical structure reduces the trained attributes but could also improve the comprehensiveness of the dataset at same time.

Mutually-exclusive attribute: From the perspective of fashion design, there are many attributes would appear in the same region of a garment but belongs to different cat-

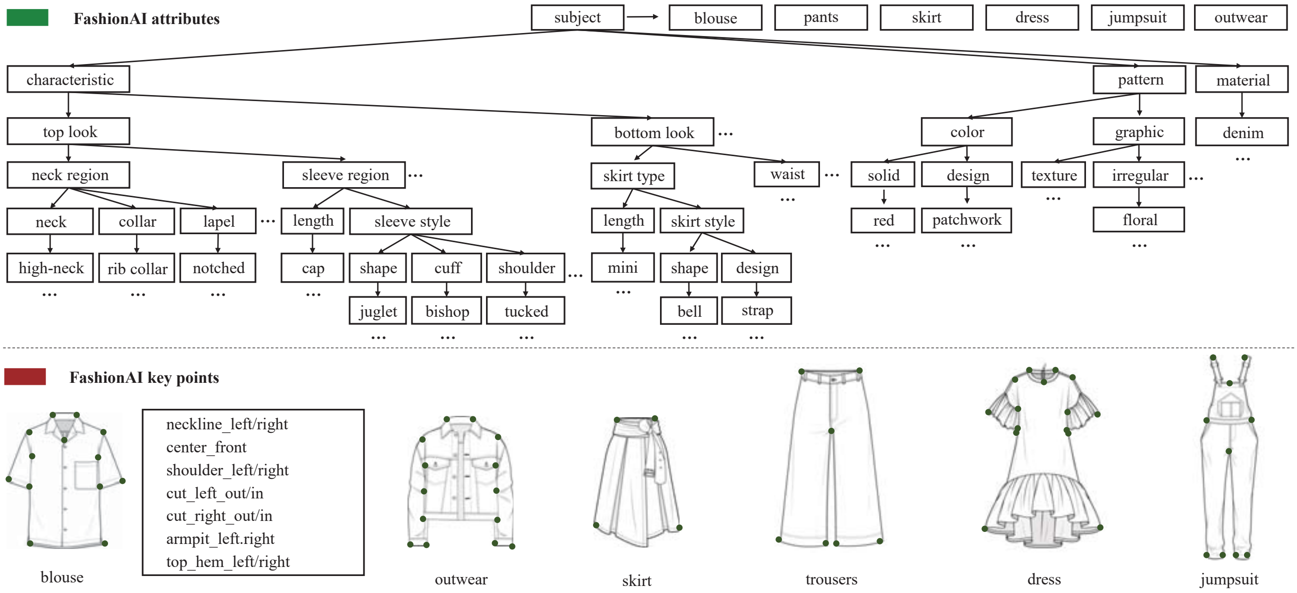


Figure 3. Part of the FashionAI attribute system for demonstration

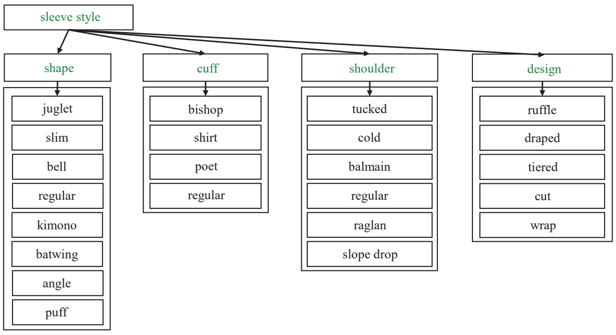


Figure 4. Demonstration the advantage of the FashionAI structure in terms of comprehensive

egories, *e.g*. V-neck and PeterPan Collar. Thus, to distinguish the overlap concepts, the definition of each attribute dimension and attribute value are clear and mutually exclusive. To realize it, fashion knowledge is decomposed clearly to ensure that they are machine-learnable. For example, the attribute dimension of "neck" is divided into four parts: "high neck", "neckline", "collar", and "lapel". Theoretically, such definition ensures that attribute values generated from each attribute dimension can exist simultaneously. The experiment results show the advance of hierarchical structure compared with the single layer one.

In the end, we obtain 24 different key points as well as 245 attribute values in 68 attribute dimensions (noted that 201 values belongs to the dimension of characteristic that covers almost all general designs of daily garments).

4. Data Preparation

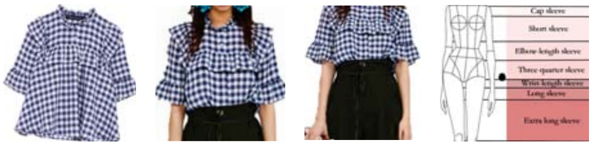
As shown in Figure 2, unlike the general method used to build fashion dataset, FashionAI Dataset is constructed in an iterative process. All images are collected from commercial website. With the back up from fashion experts, we define each attribute clearly and professionally. The standard of artificial annotation, which includes textual description and image examples, is also developed. However, since the collected data are online product images, which are complex and devise, we still face many problems, *e.g.* structure noise, attributes inconsistence *etc*.

4.1. Image collection

According to the designed standard, all related images, which we called image pool, are directionally collected online according to the key words of attributes. Two main problems have been solved in this step: scarcity collection and structure noise.

Scarcity collection: The existing common methods, like keywords searching, can retrieve most required images. However, there are still many attributes which seldom appear in the public websites. For those kind of attributes, we use model acquisition method to search similar images based on the collected small-scale image set at first. And then, those similar images are artificial checked until we collect enough images of the attributes.

Structure noise: To ensure the objectivity of constructed dataset, except the descriptive words of attribute value, we avoid using other directional keywords. For example, "balmain sleeves" are usually used in "suits" design. However,



(a) tiled (b) model (c) "invisible" (d) "uncertain"

Figure 5. Demonstrate examples in FashionAI Dataset

"suits" is not used as a keyword to search data of "balmain sleeves". Additionally, to make sure that the model recognizes "balmain sleeves" not just based on whether the apparel is "suits" or not, we train a model based on the collected data to examine its effectiveness. Based on the testing result on the overall random sampling data, the standard has been revised accordingly.

Additionally, as common in web data, the raw data from website contains certain amount of near-duplicate images. Thus, before the annotation step, the duplicate data are automatically removed.

4.2. Artificial Annotation

When the image pool is ready, we conduct artificial annotation. Since the standard contains plenty of professional knowledge, we assign parts of tasks to the crowdsourcing staffs at first and revised the standard based on their feedback and the result of labeling tasks. When the accuracy rates reaches 95%, labeling tasks are fully open to the outsourcing staffs. Meanwhile, 20% data of labeling tasks of each attribute value are checked and the accuracy with higher than 97% is regarded as qualified labeling. However, the data from image pool are complex and diverse since all of them are uploaded by different online sellers. Those images without a uniform standard cause many problems at the annotation step.

Attribute inconsistent problem: As shown in Figure 5(a) and 5(b), the tiled single apparel image and single model image are common in e-commerce fashion data. The length related attributes dimension can be easily recognized when it has a model as a reference. However, for tiled single apparel image, no model can be used for referencing. To solve this problem, we use key points of the armpits and the distance between two armpits as reference. According to the proportion of apparel, the length standard of the tiled single apparel image is defined. The established standard can be verified if same results can also be obtained by comparing the standard used in the front view of a single model image. We have tested 510 paired images, and the accuracy rate can achieve 95%.

Occlusion problem: Occlusion problem is very common in real commercial apparel images, especially the photo is uploaded by the sellers or users online. It brings troubles in attribute annotation, noticeably in the length-related attribute dimension. As shown in Figure 5(c), the length of a pair of trousers is blocked. It is impossible to determine its length just based on the single image. To solve this problem, a new attribute named "invisible" is added to label such kind of situation. Thus, the recognition result of the image in Figure 5(c) can be more reasonable and enables the trained model to have "rejection" ability instead of giving an unpredictable answer.

Uncertain problem: This kind of problem usually occurs on length related attribute dimensions. As shown in Figure 5(d), if the sleeve length of an apparel is on the position in black spot, it could be recognized as wrist-length sleeves as well as three-quarter sleeves. Thus, an annotation trick named "uncertain" is created to solve this problem instead of avoiding such uncertain images. Taking attribute dimension of sleeves length as an example, it has totally 9 attribute values, including "invisible", "sleeveless", "cap sleeves", "short sleeves", "elbow-length sleeves", "three-quarter sleeves", "wrist-length sleeves", "long sleeves", "extra-long sleeves". Therefore, the attribute dimension(length of sleeves) of Figure 5(d) is annotated as "nnnnnymnn", which means three-quarter sleeves ("y"), and can be regarded as wrist-length sleeves ("m"). Noticeably, "m" would not be punished during training.

4.3. Algorithm Design

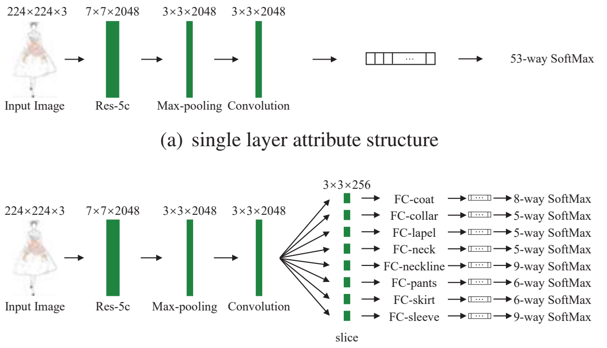
To verify the usefulness and effectiveness of our method, we propose an AttributeNet, which simultaneously predict attributes in a hierarchical and end-to-end manner. The network structure of AttributeNet is similar to that of the residual-50 network[13] shown in Figure 6(a).

However, AttributeNet, as shown in Figure 6(b), is connected to a pooling layer (3x3 stride2) Pool5 and a convolution layer (1x1 stride1) ConvNew after the res5c layer. Then, the feature map slice of ConvNew is divided into eight equal parts. Each part is used to represent an attribute dimension. The corresponding attribute dimension layer is followed by a fully connected layer and a softmax loss layer for classification. Noticeably, the AttributeNet is built on our FashionAI Dataset embedded with professional fashion knowledge. In other words, without the professional definition of fine-grained apparel semantics such as length of top, sleeves length, etc., there would be no bifurcated parallel prediction structure of AttributeNet.

The advantages of the AttributeNet include: (1) It can predict all attributes of apparel in parallel with high efficiency; (2) Such multi-classification method greatly avoids overfitting of features, which enables the trained model to obtain good generalization ability.

4.4. Model Iteration

As usual, we adopt data mining strategy to collect image for annotation. However, this method will inevitably



(a) single layer attribute structure

(b) attributeNet consisting of eight SoftMax

Figure 6. Pipeline of the proposed AttributeNet and baseline net

inject bias in terms of image distribution. For example, if we built a dataset in spring, it is unavoidable that most of our collected data are spring products. The attributes more related to autumn and winter apparel like "sweater" will be influenced. Moreover, the collected images will naturally relate to the online sellers' preference since the information about the products are written by themselves. As such, the trained model based on those images cannot perform well in real applications. Specifically, when we train a model on the attribute of lapel, the average accuracy on testing set can achieve up to 82% but only 36% on real application dataset. The classification accuracy of shawl lapel attribute is less than 10% due to its scarcity. Even though we collect enough data at multiple times, it eventually brings more bias as well. To solve the above-mentioned problems and to ensure that the trained model not only achieves good performance on testing set but also gives satisfactory results on online products, we propose the concept of model iteration.

5. Experiments and Results

5.1. Dataset validation

As depicted earlier, FashionAI is created in a different perspective from previous datasets. Thus it is hard to find a fair evaluation criterion for doing comparison with the existing attribute datasets directly, we demonstrate the effectiveness of the proposed dataset in the folowing two aspects: the knowledge structure and the practical uses.

Knowledge structure. To demonstrate the advance of the hierarchical structure compared with the single layer one, we firstly test the performance of those two different structures both in the FashionAI dataset. As described in Section 4.3, a 54-way residual-50 network is used to benchmark with the performance of the AttributeNet, as shown in Figure 6(b). We replace the fc level of the last 1000-way output of the original res-50 net with 54-way outputs. To ensure the fairness and comparability of experiment, we use the same configuration for training. Specifically, the size of each image is adjusted so that its shorter edge is 224 pixels and then the middle 224 pixels are cropped from the longer edge. There are totally 78,379 images in the training set and 1,194 images for validation. Then, we test on 10,800 images and the accuracy results are shown in Figure 7. It can be seen that ArributeNet has consistently better performance in terms of accuracy across all eight attribute dimensions. The higher accuracy indicates that the proposed AttributeNet can provide the positive sample of each attribute dimension with higher confidence level and greatly avoid over fitting of features, which proves the advantage of adopting hierarchical structure.

Further, we do the recognition task which using the data from DeepFashion[26] and FashionAI. The attributes including "sleeveless", "V-neck", and "shirt-collar" which belong to both two datasets are randomly picked. Noted that those attributes in DeepFashion all have the same definitions as FashionAI and their accuracy also has the comparability with ours. The source of data are collected from both two datasets, namely DF-sub and FAI-sub respective-

Algorithm 1: Model Iteration

```
Input: Training set: $train\_set$ Validation set: $val\_set$
Output: Testing set: $app\_set$
while do
    Training model $M$ on $train\_set$;
    Testing model $M$ on $val\_set$, the accuracy noted as $P(val\_set)$;
    Testing model $M$ on $app\_set$, the accuracy noted as $P(app\_set)$;
    if $|P(val\_set) - P(app\_set)| < 5\%$ then
        end while;
    else
        Testing model $M$ on $app\_set$, annotate samples with low confidence;
        Re-add the new annotated samples into dataset $A$, update $train\_set$ and $val\_set$;
        Continue;
    end
end
```

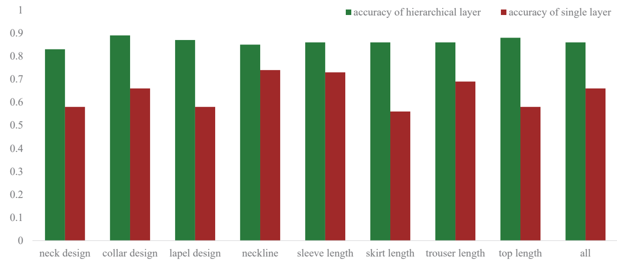


Figure 7. Accuracy comparison between the hierarchical structure and the single layer structure

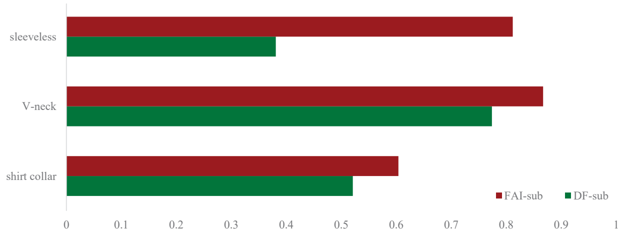


Figure 8. Accuracy of attributes recognition between DF-sub and FAI-sub

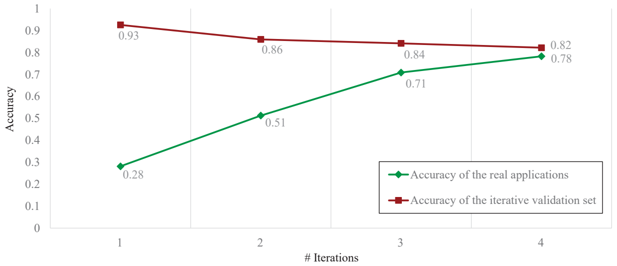


Figure 9. Accuracy of "PeterPan Collar" in test set and the online applications

ly. The number of images for sleeveless, v-neck, and shirt collar in FAI-sub are 3,576, 4,207, and 1,455 respectively, which is as same as the DF-sub for the sake of fairness. Particularly, the source of the test set with total 2,522 images is half from DF-sub and half from FAI-sub. We adopt the DenseNet161[15] for training and the recognition accuracy of those three attributes is shown in Figure 8. The model trained on FAI-sub has consistently better performance in those three attributes than the one using DF-sub for training. As discussed before, the main reason is that DF-sub exists many mixed attribute with confusion concepts while the attributes designed in FashionAI are all mutually exclusive with clear definition.

Practical uses. As described before, the structure noise and bias are unavoidable when building a dataset. Thus, an iteration process is proposed to weaken their influence and further improve the performance of the data-trained model in the most real-world applications. Considering the cost of artificial annotations, here we just taking "PeterPan Collar" as an example to verify the effectiveness of the presented process. We collect 1,000 images from real applications and 15% images from the validation set to conduct the iterative experiments. Figure 9 shows that the gap between the accuracy of the real applications and validation set becomes narrower with increase number of iterations. In other words, with the increasing number of iterations, the model obtains better performance in the online products. The number of iteration is decided by the required accuracy in practice.

In addition, we introduce a new application, which is *outfit composition generation*, depends on the FashionAI dataset. Outfit recommendation is a trendy tool for retailers to cater the consumer's pursuit of beauty and the do the cross-selling. However, currently, the outfit composition is mostly manually generated. Based on the FashionAI dataset, the key points and position of a fashion item in each subject could be recognized. As shown in Figure 10, the armpit key points of apparel in different scales are used at first. The distance of the two armpits is taken as the reference for the unity of scale. Then, the useful key points for attaching apparel are adopted. Note that the used key points in each subject are different, *e.g.* we use waist key points for skirt. Finally, the key points of the top and the bottom are merged to generate the outfit composition. We present some samples in Figure 10.

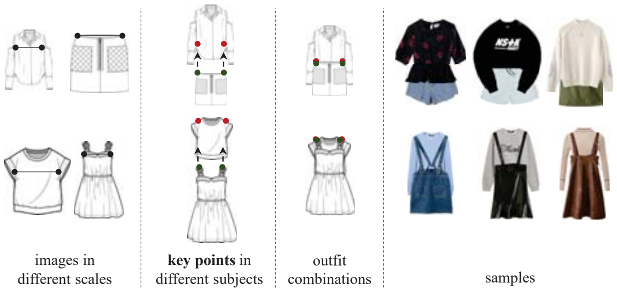


Figure 10. FashionAI for the outfit combinations generation

5.2. Data statistics

We released FashionAI dataset with 54 labels of design attributes in 8 dimensions and 24 key points in 324k images in 2018. Figure 11 presents the statistical results of the published 8 dimensions. The annotated samples are shown in Figure 12.

6. Discussion

In this paper, we present FashionAI Dataset as the basis of understanding tasks in which the fashion semantics are dissembled into different concepts which are clearly defined. The logical inner connection between those units is systematically embedded in hierarchical structure. To ensure the dataset being practically used in the real applications, we avoid discarding data when facing the problems of structure noise, attributes inconsistency, occlusion and uncertainty. Meanwhile, through the iterative process, new data is added from the online products. We ensure that the model trained on FashionAI database has a good generalization ability and practicality. The whole process of building this dataset is described in detail, which serves as a good reference for building similar professional datasets in any other field.

There are several promising directions for future annotations on our dataset. We currently only label "single model images" or "single tiled image", but labeling "multiple

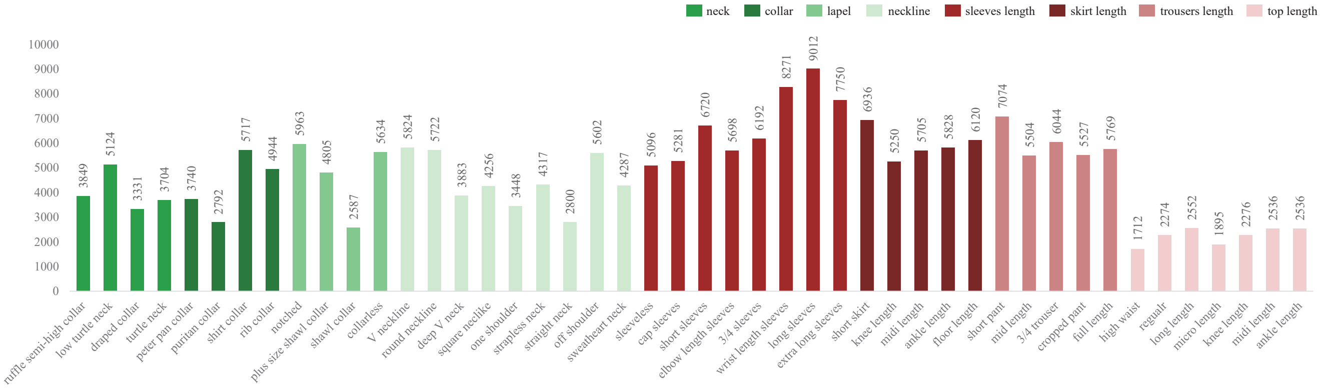


Figure 11. The number of instances per attribute for the eight attribute dimensions

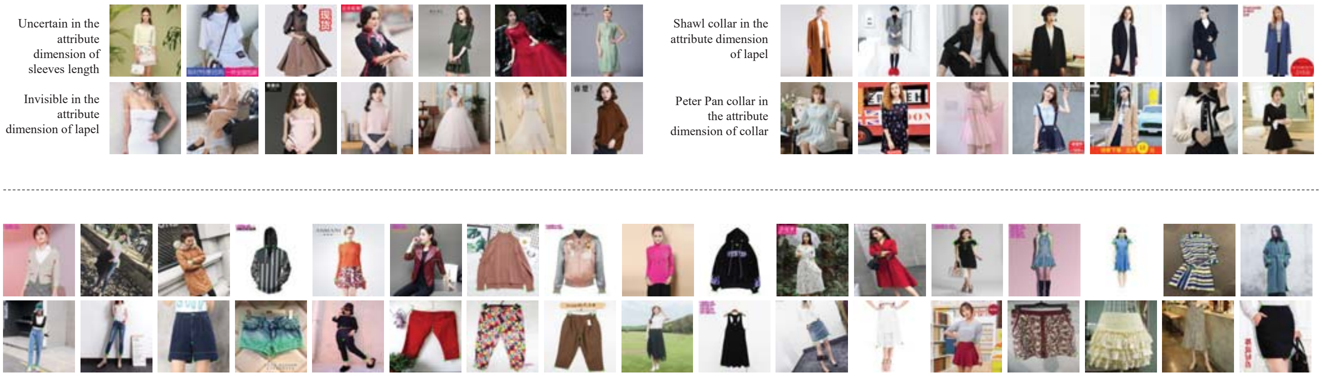

Uncertain in the attribute dimension of sleeves length
Invisible in the attribute dimension of lapel
Shawl collar in the attribute dimension of lapel
Peter Pan collar in the attribute dimension of collar

Figure 12. Samples of annotated images in the FashionAI Dataset

models image" or "multiple tiled image" may be useful for recognition. Additionally, layered wear recognition tasks that can be applied for mix and match recommendation are not considered in our current dataset. Finally, the attributes in our dataset are fine-grained enough that could provide a fundamental mapping for fashion style. We are now active in exploring building the mapping structure for fashion style based on FashionAI Dataset.

To download or learn more about FashionAI Dataset, please see the FashionAI official website[1].

7. Acknowledgement

This work was supported by the Hong Kong Polytechnic University (Grant No. RUWZ), and Alibaba Group through Alibaba Innovative Research (AIR) Program.

References

[1] K. E. Ak, A. A. Kassim, J. H. Lim, and J. Y. Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7708–7717, 2018.

[2] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Asian conference on computer vision*, pages 321–335. Springer, 2012.

[3] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Asian conference on computer vision*, pages 321–335, 2012.

[4] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5315–5324, 2015.

[5] X. Chen, Y. Zhang, H. Xu, Y. Cao, Z. Qin, and H. Zha. Visually explainable recommendation. *arXiv preprint arXiv:1801.10288*, 2018.

[6] Y. Cui, Q. Liu, C. Gao, and Z. Su. Fashiongan: Display your fashion design using conditional generative adversarial nets. In *Computer Graphics Forum*, volume 37, pages 109–119. Wiley Online Library, 2018.

[7] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 8–13, 2013.

[8] Z. Feng, Z. Yu, Y. Yang, Y. Jing, J. Jiang, and M. Song. Interpretable partitioned embedding for customized multi-item

[1] http://tianchi.aliyun.com/markets/tianchi/FashionAI

fashion outfit composition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 143–151. ACM, 2018.

[9] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *Asian conference on computer vision*, pages 420–431, 2012.

[10] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images.

[11] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1078–1086. ACM, 2017.

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

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] W.-L. Hsiao and K. Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *2017IEEE International Conference on Computer Vision (ICCV)*, pages 4203–4212, 2017.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] J. Huang, R. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1062–1070, 2015.

[17] T. Iwata, S. Wanatabe, and H. Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, volume 1, page 2, 2016.

[18] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, pages 3343–3351, 2015.

[19] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, pages 472–488. Springer, 2014.

[20] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, pages 472–488, 2014.

[21] W. H. Lin, K.-T. Chen, H. Y. Chiang, and W. Hsu. Netizen-style commenting on fashion photos: Dataset and diversity measures. *arXiv preprint arXiv:1801.10300*, 2018.

[22] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke. Explainable fashion recommendation with joint outfit matching and comment generation. *arXiv preprint arXiv:1806.08977*, 2018.

[23] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1):253–265, 2014.

[24] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM international conference on Multimedia*, pages 619–628, 2012.

[25] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3330–3337, 2012.

[26] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

[27] K. Matzen, K. Bala, and N. Snavely. Streetstyle: Exploring world-wide clothing styles from millions of photos. *arXiv preprint arXiv:1706.01869*, 2017.

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

[29] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, volume 2, page 6, 2015.

[30] X. Song, F. Feng, X. Han, X. Yang, W. Liu, and L. Nie. Neural compatibility modeling with attentive knowledge distillation. *arXiv preprint arXiv:1805.00313*, 2018.

[31] M. Takagi, E. Simo-Serra, S. Iizuka, and H. Ishikawa. What Makes a Style: Experimental Analysis of Fashion Prediction. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[32] P. Tangseng and T. Okatani. Toward explainable fashion recommendation. *arXiv preprint arXiv:1901.04870*, 2019.

[33] P. Tangseng, Z. Wu, and K. Yamaguchi. Looking at outfit to parse clothing. *arXiv preprint arXiv:1703.01386*, 2017.

[34] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4271–4280, 2018.

[35] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 3519–3526, 2013.

[36] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, 2012.

[37] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi. Mix and match: Joint model for clothing and attribute recognition. In *BMVC*, volume 1, page 4, 2015.

[38] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3182–3189, 2014.

[39] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. *arXiv preprint arXiv:1710.07346*, 2017.