Fashion Meets Computer Vision: A Survey

WEN-HUANG CHENG, National Chiao Tung University and National Chung Hsing University
SIJIE SONG, Peking University
CHIEH-YUN CHEN, National Chiao Tung University
SHINTAMI CHUSNUL HIDAYATI, Institut Teknologi Sepuluh Nopember
JIAYING LIU, Peking University

Fashion is the way we present ourselves to the world and has become one of the world's largest industries. Fashion, mainly conveyed by vision, has thus attracted much attention from computer vision researchers in recent years. Given the rapid development, this article provides a comprehensive survey of more than 200 major fashion-related works covering four main aspects for enabling intelligent fashion: (1) Fashion detection includes landmark detection, fashion parsing, and item retrieval; (2) Fashion analysis contains attribute recognition, style learning, and popularity prediction; (3) Fashion synthesis involves style transfer, pose transformation, and physical simulation; and (4) Fashion recommendation comprises fashion compatibility, outfit matching, and hairstyle suggestion. For each task, the benchmark datasets and the evaluation protocols are summarized. Furthermore, we highlight promising directions for future research.

1 INTRODUCTION

Fashion is how we present ourselves to the world. The way we dress and do our makeup defines our unique style and distinguishes us from other people. Fashion in modern society has become an indispensable part of a person's identity. Unsurprisingly, the global fashion apparel market alone

has surpassed 3 trillion U.S. dollars today and accounts for nearly 2% of the world's Gross Domestic Product.[1] Specifically, revenue in the fashion segment amounts to over U.S. $718 billion in 2020 and is expected to grow annually at a rate of 8.4%.[2]

As the revolution of **computer vision (CV)** with **artificial intelligence (AI)** is underway, AI is starting to hit the field of fashion, thereby reshaping our fashion life with a wide range of application innovations from electronic retailing to personalized stylist to the fashion design process. In this article, we term CV-enabled fashion technology as *intelligent fashion*. Technically, intelligent fashion is a challenging task, because, unlike generic objects, fashion items suffer from significant variations in style and design, and, most importantly, the long-standing semantic gap between computable low-level features and high-level semantic concepts that they encode is huge.

There are few previous works [120, 165] related to short fashion surveys. In 2014, Liu et al. [120] presented an initial literature survey focused on intelligent fashion analysis with facial beauty and clothing analysis, which introduced the representative works published during 2006–2013. However, thanks to the rapid development of computer vision, there are far more than these two domains within intelligent fashion, e.g., style transfer, physical simulation, and fashion prediction. There are a lot of related works that need to be updated. In 2018, Song and Mei [165] introduced the progress in fashion research with multimedia, which categorized fashion tasks into three aspects: low-level pixel computation, mid-level fashion understanding, and high-level fashion analysis. Low-level pixel computation aims to generate pixel-level labels on the image, such as human segmentation, landmark detection, and human pose estimation. Mid-level fashion understanding aims to recognize fashion images, such as fashion items and fashion styles. High-level fashion analysis includes recommendation, fashion synthesis, and fashion trend prediction. However, there is still a lack of a systematic and comprehensive surveys to paint the whole picture of intelligent fashion so as to summarize and classify state-of-the-art methods, discuss datasets and evaluation metrics, and provide insights for future research directions.

Current studies on intelligent fashion cover the research topics not only to detect what fashion items are presented in an image but also analyze the items, synthesize creative new ones, and, finally, provide personalized recommendations. Thus, in this article, we organize these research topics accordingly, as categorized in Figure 1, which includes fashion image detection, analysis, synthesis, and recommendation. In addition, we also give an overview of the main applications in the fashion domain, showing the power of intelligent fashion in the fashion industry. Overall, the contributions of our work can be summarized as follows:

- We provide a comprehensive survey of the current state-of-the-art research progress in the fashion domain and categorize fashion research topics into four main categories: detection, analysis, synthesis, and recommendation.
- For each category in the intelligent fashion research, we provide an in-depth and organized review of the most significant methods and their contributions. Also, we summarize the benchmark datasets as well as the links to the corresponding online portals.
- We gather evaluation metrics for different problems and also give performance comparisons for different methods.
- We list possible future directions that would help upcoming advances and inspire the research community.

This survey is organized in the following sections. Section 2 reviews fashion detection tasks, including landmark detection, fashion parsing, and item retrieval. Section 3 illustrates the works

---

[1]https://fashionunited.com/global-fashion-industry-statistics/.
[2]https://www.statista.com/outlook/244/100/fashion/worldwide.

Fig. 1. Scope of the intelligent fashion research topics covered in this survey article.

for fashion analysis containing attribute recognition, style learning, and popularity prediction. Section 4 provides an overview of fashion synthesis tasks comprising style transfer, human pose transformation, and physical texture simulation. Section 5 talks about works of fashion recommendation involving fashion compatibility, outfit matching, and hairstyle suggestion. Besides, Section 6 demonstrates selected applications and future work. Last, concluding remarks are given in Section 7.

2 FASHION DETECTION

Fashion detection is a widely discussed technology, since most fashion works need detection first. Take virtual try-on as an example [67]. It needs to early detect the human body part of the input image for knowing where the clothing region is and then synthesize the clothing there. Therefore, detection is the basis for most extended works. In this section, we mainly focus on fashion detection tasks, which are split into three aspects: landmark detection, fashion parsing, and item retrieval. For each aspect, state-of-the-art methods, the benchmark datasets, and the performance comparison are rearranged.

2.1 Landmark Detection

Fashion landmark detection aims to predict the positions of functional keypoints defined on the clothes, such as the corners of the neckline, hemline, and cuff. These landmarks not only indicate the functional regions of clothes but also implicitly capture their bounding boxes, ensuring that the design, pattern, and category of the clothes can be better distinguished. Indeed, features extracted from these landmarks greatly facilitate fashion image analysis.

It is worth mentioning the difference between fashion landmark detection and human pose estimation, which aims at locating human body joints, as Figure 2(a) shows. Fashion landmark detection is a more challenging task than human pose estimation, as the clothes are intrinsically more complicated than human body joints. In particular, garments undergo non-rigid deformations or scale variations, while human body joints usually have more restricted deformations. Moreover, the local regions of fashion landmarks exhibit more significant spatial and appearance variances than those of human body joints, as shown in Figure 2(b).
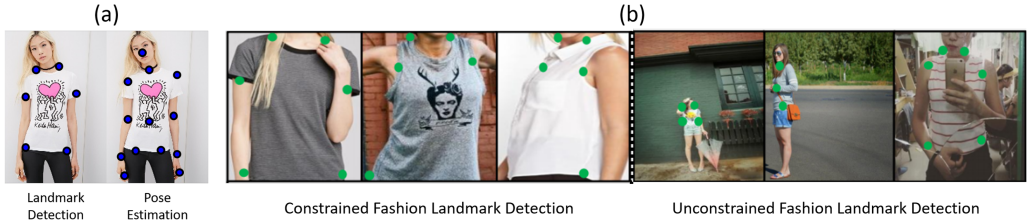
Fig. 2. (a) The visual difference between landmark detection and pose estimation. (b) The visual difference between the constrained fashion landmark detection and the unconstrained fashion landmark detection [209].

2.1.1 State-of-the-art Methods. The concept of fashion landmark was first proposed by Liu et al. [123] in 2016, under the assumption that clothing bounding boxes are given as prior information in both training and testing. For learning the clothing features via simultaneously predicting the clothing attributes and landmarks, Liu et al. introduced FashionNet [123], a deep model. The predicted landmarks were used to pool or gate the learned feature maps, which led to robust and discriminative representations for clothes. In the same year, Liu et al. also proposed the **deep fashion alignment (DFA)** framework [124], which consists of a three-stage deep **convolutional neural network (CNN)**, where each stage subsequently refined previous predictions. Yan et al. [209] further relaxed the clothing bounding box constraint, which is computationally expensive and inapplicable in practice. The proposed **Deep LAndmark Network (DLAN)** combined selective dilated convolution and recurrent spatial transformer, where bounding boxes and landmarks were jointly estimated and trained iteratively in an end-to-end manner. Both Reference [124] and Reference [209] are based on the regression model.

A more recent work [185] indicated that the regression model is highly non-linear and difficult to optimize. Instead of regressing landmark positions directly, they proposed to predict a confidence map of positional distributions (i.e., heatmap) for each landmark. Additionally, they took into account the fashion grammar to help reason the positions of landmarks. For instance, "*left collar ↔ left waistline ↔ left hemline*" connects a human-parts kinematic chain to be used as a constraint on the connected clothing parts to model the grammar topology. The human anatomical constraints were inherited in a recurrent neural network. Further, Lee et al. [96] considered contextual knowledge of clothes and proposed a global-local embedding module for achieving more accurate landmark prediction performance. Ge et al. [38] presented a versatile benchmark Deepfashion2 for four tasks, clothes detection, pose estimation, human segmentation, and clothing retrieval, which covered most significant fashion detection works. They built a strong model Match R-CNN based on Mask R-CNN [52] for solving the four tasks.

2.1.2 Benchmark Datasets. As summarized in Table 1, there are four benchmark datasets for fashion landmark detection, and the most used is Fashion Landmark Dataset [124]. These datasets differ in two major aspects: (1) standardization process of the images and (2) pose and scale variations.

2.1.3 Performance Evaluations. Fashion landmark detection algorithms output the landmark (i.e., functional key point) locations in the clothing images. The **normalized error (NE)**, which is defined as the ℓ_2 distance between detected and the ground-truth landmarks in the normalized coordinate space, is the most popular evaluation metric used in fashion landmark detection benchmarks. Typically, smaller values of NE indicates better results.

Table 1. Summary of the Benchmark Datasets for Fashion Landmark Detection Task

Dataset name	Publish time	# of photos	# of landmark annotations	Key features	Sources
DeepFashion-C [123]	2016	289,222	8	Annotated with clothing bounding box, pose variation type, landmark visibility, clothing type, category, and attributes	Online shopping sites, Google Images
Fashion Landmark Dataset (FLD) [124]	2016	123,016	8	Annotated with clothing type, pose variation type, landmark visibility, clothing bounding box, and human body joint	DeepFashion [123]
Unconstrained Land- mark Database [209]	2017	30,000	6	The unconstrained datasets are often with cluttered background and deviate from image center. The visual comparison between constrained and unconstrained dataset is presented in Figure 2(b)	Fashion blogs forums, Deep-Fashion [123]
DeepFashion2 [38]	2019	491,000	over 3.5x of Reference [123]	A versatile benchmark of four tasks including clothes detection, pose estimation, segmentation, and retrieval	Deep-Fashion [123], Online shopping sites

Table 2. Performance Comparisons of Fashion Landmark Detection Methods in Terms of NE

Dataset	Method	L. Collar	R. Collar	L. Sleeve	R. Sleeve	L. Waistline	R. Waistline	L. Hem	R. Hem	Avg.
DeepFashion -C [123]	DFA [124]	0.0628	0.0637	0.0658	0.0621	0.0726	0.0702	0.0658	0.0663	0.0660
	DLAN [209]	0.0570	0.0611	0.0672	0.0647	0.0703	0.0694	0.0624	0.0627	0.0643
	AttentiveNet [185]	0.0415	0.0404	0.0496	0.0449	0.0502	0.0523	0.0537	0.0551	0.0484
	Global-Local [96]	**0.0312**	**0.0324**	**0.0427**	**0.0434**	**0.0361**	**0.0373**	**0.0442**	**0.0475**	**0.0393**
FLD [124]	DFA [124]	0.0480	0.0480	0.0910	0.0890	–	–	0.0710	0.0720	0.0680
	DLAN [209]	0.0531	0.0547	0.0705	0.0735	0.0752	0.0748	0.0693	0.0675	0.0672
	AttentiveNet [185]	0.0463	0.0471	**0.0627**	**0.0614**	0.0635	0.0692	0.0635	**0.0527**	0.0583
	Global-Local [96]	**0.0386**	**0.0391**	0.0675	0.0672	**0.0576**	**0.0605**	**0.0615**	0.0621	**0.0568**

"L. Collar" denotes left collar, while "R. Collar" denotes right collar.

"–" denotes that detailed results are not available.

Fig. 3. Examples of semantic segmentation for fashion images [78].

We list the performance comparisons of leading methods on the benchmark datasets in Table 2. Moreover, the performances of the same method are different across datasets, but the rank is generally consistent.

2.2 Fashion Parsing

Fashion parsing, human parsing with clothes classes, is a specific form of semantic segmentation, where the labels are based on the clothing items, such as dress or pants. An example of fashion parsing is shown in Figure 3. Fashion parsing task distinguishes itself from general object or scene

segmentation problems in that fine-grained clothing categorization requires higher-level judgment based on the semantics of clothing, the deforming structure within an image, and the potentially large number of classes.

2.2.1   State-of-the-art Methods. The early work in fashion parsing was conducted by Yamaguchi et al. [206]. They exploited the relationship between clothing parsing and human pose estimation by refining two problems mutually. Specifically, clothing labels for every image segment were predicted with respect to body parts in a Conditional Random Field model. Then the predictions of clothing were incorporated as additional features for pose estimation. Their work, however, mainly focused on constrained parsing problem, where test images were parsed given user-provided tags indicating depicted clothing items. To overcome this limitation, References [205, 207] proposed clothes parsing with a retrieval-based approach. For a given image, similar images from a parsed dataset were first retrieved, and then the nearest-neighbor parsings were transferred to the final result via dense matching. Since pixel-level labels required for model training were time-consuming, Liu et al. [116] introduced the fashion parsing task with weak supervision from the color-category tags instead of pixel-level tags. They combined the human pose estimation module and (super)pixel-level category classifier learning module to generate category classifiers. They then applied the category tags to complete the parsing task.

Different from the above-mentioned works that tended to consider the human pose first, which might lead to sub-optimal results due to the inconsistent targets between pose estimation and clothing parsing, recent research studies mainly attempted to relax this constraint. Dong et al. [32] proposed to use Parselets, a group of semantic image segments obtained from a low-level over-segmentation algorithm, as the essential elements. A Deformable Mixture Parsing Model based on the "And-Or" structure of sub-trees was built to jointly learn and infer the best configuration for both appearance and structure. Next, References [106, 212] exploited contexts of clothing configuration, e.g., spatial locations and mutual relations of clothes items, to jointly parse a batch of clothing images given the image-level clothing tags. The proposed **Clothes Co-Parsing (CCP)** framework consists of two phases of inference: (1) image co-segmentation for extracting distinguishable clothes regions by applying exemplar-**Support Vector Machine (SVM)** classifiers, and (2) region co-labeling for recognizing garment items by optimizing a multi-image graphical model. Hidayati et al. [60] integrated local features and possible body positions from each superpixel as the instances of the price-collecting Steiner tree problem.

In particular, the clothing parsing approaches based on hand-crafted processing steps need to be designed carefully to capture the complex correlations between clothing appearance and structure fully. To tackle this challenge, some CNN-based approaches have been explored. Liang et al. [107] developed a framework based on active template regression to locate the mask of each semantic label, rather than assigning a label to each pixel. Two separate convolutional neural networks were utilized to build the end-to-end relation between the input image and the parsing result. Following Reference [107], Liang et al. later built a Contextualized CNN architecture [108] to simultaneously capture the cross-layer context, global image-level context, and local super-pixel contexts to improve the accuracy of parsing results.

To address the issues of parametric and non-parametric human parsing methods that relied on the hand-designed pipelines composed of multiple sequential components, such as in References [32, 116, 205, 206], Liu et al. presented a quasi-parametric human parsing framework [119]. The model inherited the merits of both parametric models and non-parametric models by the proposed Matching Convolutional Neural Network, which estimated the matching semantic region between the input image and KNN image. The works [41, 105] proposed self-supervised structure-sensitive learning approaches to explicitly enforce the consistency between the parsing results and

the human joint structures. In this way, there is no need for specifically labeling human joints in model training.

Unlike previous approaches that only focused on a single-person parsing task, References [40, 152, 224] presented different methods for solving multi-person human parsing. Zhao et al. [224] presented a deep Nested Adversarial Network[3] that contained three **Generative Adversarial Network– (GAN)** sub-nets for semantic saliency prediction, instance-agnostic parsing, and instance-aware clustering, respectively. These three sub-nets jointly learned in an end-to-end training way. Gong et al. [40] designed a detection-free **Part Grouping Network (PGN)** to deal with multi-person human parsing in an image in a single pass. The proposed PGN integrates two twinned subtasks that can be mutually refined under a unified network, i.e., semantic part segmentation, and instance-aware edge detection. Further, Ruan et al. [152] proposed CE2P framework[4] containing three key modules, a high-resolution embedding module, a global context embedding module, and an edge perceiving module, for single human parsing. This work won first place within all three human parsing tracks in the seond **Look Into Person (LIP)** Challenge.[5] For multi-person parsing, they designed a global-to-local prediction process based on CE2P cooperating with Mask R-CNN to form an M-CE2P framework and achieved the multi-person parsing goal.

In 2019, hierarchical graph was considered for human parsing tasks [39, 187]. Wang et al. [187] defined the human body as a hierarchy of multi-level semantic parts and employed three processes (direct, top-down, and bottom-up) to capture the human parsing information for better parsing performance. For tackling human parsing in various domain via a single model without retraining on various datasets, Gong et al. [39] comprised hierarchical graph transfer learning based on the conventional parsing network to constitute a general human parsing model, Graphonomy,[6] which consisted of two processes. It first learned and propagated compact high-level graph representation among the labels within one dataset and then transferred semantic information across multiple datasets.

2.2.2 Benchmark Datasets. There are multiple datasets for fashion parsing, most of which are collected from Chictopia,[7] a social networking website for fashion bloggers. Table 3 summarizes the benchmark datasets for fashion parsing in more detail. To date, the most comprehensive one is the LIP dataset [41, 105], containing over 50,000 annotated images with 19 semantic part labels captured from a wider range of viewpoints, occlusions, and background complexity.

2.2.3 Performance Evaluations. There are multiple metrics for evaluating fashion parsing methods: (1) average Pixel Accuracy as the proportion of correctly labeled pixels in the whole image, (2) mean Average Garment Recall, (3) Intersection over Union as the ratio of the overlapping area of the ground truth and predicted area to the total area, (4) mean accuracy, (5) average precision, (6) average recall, (7) average F-1 score over pixels, and (8) foreground accuracy as the number of true pixels on the body over the number of actual pixels on the body.

In particular, most of the parsing methods are evaluated on Fashionista dataset [206] in terms of accuracy, average precision, average recall, and average F-1 score over pixels. We report the performance comparisons in Table 1 of Supplementary Material.

---

[3]https://github.com/ZhaoJ9014/Multi-Human-Parsing.
[4]https://github.com/liutinglt/CE2P.
[5]https://vuhcs.github.io/vuhcs-2018/index.html.
[6]https://github.com/Gaoyiminggithub/Graphonomy.
[7]http://chictopia.com.

Table 3. Summary of the Benchmark Datasets for Fashion Parsing Task

Dataset name		Publish time	# of photos	# of classes	Key features	Sources
Fashionista dataset [206]		2012	158,235	56	Annotated with tags, comments, and links	Chictopia.com
Daily Photos [32]		2013	2,500	18	High resolution images; Parselet definition	Chictopia.com
Paper Doll dataset [205, 207]		2013	339,797	56	Annotated with metadata tags denoting characteris- tics, e.g., color, style, occasion, clothing type, brand	Fashionista [206], Chictopia.com
CCP SYSU-Clothes [106, 212]		2014	2,098	57	Annotated with superpixel-level or image-level tags	Online shopping websites
Colorful-Fashion Dataset [116]		2014	2,682	23	Annotated with 13 colors	Chictopia.com
ATR [107]	Benchmark	2015	5,867	18	Standing people in frontal/near-frontal view with good visibilities of all body parts	Fashionista [206], Daily Photos [32], Colorful-Fashion Dataset [116]
	Human Parsing in the Wild		1,833		Annotated with pixel-level labels	N/A
Chictopia10k [108]		2015	10,000	18	It contains real-world images with arbitrary postures, views and backgrounds	Chictopia.com
LIP [41, 105]		2017	50,462	20	Annotated with pixelwise and body joints	Microsoft COCO [111]
PASCAL-Person-Part [196]		2017	3,533	14	It contains multiple humans per image in uncon- strained poses and occlusions	N/A
MHP	v1.0 [98]	2017	4,980	18	There are 7 body parts and 11 clothes and accessory categories	N/A
	v2.0 [224]	2018	25,403	58	There are 11 body parts and 47 clothes and accessory categories	N/A
Crowd Instance-level Human Parsing [40]		2018	38,280	19	Multiple-person images; pixelwise annotations in instance-level	Google, Bing
ModaNet [227]		2018	55,176	13	Annotated with pixel-level labels, bounding boxes, and polygons	PaperDoll [205]
DeepFashion2 [38]		2019	491,000	13	A versatile benchmark of four tasks including clothes detection, pose estimation, segmentation, and retrieval.	DeepFashion [123], Online shopping websites
Fashionpedia [80]		2020	48,000	46	There are 294 fine-grained attributes. It contains high resolution with 1710 × 2151 to maintain more details	Flickr, Free license photo websites

N/A: there is no reported information to cite.

2.3   Item Retrieval

As fashion e-commerce has grown over the years, there has been a high demand for innovative solutions to help customers find preferred fashion items with ease. Though many fashion online shopping sites support keyword-based searches, there are many visual traits of fashion items that are not easily translated into words. It thus attracts tremendous attention from many research communities to develop cross-scenario image-based fashion retrieval tasks for matching the real-world fashion items to the online shopping image. Given a fashion image query, the goal of image-based fashion item retrieval is to find similar or identical items from the gallery.

   2.3.1   State-of-the-art Methods. The notable early work on automatic image-based clothing retrieval was presented by Wang and Zhang [190]. To date, extensive research studies have been devoted to addressing the problem of cross-scenario clothing retrieval, since there is a large domain discrepancy between daily human photo captured in general environment and clothing images taken in ideal conditions (i.e., embellished photos used in online clothing shops). Liu et al. [122] proposed to utilize an unsupervised transfer learning method based on part-based alignment and

Fig. 4. (a) An illustration of exact clothing retrieval [45]. (b) An example for clothing attributes recognition.

features derived from the sparse reconstruction. Kalantidis et al. [86] presented clothing retrieval from the perspective of human parsing. A prior probability map of the human body was obtained through pose estimation to guide clothing segmentation, and then the segments were classified through locality-sensitive hashing. The visually similar items were retrieved by summing up the overlap similarities. Notably, References [86, 122, 190] are based on hand-crafted features.

With the advances of deep learning, there has been a trend of building deep neural network architectures to solve the clothing retrieval task. Huang et al. [69] developed a **Dual Attribute-aware Ranking Network (DARN)** to represent in-depth features using attribute-guided learning. DARN simultaneously embedded semantic attributes and visual similarity constraints into the feature learning stage, while at the same time modeling the discrepancy between domains. Li et al. [103] presented a hierarchical super-pixel fusion algorithm for obtaining the intact query clothing item and used sparse coding for improving accuracy. More explicitly, an over-segmentation hierarchical fusion algorithm with human pose estimation was utilized to get query clothing items and to retrieve similar images from the product clothing dataset.

The abovementioned studies are designed for similar fashion item retrieval. But more often, people desire to find the same fashion item, as illustrated in Figure 4(a). The first attempt on this task was achieved by Kiapour et al. [45], who developed three different methods for retrieving the same fashion item in the real-world image from the online shop, the street-to-shop retrieval task. The three methods contained two deep learning baseline methods, and one method aimed to learn the similarity between two different domains, street and shop domains. For the same goal of learning the deep feature representation, Wang et al. [188] adopted a Siamese network that contained two copies of the Inception-6 network with shared weights. Also, they introduced a *robust contrastive loss* to alleviate over-fitting caused by some positive pairs (containing the same product) that were visually different, and used one multi-task fine-tuning approach to learn a better feature representation by tuning the parameters of the siamese network with product images and general images from ImageNet [27]. Further, Jiang et al. [83] extended the one-way problem, street-to-shop retrieval task, to the bi-directional problem, street-to-shop and shop-to-street clothing retrieval task. They proposed a deep bi-directional cross-triplet embedding algorithm to model the similarity between cross-domain photos, and further expanded the utilization of this approach to retrieve a series of representative and complementary accessories to pair with the shop item [84]. Moreover, Cheng et al. [24] increased the difficulty of image-based to video-based street-to-shop retrieval tasks, which was more challenging because of the diverse viewpoint or motion blur. They introduced three networks: Image Feature Network, Video Feature Network, and Similarity Network. They first did clothing detection and tracking for generating clothing trajectories. Then, Image Feature Network extracted deep visual features that would be fed into the **Long Short-Term Memory networks (LSTM)** framework for capturing the temporal dynamics in the Video Feature Network, and finally went to Similarity Network for pairwise matching. For improving the existing algorithms for retrieval tasks, which only considered global feature vectors, Reference [91] proposed a Graph Reasoning Network to build the similarity pyramid,

Fig. 5. An example of interactive fashion item retrieval [90].

which represented the similarity between a query and a gallery clothing by considering both global and local representation.

In the clothing retrieval methods described above, the retrieval is only based on query images that reflect users' needs, without considering that users may want to provide extra keywords to describe the desired attributes that are absent in the query image. Toward this goal, Kovashka et al. developed the WhittleSearch [90] that allows the user to upload a query image and provide additional relative attribute description as feedback for iteratively refining the image retrieval results (see Figure 5). Besides, for advancing the retrieval task with attribute manipulation, given a query image with attribute manipulation request, such as "I want to buy a jacket like this query image, but with plush collar instead of round collar," the memory-augmented Attribute Manipulation Network [223] updated the query representation encoding the unwanted attributes and replacing them to the desired ones. Later, Ak et al. [1] presented the FashionSearchNet to learn regions and region-specific attribute representations by exploiting the attribute activation maps generated by the global average pooling layer. Different from References [1, 223] that constructed visual representation of searched item by manipulating the visual representation of query image with the textual attributes in query text, Reference [92] inferred the semantic relationship between visual and textual attributes in a joint multimodal embedding space. To help facilitate the reasoning of search results and user intent, Liao et al. [109] built the Exclusive and Independent tree, that captured hierarchical structures of the fashion concepts, which was generated by incorporating product hierarchies of online shopping websites and domain knowledge of fashion experts. It was then applied to guide the end-to-end deep learning procedure, mapping the deep implicit features to explicit fashion concepts.

As the fashion industry attracts much attention recently, there comes a multimedia grand challenge "AI meets beauty,"[8] which aimed for competing for the top fashion item recognition methods, held in ACM Multimedia yearly from 2018. Perfect Corp., CyberLink Corp., and National Chiao Tung University in Taiwan held this grand challenge and provided a large-scale image dataset of beauty and personal care products, namely the Perfect-500K dataset [23]. Lin et al. [112] received the top performance in the challenge in 2019. They presented an unsupervised embedding learning to train a CNN model and combined the existing retrieval methods trained on different datasets to finetune the retrieval results.

2.3.2 Benchmark Datasets. The existing clothing retrieval studies mainly focused on a cross-domain scenario. Therefore, most of the benchmark datasets were collected from daily photos and online clothing shopping websites. Table 4 gives a summary and link of the download page (if publicly available) of the benchmark datasets.

2.3.3 Performance Evaluations. There are some evaluation metrics used to assess the performance of clothing retrieval methods. The different evaluation metrics used are as follows: (1) Top-k retrieval accuracy, the ratio of queries with at least one matching item retrieved within the top-k returned results, (2) Precision@k, the ratio of items in the top-k returned results that are matched

---

[8]https://challenge2020.perfectcorp.com/.

Table 4. Summary of the Benchmark Datasets for Fashion Retrieval Task

Dataset name		Publish time	# of photos (videos)	Key features	Sources
Fashion 10000 [127]		2014	32,398	Annotated with 470 fashion categories	Flickr.com
Deep Search [70]		2014	206,235	With template matching, 7 attributes and type are extract- ed from the descriptions, i.e. pattern, sleeve, button panel, collar, style, and color	Taobao.com, Tsmall.com, Amazon.com
DARN [69]		2015	545,373	Online-offline upper-clothing image pair, annotated with clothing attribute categories	Online-shopping sites and corresponding customer review pages
Exact Street-2Shop [45]	Street photos	2015	20,357	39,479 pairs of exactly matching items worn in street photos and shown in shop photos	ModCloth.com
	Shop photos		404,683		Online clothing retailers
MVC [113]		2016	161,638	Annotated with 264 attribute labels	Online shopping sites
Li et al. [103]	Product Clothing	2016	15,690	Annotated with clothing categories	Online shopping sites
	Daily Clothing		4,206		Flickr.com
DeepFashion (In-shop Clothes Retrieval Benchmark) [123]		2016	52,712	The resolution of images is 256×256	Online shopping sites (Forever21 and Mogujie), Google Images
Video2Shop [24]	Videos	2017	26,352 (526)	39,479 exact matching pairs annotated with 14 categories of clothes	Tmall MagicBox
	Online shopping		85,677		Tmall.com, Taobao.com
Dress like a star [37]		2017	7,000,000 (40)	It contains different movie genres such as animation, fantasy, adventure, comedy or drama	YouTube.com
Amazon [109]		2018	489,000	Annotated with 200 clothing categories	Amazon.com
Perfect-500K [23]		2018	500,000	It is vast in scale, rich and diverse in content to collect as many as possible beauty and personal care items	e-commerce websites
DeepFashion2 [38]		2019	491,000	a versatile benchmark of four tasks including clothes detection, pose estimation, segmentation, and retrieval	DeepFashion [123], Online shopping websites
FindFashion [91]		2019	565,041	Merge two existing datasets, Street2Shop and Deep- Fashion, and label three attributes of the most affected	Street2Shop [45], DeepFashion [123]
Ma et al. [132]		2020	180,000	Since dataset for attribute-specific fashion retrieval is lacking, this dataset rebuild three fashion dataset with attribute annotations	DARN [69], FashionAI [232], DeepFashion [123]

with the queries, (3) Recall@k, the ratio of matching items that are covered in the top-k returned results, (4) **Normalized Discounted Cumulative Gain (NDCG)** (NDCG@k), the relative orders among matching and non-matching items within the top-k returned results, and (5) **Mean Average Precision (MAP)**, which measures the precision of returned results at every position in the ranked sequence of returned results across all queries.

Table 2 of the Supplementary Material presents the performance comparisons of some retrieval methods reviewed in this survey. We are unable to give comparisons for all different retrieval methods, because the benchmarks they used are not consistent.

3 FASHION ANALYSIS

Fashion is not only about what people are wearing but also reveals personality traits and other social cues. With immense potential in the fashion industry, precision marketing, sociological analysis, and so on, intelligent fashion analysis on what style people choose to wear has thus gained

increasing attention in recent years. In this section, we mainly focus on three fields of fashion analysis: attribute recognition, style learning, and popularity prediction. For each field, state-of-the-art methods, the benchmark datasets, and the performance comparison are summarised.

3.1 Attribute Recognition

Clothing attribute recognition is a multi-label classification problem that aims at determining which elements of clothing are associated with attributes among a set of n attributes. As illustrated in Figure 4(b), a set of attributes is a mid-level representation generated to describe the visual appearance of a clothing item.

3.1.1 State-of-the-art Methods. Chen et al. [14] learned a list of attributes for clothing on the human upper body. They extracted low-level features based on human pose estimation and then combined them for learning attribute classifiers. Mutual dependencies between the attributes capturing the rules of style (e.g., neckties are rarely worn with T-shirts) were explored by a **Conditional Random Field (CRF)** to make attribute predictions. A CRF-based model was also presented in Reference [208]. The model considered the location-specific appearance with respect to a human body and the compatibility of clothing items and attributes, which was trained using a max-margin learning framework.

Motivated by the large discrepancy between images captured in constrained and unconstrained environments, Chen et al. [20] studied a cross-domain attribute mining. They mined the data from *clean* clothing images obtained from online shopping stores and then adapted it to unconstrained environments by using a deep domain adaptation approach. Lu et al. [170] further presented a study on part-based clothing attribute recognition in a search and mining framework. The method consists of three main steps: (1) Similar visual search with the assistance of pose estimation and part-based feature alignment; (2) Part-based salient tag extraction that estimates the relationship between tags and images by the analysis of intra-cluster and inter-cluster of clothing essential parts; and (3) Tag refinement by mining visual neighbors of a query image. In the meantime, Li et al. [101] learned to score the human body and attribute-specific parts jointly in a deep Convolutional Neural Network, and further improved the results by learning collaborative part modeling among humans and global scene re-scoring through deep hierarchical contexts.

Different from the above methods that conducted attribute recognition only based on annotated attribute labels, References [26, 48, 177] proposed to identify attribute vocabulary using weakly labeled image-text data from shopping sites. They used the neural activations in the deep network that generated attribute activation maps through training a joint visual-semantic embedding space to learn the characteristics of each attribute. In particular, Vittayakorn et al. [177] exploited the relationship between attributes and the divergence of neural activations in the deep network. Corbiere et al. [26] trained two different and independent deep models to perform attribute recognition. Han et al. [48] derived spatial-semantic representations for each attribute by augmenting semantic word vectors for attributes with their spatial representation.

Besides, there are research works exploring attributes for recognizing the type of clothing items. Hidayati et al. [56] introduced clothing genre classification by exploiting the discriminative attributes of style elements, with an initial focus on the upperwear clothes. The work in Reference [61] later extended [56] to recognize the lowerwear clothes. Yu and Grauman [218] proposed a local learning approach for fine-grained visual comparison to predict which image is more related to the given attribute. Jia et al. [79] introduced a notion of using the two-dimensional continuous image-scale space as an intermediate layer and formed a three-level framework, i.e., visual features of clothing images, image-scale space based on the aesthetic theory, and aesthetic words space consisting of words like "formal" and "casual." A Stacked Denoising Autoencoder Guided by

Table 5. Summary of the Benchmark Datasets for Clothing Attribute Recognition Task

Dataset name		Publish time	# of photos	# of categories	# of attributes	Key features	Sources
Clothing Attributes [14]		2012	1,856	7	26	Annotated with 23 binary-class attri- butes and 3 multi-class attributes	Thesartorialist.com, Flickr.com
Hidayati et al. [56]		2012	1,077	8	5	Annotated with clothing categories	Online shopping sites
UT-Zap50K shoe [218]		2014	50,025	N/A	4	Shoe images annotated with associated metadata (shoe type, materials, gender, manufacturer, *etc.*)	Zappos.com
Chen et al. [20]	Online-data	2015	341,021	15	67	Each attribute has 1000+ images	Online shopping sites
	Street-data-a		685	N/A	N/A	Annotated with fine-grained attributes	Fashionista [206]
	Street-data-b		8,000	N/A	N/A		Parsing [31]
	Street-data-c		4,200	N/A	N/A		Surveillance videos
Lu et al. [170]		2016	~1,1 M	N/A	N/A	Annotated with the associated tags	Taobao.com
WIDER Attribute [101]		2016	13,789	N/A	N/A	Annotated with 14 human attribute la- bels and 30 event class labels	The 50574 WIDER images [200]
Vittayakorn et al. [177]	Etsy	2016	173,175	N/A	250	Annotated with title and description of the product	Etsy.com
	Wear		212,129	N/A	250	Annotated with the associated tags	Wear.jp
DeepFashion-C [123]		2016	289,222	50	1,000	Annotated with clothing bounding box, type, category, and attributes	Online shopping sites, Google Images
Fashion200K [48]		2017	209,544	5	4,404	Annotated with product descriptions	Lyst.com
Hidayati et al. [61]		2018	3,250	16	12	Annotated with clothing categories	Online shopping sites
CatalogFashion-10x [8]		2019	1M	43	N/A	The categories are identical to the DeepFashion dataset [123]	Amazon.com

N/A: there is no reported information to cite.

Correlative Labels was proposed to map the visual features to the image-scale space. Ferreira et al. [146] designed a visual semantic attention model with pose guided attention for multi-label fashion classification. Besides the clothing attribute recognition works, micro expression recognition is also an interesting task [126, 199].

3.1.2 Benchmark Datasets. There have been several clothing attribute datasets collected, as the older datasets are not capable of meeting the needs of the research goals. In particular, more recent datasets have a more practical focus. We summarize the clothing attribute benchmark datasets in Table 5 and provide the links to download if they are available.

3.1.3 Performance Evaluations. Metrics that are used to evaluate the clothing attribute recognition models include the top-k accuracy, MAP, and **Geometric Mean (G-Mean)**. The top-k accuracy and MAP have been described in Section 2.3.3, while G-Mean measures the balance between classification performances on both the majority and minority classes. Most authors opt for a measure based on top-k accuracy. We present the evaluation for general attribute recognition on the DeepFashion-C Dataset with different methods in Table 6. The evaluation protocol is released in Reference [123].

3.2 Style Learning

A variety of fashion styles is composed of different fashion designs. Inevitably, these design elements and their interrelationships serve as the powerful source-identifiers for understanding fashion styles. The key issue in this field is thus how to analyze discriminative features for different styles and also learn what style makes a trend.

Table 6. Performance Comparisons of Attribute Recognition Methods in Terms of
Top-*k* Classification Accuracy

Method	Category		Texture		Fabric		Shape		Part		Style		All	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
Chen et al. [14]	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN [69]	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
FashionNet [123]	82.58	90.17	37.46	49.52	39.30	**49.84**	39.47	48.59	**44.13**	54.02	66.43	73.16	45.52	54.61
Corbiere et al. [26]	86.30	92.80	**53.60**	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40
AttentiveNet [185]	**90.99**	**95.78**	50.31	**65.48**	**40.31**	48.23	**53.32**	**61.05**	40.65	**56.32**	**68.70**	**74.25**	**51.53**	**60.95**

The best result is marked in bold.

3.2.1 State-of-the-art Methods. An early attempt in fashion style recognition was presented by Kiapour et al. [89]. They evaluated the concatenation of hand-crafted descriptors as style representation. Five different style categories are explored, including hipster, bohemian, pinup, preppy, and goth.

In the following studies [63, 82, 130, 164, 172], deep learning models were employed to represent fashion styles. In particular, Simo-Serra and Ishikawa [164] developed a joint ranking and classification framework based on the Siamese network. The proposed framework was able to achieve outstanding performance with features being the size of a SIFT descriptor. To further enhance feature learning, Jiang et al. [82] used a consensus style centralizing auto-encoder to centralize each feature of certain style progressively. Ma et al. introduced **Bimodal Correlative Deep Autoencoder (BCDA)** [130], a fashion-oriented multimodal deep learning based model adopted from Bimodal Deep Autoencoder [138], to capture the correlation between visual features and fashion styles. The BCDA learned the fundamental rules of tops and bottoms as two modals of clothing collocations. The shared representation produced by BCDA was then used as input of the regression model to predict the coordinate values in the fashion semantic space that describes styles quantitatively. Vaccaro et al. [172] presented a data-driven fashion model that learned the correspondences between high-level style descriptions (e.g., "valentines day" and low-level design elements (e.g., "red cardigan" by training polylingual topic modeling. This model adapted a natural language processing technique to learn latent fashion concepts jointly over the style and element vocabularies. Different from previous studies that sought coarse style classification, Hsiao and Grauman [63] treated styles as discoverable latent factors by exploring style-coherent representation. An unsupervised approach based on polylingual topic models was proposed to learn the composition of clothing elements that are stylistically similar. Further, interesting work for learning the user-centric fashion information based on occasions, clothing categories, and attributes was introduced by Ma et al. [131]. Their main goal is to learn the information about "what to wear for a specific occasion?" from social media, e.g., Instagram. They developed a contextualized fashion concept learning model to capture the dependencies among occasions, clothing categories, and attributes.

Fashion Trends Analysis. A research pioneer in automatic fashion trend analysis was presented by Hidayati et al. [59]. They investigated fashion trends at ten different seasons of New York Fashion Week by analyzing the coherence (to occur frequently) and uniqueness (to be sufficiently different from other fashion shows) of visual style elements. Following Reference [59], Chen et al. [17] utilized a learning-based clothing attributes approach to analyze the influence of the New York Fashion Show on people's daily life. Later, Gu et al. [43] presented QuadNet for analyzing fashion trends from street photos. The QuadNet was a classification and feature embedding learning network that consists of four identical CNNs, where the shared CNN was jointly

optimized with a multi-task classification loss and a neighbor-constrained quadruplet loss. The multi-task classification loss aimed to learn the discriminative feature representation, while the neighbor-constrained quadruplet loss aimed to enhance the similarity constraint. A study on modeling the temporal dynamics of the popularity of fashion styles was presented in Reference [53]. It captured fashion dynamics by formulating the visual appearance of items extracted from a deep convolutional neural network as a function of time.

Further, for statistics-driven fashion trend analysis, Reference [18] analyzed the best-selling clothing attributes corresponding to specific season via the online shopping websites statistics, e.g., (1) winter: gray, black, and sweater; (2) spring: white, red, and v-neckline. They designed a machine learning based method considering the fashion item sales information and the user transaction history for measuring the real impact of the item attributes to customers. Besides the fashion trend learning related to online shopping websites was considered, Ha et al. [46] used fashion posts from Instagram to analyze visual content of fashion images and correlated them with likes and comments from the audience. Moreover, Chang et al. achieved an interesting work [12] on what people choose to wear in different cities. The proposed "Fashion World Map" framework exploited a collection of geo-tagged street fashion photos from an image-centered social media site, Lookbook.nu. They devised a metric based on deep neural networks to select the potential iconic outfits for each city and formulated the detection problem of iconic fashion items of a city as the prize-collecting Steiner tree problem. A similar idea was proposed by Mall et al. [133] in 2019; they established a framework[9] for automatically analyzing the fashion trend worldwide in item attributes and style corresponding to the city and month. Specifically, they also analyzed how social events impact people wear, e.g., "new year in Beijing in February 2014" leads to red upper clothes.

Temporal estimation is an interesting task for general fashion trend analysis, the goal of which is to estimate when a style was made. Although visual analysis of fashion styles have been much investigated, this topic has not received much attention from the research community. Vittayakorn et al. [176] proposed an approach to the temporal estimation task using CNN features and fine-tuning new networks to predict the time period of styles directly. This study also provided estimation analyses of what the temporal estimation networks have learned.

3.2.2 Benchmark Datasets. Table 7 compares the benchmark datasets used in the literature. Hipster Wars [89] is the most popular dataset for style learning. We note that, instead of using the same dataset for training and testing, the study in Reference [164] used Fashion144k dataset [163] to train the model and evaluated it on Hipster Wars [89]. Moreover, the datasets related to fashion trend analysis were mainly collected from online media platforms during a specific time interval.

3.2.3 Performance Evaluations. The evaluation metrics used to measure the performance of existing fashion style learning models are precision, recall, and accuracy. Essentially, precision measures the ratio of retrieved results that are relevant, recall measures the ratio of relevant results that are retrieved, while accuracy measures the ratio of correct recognition. Although several approaches for fashion temporal analysis have been proposed, there is no extensive comparison between them. It is an open problem to determine which of the approaches perform better than others.

3.3 Popularity Prediction

Precise fashion trend prediction is not only essential for fashion brands to strive for the global marketing campaign but also crucial for individuals to choose what to wear corresponding to the

---

[9]https://github.com/kavitabala/geostyle.

Table 7. Summary of the Benchmark Datasets for Style Learning Task

Dataset name		Publish time	# of photos	Key features	Sources
Hipster Wars [89]		2014	1,893	Annotated with Bohemian, Goth, Hipster, Pinup, or Preppy style	Google Image Search
Hidayati et al. [59]		2014	3,276	Annotated with 10 seasons of New York Fashion Week (Spring/Summer 2010 to Autumn/Winter 2014)	FashionTV.com
Fashion144k [163]		2015	144,169	Each post contains text in the form of descriptions and gar- ment tags. It also consists of "likes" for scaling popularity	Chictopia.com
Chen et al. [17]	Street-chic	2015	1,046	Street-chic images in New York from April 2014 to July 2014 and from April 2015 to July 2015	Flickr.com, Pinterest.com
	New York Fashion Show		7,914	2014 and 2015 summer/spring New York Fashion Show	Vogue.com
Online Shopping [82]		2016	30,000	The 12 classes definitions are based on the fashion maga- zine [11]	Nordstrom.com, barneys.com
Fashion Data [172]		2016	590,234	Annotated with a title, text description, and a list of items the outfit comprises	Polyvore.com
He and McAuley et al. [53]	Women	2016	331,173	Annotated with users' review histories, time span between March 2003 and July 2014	Amazon.com
	Men		100,654		
Vittayakorn et al. [176]	Flickr Clothing	2017	58,350	Annotated with decade label and user provided metadata	Flickr.com
	Museum Dataset		9,421	Annotated with decade label	museum
Street Fashion Style [43]		2017	293,105	Annotated with user-provided tags, including geographical and year information	Chictopia.com
Global Street Fashion (GSFashion) [12]		2017	170,180	Annotated with (1) city name, (2) anonymized user ID, (3) anonymized post ID, (4) posting date, (5) number of likes of the post, and (6) user-provided metadata descri- bing the categories of fashion items along with the brand names and the brand-defined product categories	Lookbook.nu
Fashion Semantic Space [130]		2017	32,133	Full-body fashion show images; annotated with visual features (e.g., collar shape, pants length, or color theme.) and fashion styles (e.g., casual, chic, or elegant).	Vogue.com
Hsiao and Grauman [63]		2017	18,878	Annotated with associated attributes	Google Images
Ha et al. [46]		2017	24,752	It comprises description of images, associated metadata, and annotated and predicted visual content variables	Instagram
Geostyle [133]		2019	7.7M	This dataset includes categories of 44 major world cities across 6 continents, person body and face detection, and canonical cropping	Street Style [135], Flickr100M [178]
FashionKE [131]		2019	80,629	With the help of fashion experts, it contains 10 common types of occasion concepts, e.g., dating, prom, or travel	Instagram

M means million.

specific occasion. Based on the fashion style learning via the existing data (e.g., fashion blogs), it can better predict fashion popularity and further forecast future trend, which profoundly affect the fashion industry about multi-trillion U.S. dollars.

3.3.1 State-of-the-art Methods. Despite the active research in popularity prediction of online content on general photos with diverse categories [55, 134, 186, 192, 194], the popularity predic-tion on online social network specialized in fashion and style is currently understudied. The work in Reference [204] presented a vision-based approach to quantitatively evaluate the influence of visual, textual, and social factors on the popularity of outfit pictures. They found that the combina-tion of social and content information yields a good predictory for popularity. In Reference [142],

Park et al. applied a widely-used machine learning algorithm to uncover the ingredients of success of fashion models and predict their popularity within the fashion industry by using data from social media activity. Recently, Lo et al. [125] proposed a deep temporal sequence learning framework to predict the fine-grained fashion popularity of an outfit look.

Discovering the visual attractiveness has been the pursuit of artists and philosophers for centuries. Nowadays, the computational model for this task has been actively explored in the multimedia research community, especially with the focus on clothing and facial features. Nguyen et al. [139] studied how different modalities (i.e., face, dress, and voice) individually and collectively affect the attractiveness of a person. A tri-layer learning framework, namely Dual-supervised Feature-Attribute-Task network, was proposed to learn attribute models and attractiveness models simultaneously. Chen et al. [21] focused on modeling fashionable dresses. The framework was based on two main components: basic visual pattern discovery using active clustering with humans in the loop, and latent structural SVM learning to differentiate fashionable and non-fashionable dresses. Next, Simo-Serra et al. [163] not only predicted the fashionability of a person's look on a photograph but also suggested what clothing or scenery the user should change to improve the look. For this purpose, they utilized a Conditional Random Field model to learn correlation among fashionability factors, such as the type of outfit and garments, the type of the user, and the scene type of the photograph.

Meanwhile, the aesthetic quality assessment of online shopping photos is a relatively new area of study. A set of features related to photo aesthetic quality is introduced in Reference [181]. To be more specific, in this work, Wang and Allebach investigated the relevance of each feature to aesthetic quality via the elastic net. They trained an SVM predictor with an optimal feature subset constructed by a wrapper feature selection strategy with the best-first search algorithm.

Other studies built computational attractiveness models to analyze facial beauty. A previous survey on this task was presented in Reference [115]. Since some remarkable progress have been made on this subject, we extend [115] to cover recent advancements. Chen and Zhang [13] introduced a causal effect criterion to evaluate facial attractiveness models. It proposed two-way measurements, i.e., by imposing interventions according to the model and by examining the change of attractiveness. To alleviate the need for rating history for the query, which prior works could not cope when there are none or few, Rothe et al. [150] proposed to regress visual query to a latent space derived through matrix factorization for the known subjects and ratings. Besides, they employed a visual regularized collaborative filtering approach to infer inter-person preferences for attractiveness prediction. A **psychologically inspired deep convolutional neural network (PI-CNN)**, which is a hierarchical model that facilitates both the facial beauty representation learning and predictor training, was later proposed in Reference [201]. To optimize the performance of the PI-CNN facial beauty predictor, Reference [201] introduced a cascaded fine-tuning scheme that exploits appearance features of facial detail, lighting, and color. To further consider the facial shape, Gao et al. [36] designed a multi-task learning framework that took appearance and facial shape into account simultaneously and jointly learned facial representation, landmark location, and facial attractiveness score. They proved that learning with landmark localization is effective for facial attractiveness prediction. For building flexible filters to learn the mapping adaptive for different attributes within a deep modal, Lin et al. [110] proposed an attribute-aware convolutional neural network whose filter parameters were controlled adaptively by facial attributes. They also considered the cases without attribute labels and presented a pseudo attribute-aware network, which learned to utilize image context information for generating attribute-like knowledge. Moreover, Shi et al. [158] introduced a co-attention learning mechanism that employed facial parsing masks for learning accurate representation of facial composition to improve facial attractiveness prediction.

Table 8. Summary of the Benchmark Datasets for Popularity Prediction Task

Dataset name		Publish time	# of photos	Key features	Source
Yamaguchi et al. [204]		2014	328,604	Annotated with title, description, and user-provided labels	Chictopia.com
SCUT-FBP [198]		2015	500	Asian female face images with attractiveness ratings	Internet
Fashion144k [163]		2015	144,169	Each post contains text in the form of descriptions and garment tags. It also consists of votes or "likes" for scaling popularity	Chictopia.com
Park et al. [142]	Fashion Model Directory	2016	N/A	Profile of fashion models, including name, height, hip size, dress size, waist size, shoes size, list of agencies, and details about all runways the model walked on (year, season, and city)	Fashionmodel-directory.com
	Instagram			Annotated with the number of "likes" and comments, as well as the metadata of the first 125 "likes" of each post	Instagram
TPIC17 [192]		2017	680,000	With time information	Flickr.com
Massip et al. [134]		2018	6,000	Using textual queries with hashtags of targeted image categories to collect, e.g., #selfie, or #friend	Instagram
SCUT-FBP5500 [104]		2018	5,500	Frontal faces with various properties (gender, race, ages) and div- erse labels (face landmark and beauty score)	Internet
Lo et al. [125]		2019	380,000	Within 52 different cities and the timeline was between 2008−2016	lookbook.nu
SMPD2019 [193]		2019	486,000	It contains rich contextual information and annotations	Flickr.com

N/A: there is no reported information to cite.

There was a Social Media Prediction challenge[10] held by Wu et al. [193] in ACM Multimedia 2019 that aimed at the work focused on predicting future clicks of new social media posts before they were posted in social feeds. The participated teams need to build a new algorithm based on understanding and learning techniques and automatically predict popularity (formulated by clicks or visits) to achieve better performances.

Fashion Forecasting. There are strong practical interests in fashion sales forecasting, either by utilizing traditional statistical models [25, 140], applying artificial intelligence models [6], or combining the advantages of statistics-based methods and artificial intelligence-based methods into hybrid models [16, 88, 148]. However, the problem of visual fashion forecasting, where the goal is to predict the future popularity of styles discovered from fashion images, has gained limited attention in the literature. In Reference [2], Al-Halah et al. proposed to forecast the future visual style trends from fashion data in an unsupervised manner. The proposed approach consists of three main steps: (1) learning a representation of fashion images that captures clothing attributes using a supervised deep convolutional model; (2) discovering the set of fine-grained styles that are distributed across images using a non-negative matrix factorization framework; and (3) constructing styles' temporal trajectories based on statistics of past consumer purchases for predicting the future trends.

3.3.2 Benchmark Datasets. We summarize the benchmark datasets used to evaluate the popularity prediction models reviewed above in Table 8. There are datasets focused on face attractiveness called SCUT-FBP [198] and SCUT-FBP5500 [104]. Also, SMPD2019 [193] is specific for social media prediction.

3.3.3 Performance Evaluations. Mean Absolute Percentage Error, Mean Absolute Error, Mean Square Error, and **Spearman Ranking Correlation (SRC)** are the most used metrics to evaluate the popularity prediction performance. SRC is to measure the ranking correlation between ground-truth popularity set and predicted popularity set, varying from 0 to 1.

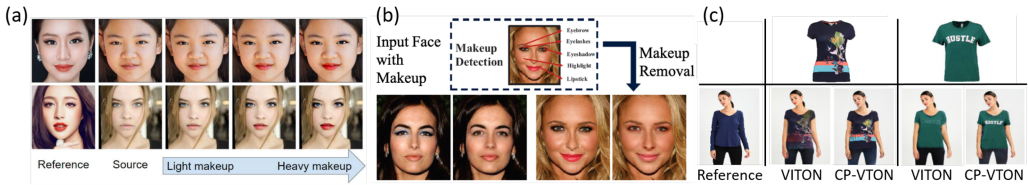
[10]http://www.smp-challenge.com/.

Fig. 6. (a) Facial makeup transfer [15]. (b) Facial makeup detection and removal [184]. (c) Comparison of VITON [51] and CP-VTON [179].

4 FASHION SYNTHESIS

Given an image of a person, we can imagine what that person would like in a different makeup or clothing style. We can do this by synthesizing a realistic-looking image. In this section, we review the progress to address this task, including style transfer, pose transformation, and physical simulation.

4.1 Style Transfer

Style transfer is transferring an input image into a corresponding output image such as transferring a real-world image into a cartoon-style image, transferring a non-makeup facial image into a makeup facial image, or transferring the clothing, which is tried on the human image, from one style to another. Style transfer in image processing contains a wide range of applications, e.g., facial makeup and virtual try-on.

4.1.1 State-of-the-art Methods. The most popular style transfer work is pix2pix [73], which is a general solution for style transfer. It learns not only the mapping from the input image to output image but also a loss function to train the mapping. For specific goal, based on a texture patch, References [81, 197][11] transferred the input image or sketch to the corresponding texture. Based on the human body silhouette, References [49, 94] inpainted compatible style garments to synthesize realistic images. An interesting work [159] learned to transferred the facial photo into the game character style.

Facial Makeup. Finding the most suitable makeup for a particular human face is challenging, given the fact that a makeup style varies from face-to-face due to the different facial features. Studies on how to automatically synthesize the effects of with or without makeup on one's facial appearance have aroused interest recently. There is a survey on computer analysis of facial beauty provided in Reference [95]. However, it is limited to the context of the perception of attractiveness.

Facial makeup transfer refers to translate the makeup from a given face to another one while preserving the identity as Figure 6(a). It provides an efficient way for virtual makeup try-on and helps users select the most suitable makeup style. The early work achieved this task by image processing methods [97], which decomposed images into multiple layers and transferred information from each layer after warping the reference face image to a corresponding layer of the non-makeup one. One major disadvantage of this method was that it required warping the reference face image to the non-makeup face image, which was very challenging and inclined to generate artifacts in many cases. Liu et al. [121] first adopted a deep learning framework for makeup transfer. They employed several independent networks to transfer each cosmetic on the corresponding facial part. The simple combination of several components applied in this framework, however, leads to unnatural artifacts in the output image. To address this issue, Alashkar et al. [3] trained a deep neural

---

[11]https://github.com/janesjanes/Pytorch-TextureGAN.

network based makeup recommendation model from examples and knowledge base rules jointly. The suggested makeup style was then synthesized on the subject face.

Different makeup styles result in significant facial appearance changes, which brings challenges to many practical applications, such as face recognition. Therefore, researchers also devoted themselves to makeup removal as Figure 6(b), which is an ill-posed problem. Wang and Fu [184] proposed a makeup detector and remover framework based on locality-constrained dictionary learning. Li et al. [102] later introduced a bi-level adversarial network architecture, where the first adversarial scheme was to reconstruct face images, and the second was to maintain face identity.

Unlike the aforementioned facial makeup synthesis methods that treat makeup transfer and removal as separate problems, References [10, 15, 42, 99] performed makeup transfer and makeup removal simultaneously. Inspired by CycleGAN architecture [230], Chang et al. [10] introduced the idea of asymmetric style transfer and a framework for training both the makeup transfer and removal networks together, each one strengthening the other. Li et al. [99] proposed a dual input/output generative adversarial network called BeautyGAN for instance-level facial makeup transfer. More recent work by Chen et al. [15] presented BeautyGlow that decomposed the latent vectors of face images derived from the Glow model into makeup and non-makeup latent vectors. For achieving better makeup and de-makeup performance, Gu et al. [42] focused on local facial details transfer and designed a local adversarial disentangling network that contained multiple and overlapping local adversarial discriminators.

Virtual Try-On. Data-driven clothing image synthesis is a relatively new research topic that is gaining more and more attention. In the following, we review existing methods and datasets for addressing the problem of generating images of people in clothing by focusing on the styles. Han et al. [51] utilized a coarse-to-fine strategy. Their framework, **VIrtual Try-On Network (VITON)**, focused on trying an in-shop clothing image on a person image. It first generated a coarse tried-on result and predicted the mask for the clothing item. Based on the mask and coarse result, a refinement network for the clothing region was employed to synthesize a more detailed result. However, Reference [51] fails to handle large deformation, especially with more texture details, due to the imperfect shape-context matching for aligning clothes and body shape. Therefore, a new model called **Characteristic-Preserving Image-based Virtual Try-On Network (CP-VTON)** [179] was proposed. The spatial deformation can be better handled by a Geometric Matching Module, which explicitly aligned the input clothing with the body shape. The comparisons of VITON and CP-VTON are given in Figure 6(c). There were several improved works [4, 76, 210, 220] based on CP-VTON. Different from the previous works that needed the in-shop clothing image for virtual try-on, FashionGAN [231] and M2E-TON [195] presented target try-on clothing image based on text description and model image respectively. Given an input image and a sentence describing a different outfit, FashionGAN was able to "redress" the person. A segmentation map was first generated with a GAN according to the description. Then, the output image was rendered with another GAN guided by the segmentation map. M2E-TON was able to try on clothing from human$_A$ image to human$_B$ image, and two people can perform in different poses. Considering the runtime efficacy, Issenhuth et al. [74] proposed a parser free virtual try-on network. It designs a teacher-student architecture to free the parsing process during the inference time for improving efficiency.

Viewing the try-on performance from different views is also essential for virtual try-on task, Fit-Me [66] was the first work to do virtual try-on with arbitrary poses in 2019. They designed a coarse-to-fine architecture for both pose transformation and virtual try-on. Further, FashionOn [67] applied the semantic segmentation for detailed part-level learning and focused on refining the facial part and clothing region to present more realistic results. They succeeded in preserving detailed facial and clothing information, perform dramatic posture, and also resolve the human limb

Table 9. Summary of the Benchmark Datasets for Style Transfer Task

Task	Dataset name	Publish time	# of photos	Key features	Sources
Facial Makeup	Liu et al. [121]	2016	2,000	1000 non-makeup faces and 1000 reference faces	N/A
	Stepwise Makeup [184]	2016	1,275	Images in 3 sub-regions (eye, mouth, and skin); labeled with procedures of makeup	N/A
	Beauty [228]	2017	2,002	1,001 subjects, where each subject has a pair of photos being with and without makeup	The Internet
	Before-After Makeup[3]	2017	1,922	961 different females (224 Caucasian, 187 Asian, 300 African, and 250 Hispanic) where one with clean face and another after pro- fessional makeup; annotated with facial attributes	N/A
	Chang et al. [10]	2018	2,192	1,148 non-makeup images and 1,044 makeup images	Youtube makeup tutorial videos
	Makeup Transfer [99]	2018	3,834	1,115 non-makeup images and 2,719 makeup images; assembled with 5 different makeup styles (smoky-eyes, flashy, Retro, Ko- rean, and Japanese makeup styles), varying from subtle to heavy	N/A
	LADN [42]	2019	635	333 non-makeup images and 302 makeup images	The Internet
	Makeup-Wild [85]	2020	772	369 non-makeup images and 403 makeup images	The Internet
Virtual Try-On	LookBook [217]	2016	84,748	The 9,732 top product images are associated with 75,016 fashion model images	Bongjashop, Jogunshop, Stylenanda, SmallMan, WonderPlace
	DeepFashion [123]	2016	78,979	The corresponding upper-body images, sentence descriptions, and human body segmentation maps	Forever21
	CAGAN [77]	2017	15,000	Frontal view human images and paired upper-body garments (pullovers and hoodies)	Zalando.com
	VITON [51]	2018	32,506	Pairs of frontal-view women and top clothing images	N/A
	FashionTryOn [226]	2019	28,714	Pairs of same person with same clothing in 2 different poses and one corresponding clothing image	Zalando.com
	FashionOn [67]	2019	22,566	Pairs of same person with same clothing in 2 different poses and one corresponding clothing image	The Internet, DeepFashion [123]
	Video Virtual Try-on [30]	2019	791 videos	Each video contains 250–300 frame numbers	fashion model catwalk

N/A: there is no reported information to cite.

occlusion problem in CP-VTON. Similar architecture to CP-VTON for virtual try-on with arbitrary poses was presented in Reference [226]. They further made body shape mask prediction at the beginning of the first stage for pose transformation, and, in the second stage, they presented an attentive bidirectional GAN to synthesize the final result. For pose-guided virtual try-on, Dong et al. [29] further improved VITON and CP-VTON, which tackled the virtual try-on for different poses. Han et al. [47] proposed ClothFlow to focus on the clothing regions and model the appearance flow between source and target for transferring the appearance naturally and synthesizing novel result. Beyond the abovementioned image-based virtual try-on works, Dong et al. [30] presented a video virtual try-on system, FWGAN, which learned to synthesize a video of virtual try-on results based on a person image, a target try-on clothing image and a series of target poses. Increasing the image resolution of virtual try-on, References[137, 215] further design novel architectures for achieving multi-layer virtual try-on.

4.1.2 Benchmark Datasets. Style transfer for fashion contains two hot tasks, facial makeup and virtual try-on. Existing makeup datasets applied for studying facial makeup synthesis typically assemble pairs of images for one subject: non-makeup image and makeup image pair. As for virtual try-on, since it is a highly diverse topic, there are several datasets for different tasks and settings. We summarize the datasets for style transfer in Table 9.
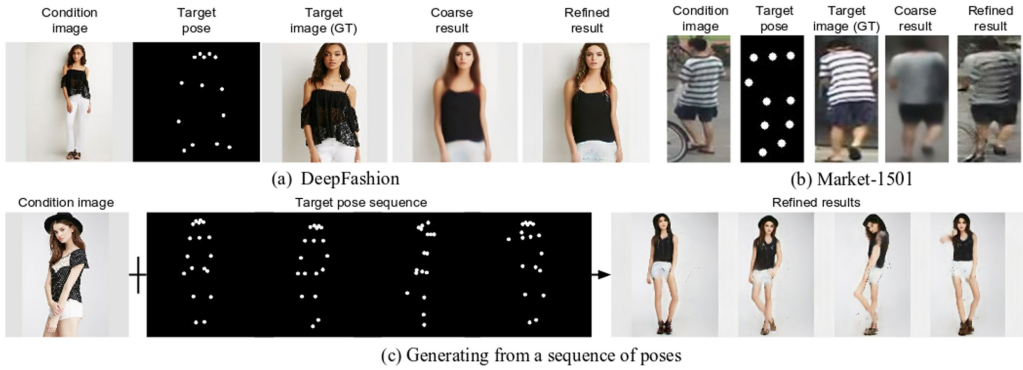
(a) DeepFashion   (b) Market-1501

(c) Generating from a sequence of poses

Fig. 7. Examples of pose transformation [128].

4.1.3 Performance Evaluations. The evaluation for style transfer is generally based on subjective assessment or user study. That is, the participants rate the results into some certain degrees, such as "Very bad," "Bad," "Fine," "Good," and "Very good." The percentages of each degree are then calculated to quantify the quality of results. Besides, there are objective comparisons for virtual try-on, in terms of **inception score (IS)** or **structural similarity (SSIM)**. IS is used to evaluate the synthesis quality of images quantitatively. The score will be higher if the model can produce visually diverse and semantically meaningful images. However, SSIM is utilized to measure the similarity between the reference image and the generated image ranging from zero (dissimilar) to one (similar).

4.2 Pose Transformation

Given a reference image and a target pose only with keypoints, the goal of pose transformation is to synthesize pose-guided person image in different posture while keeping personal information. A few examples of pose transformation are shown in Figure 7. Pose transformation, in particular, is a challenging task, since the input and output are not spatially aligned.

4.2.1 State-of-the-art Methods. A two-stage adversarial network PG^2 [128] achieved an early attempt on this task. A coarse image under the target pose was generated in the first stage and then refined in the second stage. The intermediate results and final results are shown in Figure 7(a) and (b) with two benchmark datasets. However, the results were highly blurred, especially for texture details. To tackle the problem, the affine transform was employed to keep textures in the generated images better. Siarohin et al. [162] designed a deformable GAN, in which the key deformable skip elegantly transforms high-level features for each body part. Similarly, the work in Reference [5] employed body part segmentation masks to guide the image generation. The proposed framework contained four modules, including source image segmentation, spatial transformation, foreground synthesis, and background synthesis that can be trained jointly. Further, Si et al. [161] introduced a multistage pose-guided image synthesis framework, which divided the network into three stages for pose transform in a novel two-dimensional (2D) view, foreground synthesis, and background synthesis.

To break the data limitation of previous studies, Pumarola et al. [145] borrowed the idea from Reference [230] by leveraging cycle consistency. In the meantime, the works in References [34, 129] formulated the problem from the perspective of variational auto-encoder. They can successfully model the body shape; however, their results were less faithful to the appearance of reference

Table 10. Summary of the Benchmark Datasets for Pose Transformation Task

Dataset name	Publish time	# of photos	Key features	Sources
Human3.6M [72]	2014	3.6M	It contains 32 joints for each skeleton	self-collected
Market-1501 [225]	2015	32,668	Multiple viewpoints; the resolution of images is 128×64	A supermarket in Tsinghua University
DeepFashion [123]	2016	52,712	The resolution of images is 256×256	Online shopping sites (Forever21 and Mogujie), Google Images
Balakrishnan et al. [5]	2018	N/A	Action classes: golf swings (136 videos), yoga/workout routines (60 videos), and tennis actions (70 videos)	YouTube

N/A: there is no reported information to cite. M means million.

Table 11. Summary of the Benchmark Datasets for Physical Simulation Task

Dataset name	Publish time	# of photos	Key features	Sources
DeepWrinkles [93]	2018	9,213	Each image contains a colored mesh with 200K vertices.	N/A
Santesteban et al. [156]	2019	7,117	It was simulated with 17 body shapes with only one garment.	SMPL
MGN [7]	2019	N/A	It contains 356 3D scans of people with different body shapes, poses and in diverse clothing.	SMPL+G
RenderPeople	2019	N/A	It consists of 500 high-resolution photogrammetry scans. It was used by PIFu [153, 154] and ARCH [71]	RenderPeople
DeepFashion3D [229]	2020	N/A	It contains 2,078 3D garment models with 10 different clothing categories and 563 garment instances.	Reconstructed from real garments
TailorNet [143]	2020	55,800	It contains 20 aligned real static garments with 1,782 different poses and 9 body shapes.	Simulated by the Marvelous Designer
Sizer [171]	2020	N/A	It includes 100 different subjects with 10 casual clothing classes in various sizes in total of over 2,000 scans.	self-collected

N/A: there is no reported information to cite. M means million.

images, since they generated results from highly compressed features sampled from the data distribution. To improve the appearance performance, Song et al. [166][12] designed a novel pathway to decompose the hard mapping into two accessible subtasks, semantic parsing transformation and appearance generation. First, for simplifying the non-rigid deformation learning, it transformed the posture in semantic parsing maps. Then, synthesizing the semantic-aware human information to the previous synthesized semantic maps formed the realistic final results. Conducting the concept of optical flow, Ren et al. [149] proposed a differentiable global-flow local-attention framework to ensemble the features between source human, source pose, and target pose.

4.2.2 Benchmark Datasets. The benchmark datasets for pose transformation are very limited. The most used two benchmark datasets are Market-1501 [225] and DeepFashion (In-shop clothes retrieval benchmark) [123]. Besides, there is one dataset collected in videos [5]. All of the three datasets are summarized in Table 10.

4.2.3 Performance Evaluations. There are objective comparisons for pose transformation, in terms of IS and SSIM, which have been introduced in Section 4.1.3. Additionally, to eliminate the effect of background, mask-IS and mask-SSIM were proposed in Reference [128].

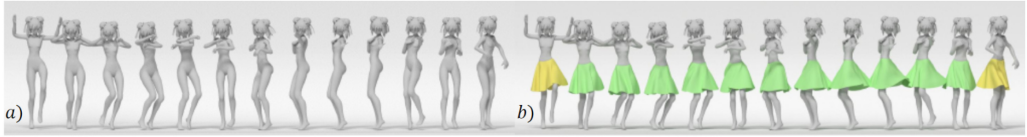
[12]https://github.com/SijieSong/person_generation_spt.

Fig. 8. Examples of physical simulation [191].

4.3 Physical Simulation

For more vivid fashion synthesis performance, physical simulation plays a crucial role. The above-mentioned synthesis works are within the 2D domain, limited in the simulation of the physical deformation, e.g., shadow, pleat, or hair details. For advancing the synthesis performance with dynamic details (drape, or clothing-body interactions), there are physical simulation works based on 3D information. Take Figure 8 for example. Based on a body animation sequence (a), the shape of the target garment and keyframes (marked by yellow), Wang et al. [191] learned the intrinsic physical properties and simulated to other frames with different postures that are shown in (b).

4.3.1 State-of-the-art Methods. The traditional pipeline for designing and simulating realistic clothes is to use computer graphics to build 3D models and render the output images [44, 144, 180, 211]. For example, Wang et al. [180] developed a piecewise linear elastic model for learning both stretching and bending in real cloth samples. For learning the physical properties of clothing on different human body shapes and poses, Guan et al. [44] designed a pose-dependent model to simulate clothes deformation. For simulating regular clothing on fully dressed people in motion, Pons-Moll et al. [144] designed a multi-part 3D model called ClothCap. First, it separated different garments from the human body for estimating the clothed body shape and pose under the clothing. Then, it tracked the 3D deformations of the clothing over time from 4D scans to help simulate the physical clothing deformations in different human posture. To enhance the realism of the garment on human body, Lähner et al. [93] proposed a novel framework, which composed of two complementary modules: (1) A statistical model learned to align clothing templates based on 3D scans of clothed people in motion and a linear subspace model factored out the human body shape and posture. (2) A cGAN added high-resolution geometric details to normal maps and simulated the physical clothing deformations.

For advancing the physical simulation with non-linear deformations of clothing, Santesteban et al. [156] presented a two-level learning-based clothing animation method for highly efficient virtual try-on simulation. There were two fundamental processes: It first applied global body-shape-dependent deformations to the garment and then predicted dynamic wrinkle deformations based on the body shape and posture. Further, Wang et al. [191] introduced a semi-automatic method for authoring garment animation, which first encoded essential information of the garment shape and based on the intrinsic garment representation and target body motion, it learned to reconstruct garment shape with physical properties automatically. Based only on a single-view image, Yang et al. [211] proposed a method to recover a 3D mesh of garment with the 2D physical deformations. Given a single-view image, a human-body database, and a garment-template database as input, it first preprocessed with garment parsing, human body reconstruction, and features estimation. Then, it synthesized the initial garment registration and garment parameter identification for reconstructing body and clothing models with physical properties. Besides, Yu et al. [221] enhanced the simulation performance with a two-step model called SimulCap, which combined the benefits of capture and physical simulation. The first step aimed to get a multi-layer avatar via double-layer reconstruction and multi-layer avatar generation. Then, it captured the human performance by body tracking and cloth tracking for simulating the physical clothing-body interactions.
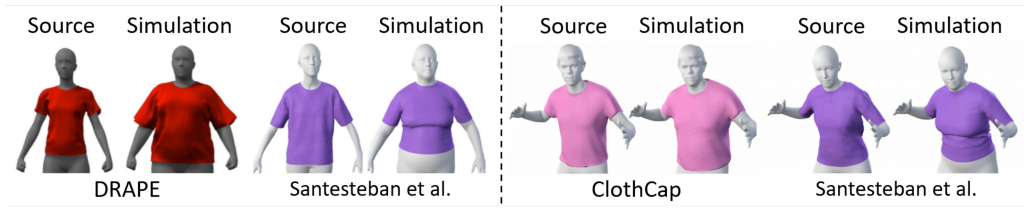
Fig. 9. Left comparison is between DRAPE [44] and Santesteban et al. [156], while the right one compares ClothCap [144] and Santesteban et al. [156]. Both are given a source and simulate the physical clothing deformation in different body shapes.

4.3.2 Benchmark Datasets. Datasets for physical simulation are different from other fashion tasks, since the physical simulation is more related to computer graphics than computer vision. Here, we discuss the types of input data used by most physical simulation works. Physical simulation working within the fashion domain focus on clothing-body interactions, and datasets can be categorized into real data and created data. For an example of real data, for each type of clothing, Reference [93] captured 4D scan sequences at 60 fps in motion and dressed in a full-body suit. As for created data, Reference [156] was based on one garment to create dressed character animations with diverse motions and body shapes, and it can be applied to other garments. We list the benchmark datasets for physical simulation in Table 11.

4.3.3 Performance Comparison. There are limited quantitative comparisons between physical simulation works. Most of them tend to calculate the qualitative results only within their work (e.g., per-vertex mean error) or show the vision comparison with state-of-the-art methods. Take the comparison done by Santesteban et al. [156] for example in Figure 9.

5 FASHION RECOMMENDATION

Dressing well will make everyone look and feel confident and energetic. Tom Ford, a fashion expert and film director, once even said, "dressing well is a form of good manners," but not everyone is a natural-born fashion stylist. In support of this need, fashion recommendation has attracted increasing attention, given its ready applications to online shopping for fashion products. Relevant literature on the research progress of fashion recommendation falls into three main tasks: fashion compatibility, outfit matching, and hairstyle suggestion.

5.1 Fashion Compatibility

Fashion recommendation works based on fashion compatibility, which performs how well items of different types can collaborate to form fashionable outfits. Also, it is worth mentioning that the main difference between fashion item retrieval (Section 2.3) and fashion recommendation (Section 5) is that the former learns the visual similarity between the same clothing type. In contrast, the latter learns both visual similarity and visual compatibility between different clothing types.

5.1.1 State-of-the-art Methods. Veit et al. introduced **Conditional Similarity Networks (CSNs)** [174], which learned non-linear feature embeddings that incorporated various notions of similarity within a shared embedding using a shared feature extractor. The CSNs addressed the issue of a standard triplet embedding that treated all triplets equally and ignored the sources of similarity. Song et al. [168] integrated visual and contextual modalities of fashion items by employing the autoencoder neural model to seek the non-linear latent compatibility space. Following Reference [168], Song et al. [167] integrated fashion domain knowledge to the neural networks to boost the performance. Vasileva et al. [173] presented to learn an image embedding that respected

Table 12. Summary of the Benchmark Datasets for Fashion Compatibility Task

Dataset name		Publish time	# of outfits	# of item categories	Key features	Sources
FashionVC [168]		2017	20,726	2	Annotated with title, category, and description	Polyvore.com
Vasileva et al. [173]		2018	68,306	19	Annotated with outfit and item ID, fine-grained item type, title, and text descriptions	Polyvore.com
Style for Body Shape [58]	Stylish Celeb-rity Names	2018	N/A	N/A	270 names of the top stylish female celebrities	Ranker.com, fashion magazine sites
	Body Measurements				Body measurements of 3,150 female celebrities	Bodymeasurements.org
	Stylish Celeb-rity Images		347,948		Annotated with clothing categories and celebrity names	Google Image Search
PolyVore-T [189]		2019	19,835	5	Categories includes top, bottom, shoes, bag, and accessory	Dataset collected by [50] from Polyvore
IQON3000 [169]		2019	308,747	6	Categories contains top, bottom, shoes, accessory, dress and tunic, and coat. The outfits within this dataset was created by 3,568 users	The fashion web service IQON

N/A: there is no reported information to cite.

item type. They first learned a single, shared embedding space to measure item similarity, then projected from that shared embedding to subspaces identified by type. For learning the compatibility between clothing styles and body shapes, Hidayati et al. [57, 58] exploited human body measurements and images of stylish celebrities. They presented a body shape calculator to determine the type of body shape based on a set of body measurements, and a style propagation mechanism to model the correlation between body shapes and clothing styles by constructing graphs of images based on the body shape information and the visual appearance of clothing items, respectively. For using category complementary relations to model compatibility, Yang et al. [214] proposed a translation-based neural fashion compatibility model that contained three parts: (1) first mapped each item into a latent space via two CNN for visual and textual modality, (2) encoded the category complementary relations into the latent space, and (3) minimized a margin-based ranking criterion to optimize both item embeddings and relation vectors jointly.

For making the fashion compatibility task more user-friendly, Wang et al. [189] introduced a diagonal process for giving information about which item made the outfit incompatible. They presented an end-to-end multi-layered comparison network to predict the compatibility between different items at different layers and use the backpropagation gradient for diagnosis. Hsiao et al. [65] proposed Fashion++ to make minimal adjustments to a full-body clothing outfit that have a maximal impact on its fashionability. Besides, Song et al. [169] took user preferences into account to present a personalized compatibility modeling scheme GP-BPR. It utilized two components, general compatibility modeling and personal preference modeling, for evaluating the item–item and user–item interactions, respectively.

5.1.2 Benchmark Datasets. The most used source for fashion compatibility datasets is the Polyvore fashion website. It is an online shopping website, where the fashion items contain rich annotations, e.g., clothing color, text description, and multi-view outfit images. We list the benchmark datasets for fashion compatibility in Table 12.

5.1.3 Performance Evaluations. For measuring the performance of fashion compatibility works, **area under the receiver operating characteristic curve (AUC)** is the most used metric. AUC measures the probability that the evaluated work would recommend higher compatibility for positive set than negative set. The AUC scores range between 0 and 1.

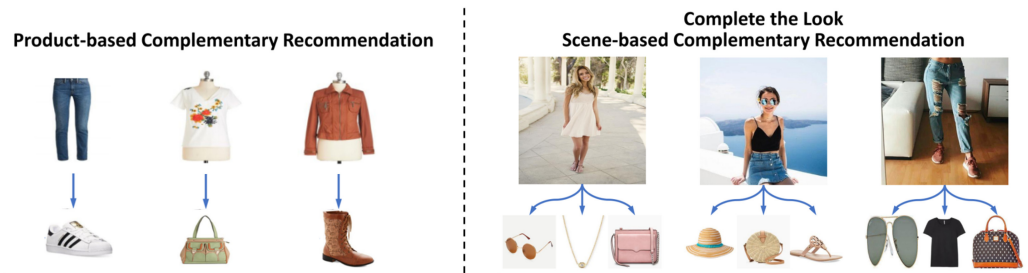Fig. 10. Examples of outfit matching task [64].

Fig. 11. A comparison between product-based and scene-based complementary recommendation [87].

5.2 Outfit Matching

Each outfit generally involves multiple complementary items, such as tops, bottoms, shoes, and accessories. A key to a stylish outfit lies in the matching fashion items, as illustrated in Figure 10. However, generating harmonious fashion matching is challenging due to three main reasons. First, the fashion concept is subtle and subjective. Second, there are a large number of attributes for describing fashion. Third, the notion of fashion item compatibility generally goes across categories and involves complex relationships. In the past several years, this problem has attracted a great deal of interest, resulting in a long list of algorithms and techniques.

5.2.1 State-of-the-art Methods. Fashion recommendation for outfit matching was initially introduced by Iwata et al. [75]. They proposed a probabilistic topic model for learning information about fashion coordinates. Liu et al. [117] explored occasion-oriented clothing recommendation by considering the *wearing properly* and *wearing aesthetically* principles. They adopted a unified latent SVM to learn the recommendation model that incorporates clothing matching rules among visual features, attributes, and occasions. A similar idea of location-oriented recommendation system was proposed by Zhang et al. [222]. They considered the visual match between the foreground clothing and the background scenery and proposed a hybrid multi-label convolutional neural network combined with the SVM (mCNN-SVM), which captured the uneven distribution of clothing attributes and explicitly formulated the correlations between clothing attributes and location attributes. More recently, Kang et al. [87] introduced "Complete the Look" aiming at recommending fashion items that go well with the given scene. They measured both global compatibility (i.e., the compatibility between the scene and product images) and local compatibility (i.e., the compatibility between every scene patch and product image) via Siamese networks and category-guided attention mechanisms. The comparison of the product-based and the scene-based complementary recommendation is shown in Figure 11.

A line with metric-based works then proposed to model item-to-item compatibility based on co-purchase behavior. Veit et al. [175] utilized the co-purchase data from Amazon.com to train

a Siamese CNN to learn style compatibility across categories and used a robust nearest neighbor retrieval to generate compatible items. The study in Reference [136] modeled human preference to uncover the relationships between the appearances of pairs of items by Low-rank Mahalanobis Transform that mapped compatible items to embeddings close in the latent space. He and McAuley [54] later extended the work of Reference [136] by combining visual and historical user feedback data. The proposed study incorporated visual signals into Bayesian Personalized Ranking with Matrix Factorization as the underlying predictor. Besides, Hu et al. [68] addressed the personalized issue by utilizing a functional tensor factorization method to model the user–item and item–item interactions.

Most previous works mainly focused on top-bottom matching. However, an outfit generally includes more items, such as shoes and bags. To address this issue, Chen and He [19] extended the traditional triplet neural network, which usually receives three instances, to accept multiple instances. A mixed-category metric learning method was proposed to adapt the multiple inputs. A similar idea was also employed in Reference [160], where the proposed generator, referred to as metric-regularized cGAN, was regularized by a projected compatibility distance function. It ensured the compatible items were closer in the learned space compared to the incompatible ones.

As for methods with the end-to-end framework, Li et al. [100] formulated the problem as a classification task, where a given outfit composition was labeled as a popular or unpopular one. They designed a multi-modal multi-instance model, that exploited images and meta-data of fashion items, and information across fashion items, to evaluate instance aesthetics and set compatibility simultaneously. Inspired by image captioning of Reference [28], Han et al. [50] built a model based on **bidirectional LSTM (Bi-LSTM)** to treat an outfit as a sequence of fashion items and each item in the outfit as a timestep. A Bi-LSTM model was then utilized to predict the next item conditioned on previously seen ones, where the objective was to maximize the total probability of the given positive sequence. The model was able to achieve three tasks, i.e., fill-in-the-blank (fashion item recommendation given an existing set), outfit generation conditioned on users' text/image inputs, and outfit compatibility prediction. Another work [64] also borrowed the idea from natural language processing, which meant that an outfit was regarded as a "document," an inferred clothing attribute was taken to be a "word," and a clothing style was referred to the "topic." The problem of outfit matching was formulated with the topic model. The combinations similar to previously assembled outfits should have a higher probability, which can be employed as the prediction for compatibility, and then solve the outfit matching problem.

For building the bridge between fashion compatibility and personalized preference in outfit matching tasks, there are a few methods for this goal. A personalized clothing recommendation system, namely i-Stylist that retrieved clothing items through the analysis of user's images, was developed in Reference [155]. The i-Stylist organized the deep learning features and clothing properties of user's clothing items as a fully connected graph. The user's personalized graph model later derived the probability distribution of the likability of an item in shopping websites. Dong et al. [33] took user preference and body shape into account for measuring the user-garment compatibility to deal with personalized capsule wardrobe creation task. They introduced an optimization-based framework with dual compatibility modeling, which can both evaluate the garment-garment compatibility and user-garment compatibility. Besides, Yu et al. [219] worked for synthesizing new items automatically for recommendation. Given a query item, the personalized fashion design network they proposed would generate a fashion item for the specific user based on fashion compatibility and user preference. Furthermore, Chen et al. [22] presented an industrial-scale **Personalized Outfit Generation (POG)** model. They deployed POG on platform *Dida* in *Alibaba* to recommend personalized fashionable outfits for users. For providing more concrete recommendation for users, Hou et al. [62] proposed a semantic attribute explainable recommender system to

Table 13. Summary of the Benchmark Datasets for Outfit Matching Task

Dataset name	Publish time	# of outfits	# of item categories	Key features	Sources
What-to-Wear [117]	2012	24,417	2	Occasion-oriented work; Annotated with full-body, upper-body, or lower-body	Online shopping photo sharing sites
Styles and Substitutes [136]	2015	773,465	N/A	Annotated with 4 categories of relationship: (1) users who viewed X also viewed Y, (2) users who viewed X eventually bought Y, (3) users who bought X also bought Y, (4) users bought X and Y simultaneously	Amazon.com
Hu et al. [68]	2015	28,417	3	Annotated with categories, names, and descriptions	Polyvore.com
He et al. [54]	2016	598.353	N/A	Annotated with users' review histories	Amazon.com, Tradesy.com
Journey Outfit [222]	2017	3,392	N/A	Location-oriented work; Annotated with 14 travel destinations	Online travel review sites
Li et al. [100]	2017	195,262	4	Annotated with title, category, and number of likes	Polyvore.com
Han et al. [50]	2017	21,889	8	Each item contains a pair – product image and a corre sponding text description	Polyvore.com
Hsiao et al. [64]	2018	3,759	≥ 2	Annotated with meta-labels, e.g., season (winter, spring, summer, fall), occasion (work, vacation), and function (date, hike)	Polyvore.com
Fashion Collocation[19]	2018	220,000	5	Annotated with independent and ready for wearing (off- body module) or dependent (on-body module), or a bounding box	Chictopia.com, Wear.net, FashionBeans.com
Pinterest's Shop The Look [87]	2019	38,111	10	Annotated with categories of shoes, tops, pants, hand- bags, coats, sunglasses, shorts, skirts, earrings, necklaces	Pinterest.com
bodyFashion [33]	2019	75,695	N/A	It contains 11,784 users with his/her latest historical pur- chase records in total of 116,532 user–item records	Amazon.com
Yu et al. [219]	2019	208,814	N/A	It contains 797 users with 262 outfits and each outfit with 2 items, i.e., a top and a bottom	Polyvore.com
POG [33]	2019	1.01M	80	It is composed of 583,000 individual items	Taobao.com, iFashion

N/A: There is no reported information to cite; M means million.

not only recommend for personalized compatible items but also explain the reason why the system recommends it.

5.2.2 Benchmark Datasets. Since different papers are under various settings and most of the datasets for outfit matching are not publicly available, almost every work collected their own dataset. We list the benchmark datasets for outfit matching in Table 13. Note that the outfit database in References [50, 54, 64, 68, 87, 100, 136] comprises the images of each single item, while in References [19, 75, 117] consists of an outfit in a single image.

5.2.3 Performance Evaluations. As the evaluation protocol for fashion compatibility (Section 5.1.3), AUC is the most used metric for outfit matching methods. While some methods are also evaluated with NDCG, i.e., References [62, 68, 117], and fill-in-the-blank accuracy, i.e., Reference [22, 50]. Unfortunately, there is no unified benchmark for outfit matching, both in datasets and evaluation metrics. Therefore, we are unable to give a comparison of different methods.

5.3 Hairstyle Suggestion

Hairstyle plays an essential role in physical appearance. People can look completely different with a different hairstyle. The right hairstyle can enhance the best facial features while concealing the

Table 14. Summary of the Benchmark Datasets for Hairstyle Suggestion Task

Dataset name	Publish time	# of photos	Key features	Sources
Yang et al. [213]	2012	84	Frontal face photos	Hair stylists from three salons
Beauty e-Experts [118]	2013	1,605	It consists of 20 hair color classes, three different hair length attributes (long, medium, or short), three differ- ent hair shape attributes (straight, curled, or wavy), and two kinds of hair volume (dense or normal)	Professional hairstyle and makeup websites (e.g., Stylebistro.com)
Hairstyle30k [216]	2017	30,000	It contains 64 different hairstyles	Various web search engines (e.g., Google, Flicker, and Bing)

flaws, bringing out natural beauty and style. However, choosing the right hairstyle requires careful consideration as not all hairstyles suit all facial features.

5.3.1 State-of-the-art Methods. In recent years, many papers have been published related to hairstyles focusing on how to model and render hairstyles with computer graphics [9, 35, 141, 157, 202] or on how to segment hair automatically [147, 151, 182, 183, 203]. Only a few studies have been devoted to finding the right hairstyle to suit one's face. In the following, we review the literature concerning hairstyle recommendation.

The pioneering work in Reference [213] suggested suitable hairstyles for a given face by learning the relationship between facial shapes and successful hairstyle examples. The proposed example-based framework consisted of two steps: the statistical learning step and the composition step. The purpose of the statistical learning step was to find the most suitable hairstyle through Bayesian inference-based model that estimated the probability distributions of hairstyles to a face image. They proposed to use the ratio of line segments as the feature vector for characterizing the shape of each face, and α-matting-based method to indicate hair area in the image. The most suitable hairstyle obtained from the statistical learning step was further superimposed over a given face image to output the face in a suitable hairstyle.

Liu et al. later developed the Beauty e-Experts system [114] to automatically recommend the most suitable facial hairstyle and makeup, and synthesize the visual effects. They proposed to use the extracted facial and clothing features to simultaneously learn multiple tree-structured super-graphs to collaboratively model the underlying relationships among the high-level beauty attributes (e.g., hair length and eye shadow color), mid-level beauty-related attributes (e.g., eye shape and mouth width), and low-level image features. Besides, they also proposed a facial image synthesis module to synthesize the beauty attributes recommended by the multiple tree-structured super-graphs model.

5.3.2 Benchmark Datasets. Table 14 provides benchmark datasets for assessing the performance of hairstyle suggestion methods. It is worth mentioning that Hairstyle30k [216] is by far the largest dataset for hairstyle-related problems thought the proposed method is not for hairstyle suggestion.

5.3.3 Performance Evaluations. Yang et al. [213] conducted a user study to evaluate the effectiveness of their proposed system, while Liu et al. [114] computed the NDCG that measures how close the ranking of the top-k recommended styles is to the optimal ranking. However, we are unable to give comparisons on different hairstyle suggestion methods due to inconsistent benchmarks for different papers.

6 APPLICATIONS AND FUTURE WORK

The future of the fashion world will be shaped in large part by advancements in the technology, which is currently creeping into the creative domains, because it is starting to mimic human neurons. In the following, we discuss emerging uses of fashion technology in some application areas and future work that is needed to achieve the promised benefits.

6.1 Applications

The most popular AI application from the top global apparel industry leaders currently implementing AI appears to be AI chatbots, also called smart assistants, which is used to interact with their customers. The common use-cases that are covered are as follows: (1) Responding to customer service inquiries and providing suggestions related to product searches through a social media messaging platform, e.g., "Dior Insider"; (2) helping customers navigate products online or in-store to find product(s) that align with their interests, e.g., "Levi's Virtual Stylist," "VF Corporation," "Macy's On Call," and "Nordstrom"; and (3) virtual assistant to encourage exercise/behavior adherence, e.g., "Nike on Demand."

Moving forward, AI technology will have explosive growth to power fashion industry. In addition to connecting with the customers with the use of AI chatbots, we identify there are four other ways that AI is transforming the future of fashion and beauty, which include the following: (1) *Improving product discovery.* Visual search makes it easier for shoppers to find, compare, and purchase products by taking or uploading a photo. One example is Google Lens,[13] which allows mobile users to perform searches for similar styles through Google Shopping from the photos they take. (2) *Tailor recommendation.* To keep costs low, brands need to better predict customer preferences by gathering and analyzing purchase behavior, customer profile, as well as customer feedback. Using this data alongside AI and machine learning allows fashion retailers to deliver a personalized selection of clothes to customers. Stitch Fix[14] is one of the most popular AI fashion companies in this category. (3) *Reducing product return.* Customers have more options to choose from than ever when it comes to making purchases. To gain the loyalty of the consumers, one recent focus of the fashion retailer has been to extend an excellent customer service experience not only at the point of purchase but at the point of return as well. AI technology has the power to better engage customers with the personalized shopping experience that leads them to make more informed and confident purchase decisions, which in turn helps retailers lower return rates. Sephora[15] is an example of a retailer that has developed this strategy. (4) *Powering productivity and creativity.* The promise of AI for fashion brands that can marry design creativity with digital innovation has a powerful competitive advantage in the market. The AI technology enables fashion brands to sift through consumer data to gain insights on which product features their customers are likely to prefer.

6.2 Breaking the Limitations of Annotation

Large-scale data annotation with high quality is indispensable. However, current studies are mainly based on relatively small-scale datasets, which is usually constrained by annotation costs and efforts. Faced with such enormous amounts of fashion- and beauty-related data, how to generate high coverage and high precision annotations to considerably reduce the cost while preserving quality remains a hot issue. Therefore, more efforts in the development of cost-effective annotations approach on fashion- and beauty-related data are necessary to address the problem.

---

[13]https://lens.google.com.
[14]https://www.stitchfix.com.
[15]https://www.sephora.com.

6.3 Overcoming Industry Challenges

There are still many challenges in adopting fashion and beauty technologies in industry, because real-world fashion and beauty are much more complex and strict than in the experiments. The main issue is related to system performance, which is still far from human performance in real-world settings. The demand for a more robust system consequently grows with it. Accordingly, it is crucial to continue to pay attention to handling data bias and variations to improve the true-positive rate while maintaining a low false-positive rate. Moreover, with the rising interest in mobile applications, there is a definite need to perform the task in a light but timely fashion. It is thus also beneficial to consider how to optimize the model to achieve higher performance and better computation efficiency.

7 CONCLUSION

With the significant advancement of information technology, research in CV and its applications in fashion have become a hot topic and received a great deal of attention. Meanwhile, the enormous amount of data generated by social media platforms and e-commerce websites provide an opportunity to explore knowledge relevant to support the development of intelligent fashion techniques. Arising from the above, there has much CV-based fashion technology been proposed to handle the problems of fashion image detection, analysis, synthesis, recommendation, and its applications. The long-standing semantic gap between computable low-level visual features and high-level intents of customers now seems to be narrowing down. Despite recent progress, investigating and modeling complex real-world problems when developing intelligent fashion solutions remain challenging. Given the enormous profit potential in the ever-growing consumer fashion and beauty industry, the studies on intelligent fashion-related tasks will continue to grow and expand.

REFERENCE

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

[2] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. 2017. Fashion forward: Forecasting visual style in fashion. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 388–397. https://doi.org/10.1109/ICCV.2017.50

[3] T. Alashkar, S. Jiang, S. Wang, and Y. Fu. 2017. Examples-rules guided deep neural network for makeup recommendation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 31. 941–947.

[4] K. Ayush, S. Jandial, A. Chopra, and B. Krishnamurthy. 2019. Powering virtual try-on via auxiliary human segmentation learning. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[5] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. 2018. Synthesizing images of humans in unseen poses. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8340–8348.

[6] L. Banica, D. Pirvu, and A. Hagiu. 2014. Neural networks based forecasting for romanian clothing sector. In *Intelligent Fashion Forecasting Systems: Models and Applications (IFFS)*. 161–194.

[7] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. 2019. Multi-Garment net: Learning to dress 3D people from images. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 5420–5430.

[8] F. C. Heilbron, B. Pepik, Z. Barzelay, and M. Donoser. 2019. Clothing recognition in the wild using the amazon catalog. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 3145–3148.

[9] M. Chai, T. Shao, H. Wu, Y. Weng, and K. Zhou. 2016. AutoHair: Fully automatic hair modeling from a single image. *ACM Transactions on Graphics* 35, 4 (2016), 116:1–116:12.

[10] H. Chang, J. Lu, F. Yu, and A. Finkelstein. 2018. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 40–48.

[11] Y. Chang, M. Chuang, S. Hung, S. Shen, and B. Chu. 2003. A Kansei Study on the Style Image of Fashion Design. *Asian Design Conference* (2003).

[12] Y.-T. Chang, W.-H. Cheng, B. Wu, and K.-L. Hua. 2017. Fashion world map: Understanding cities through streetwear fashion. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 91–99.

[13] F. Chen and D. Zhang. 2016. Combining a causal effect criterion for evaluation of facial attractiveness models. *Neurocomputing* 177, C (2016), 98–109.

[14] H. Chen, A. Gallagher, and B. Girod. 2012. Describing clothing by semantic attributes. In *Proc. European Conference on Computer Vision (ECCV)*. 609–623.

[15] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng. 2019. BeautyGlow: On-demand makeup transfer framework with reversible generative 2 network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10042–10050.

[16] I. Chen and C. Lu. 2017. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Comput. Appl.* 28, 9 (2017), 2633–2647.

[17] K. Chen, K. Chen, P. Cong, H. Hsu, and J. Luo. 2015. Who are the Devils Wearing Prada in New York City? In *Proceedings of the 23rd ACM International Conference on Multimedia (MM'15)*. 177–180.

[18] K. Chen and J. Luo. 2017. When fashion meets big data: Discriminative mining of best selling clothing features. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW)*. 15–22.

[19] L. Chen and Y. He. 2018. Dress fashionably: Learn fashion collocation with deep mixed-category metric learning. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 2103–2110.

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

[21] Q. Chen, G. Wang, and C. L. Tan. 2013. Modeling fashion. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

[22] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, and B. Zhao. 2019. POG: Personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2662–2670.

[23] W.-H. Cheng, J. Jia, S. Liu, J. Fu, J. Liu, S. C. Hidayati, J. Tseng, and J. Huang. 2019. Perfect Corp. Challenge 2019: Half Million Beauty Product Image Recognition. https://challenge2019.perfectcorp.com/.

[24] Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua. 2017. Video2Shop: Exact matching clothes in videos to online shopping images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4048–4056.

[25] T. Choi, C. Hui, S. Ng, and Y. Yu. 2012. Color trend forecasting of fashionable products with very few historical data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1003–1010.

[26] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion. 2017. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2268–2274.

[27] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[28] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2625–2634.

[29] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin. 2019. Towards multi-pose guided virtual try-on network. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 9026–9035.

[30] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin. 2019. FW-GAN: Flow-navigated warping GAN for video virtual try-on. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 1161–1170.

[31] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. 2014. Towards unified human parsing and pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4321–4328.

[32] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. 2013. A deformable mixture parsing model with parselets. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 3408–3415.

[33] X. Dong, X. Song, F. Feng, P. Jing, X. Xu, and L. Nie. 2019. Personalized capsule wardrobe creation with garment and user modeling. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 302–310.

[34] P. Esser, E. Sutter, and B. Ommer. 2018. A variational U-Net for conditional appearance and shape generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8857–8866.

[35] Y. Fei, H. T. Maia, C. Batty, C. Zheng, and E. Grinspun. 2017. A multi-scale model for simulating liquid-hair interactions. *ACM Transactions on Graphics* 36, 4 (2017), 56:1–56:17.

[36] L. Gao, W. Li, Z. Huang, D. Huang, and Y. Wang. 2018. Automatic facial attractiveness prediction by deep multi-task learning. In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*. 3592–3597.

[37] N. Garcia and G. Vogiatzis. 2017. Dress like a star: Retrieving fashion products from videos. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2293–2299.

[38] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo. 2019. A versatile benchmark for detection, pose estimation, segmentation and Re-Identification of clothing images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5337–5345.

[39] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin. 2019. Graphonomy: Universal human parsing via graph transfer learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7450–7459.

[40] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. 2018. Instance-level human parsing via part grouping network. In *Proc. European Conference on Computer Vision (ECCV)*. 805–822.

[41] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. 2017. Look into Person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 932–940.

[42] Q. Gu, G. Wang, M. T. Chiu, Y. Tai, and C. Tang. 2019. LADN: Local adversarial disentangling network for facial makeup and de-makeup. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 10481–10490.

[43] X. Gu, Y. Wong, P. Peng, L. Shou, G. Chen, and M. S. Kankanhalli. 2017. Understanding fashion trends from street photos via neighbor-constrained embedding learning. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 190–198.

[44] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. 2012. DRAPE: DRessing Any PErson. *ACM Transactions on Graphics* 31, 4 (July 2012), 35:1–35:10.

[45] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. 2015. Where to buy It: Matching street clothing photos in online shops. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 3343–3351.

[46] Y. Ha, S. Kwon, M. Cha, and J. Joo. 2017. Fashion conversation data on instagram. In *Proc. AAAI Conference on Web and Social Media (ICWSM)*. 418–427.

[47] X. Han, X. Hu, W. Huang, and M. R. Scott. 2019. ClothFlow: A flow-based model for clothed person generation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 10471–10480.

[48] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 1472–1480.

[49] X. Han, Z. Wu, W. Huang, M. R. Scott, and L. S. Davis. 2019. Compatible and diverse fashion image inpainting. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 4481–4491.

[50] X. Han, Z. Wu, Y. Jiang, and L. S. Davis. 2017. Learning fashion compatibility with bidirectional LSTMs. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 1078–1086.

[51] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S Davis. 2018. VITON: An image-based virtual Try-on network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7543–7552.

[52] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2961–2969.

[53] R. He and J. McAuley. 2016. Ups and Downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*. 507–517.

[54] R. He and J. McAuley. 2016. VBPR: Visual bayesian personalized ranking from implicit feedback. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 144–150.

[55] S. C. Hidayati, Y.-L. Chen, C.-L. Yang, and K.-L. Hua. 2017. Popularity Meter: An influence- and aesthetics-aware social media popularity predictor. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 1918–1923.

[56] S. C. Hidayati, W.-H. Cheng, and K.-L. Hua. 2012. Clothing genre classification by exploiting the style elements. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. 1137–1140.

[57] S. C. Hidayati, T. W. Goh, J.-S. Gary Chan, C.-C. Hsu, J. See, L. Wong, K.-L. Hua, Y. Tsao, and W.-H. Cheng. 2021. Dress with Style: Learning style from joint deep embedding of clothing styles and body shapes. *IEEE Transactions on Multimedia* 23 (2021), 365–377. https://doi.org/10.1109/TMM.2020.2980195

[58] S. C. Hidayati, C.-C. Hsu, Y.-T. Chang, K.-L. Hua, J. Fu, and W.-H. Cheng. 2018. 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 (MM'18)*. 438–446.

[59] S. C. Hidayati, K.-L. Hua, W.-H. Cheng, and S.-W. Sun. 2014. What are the fashion trends in New York?. In *Proceedings of the 22th ACM International Conference on Multimedia (MM'14)*. 197–200.

[60] S. C. Hidayati, K.-L. Hua, Y. Tsao, H.-H. Shuai, J. Liu, and W.-H. Cheng. 2019. Garment detectives: Discovering clothes and its genre in consumer photos. In *International Conference on Multimedia Information Processing and Retrieval (MIPR)*. 471–474.

[61] S. C. Hidayati, C.-W. You, W.-H. Cheng, and K.-L. Hua. 2018. Learning and recognition of clothing genres from full-body images. *IEEE Transactions on Cybernetics* 48, 5 (May 2018), 1647–1659.

[62] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu. 2019. Explainable fashion recommendation: A semantic attribute region guided approach. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*. 4681–4688.

[63] W. Hsiao and K. Grauman. 2017. Learning the latent "Look": Unsupervised discovery of a style-coherent embedding from fashion images. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 1109–1112.

[64] W. Hsiao and K. Grauman. 2018. Creating capsule wardrobes from fashion images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7161–7170.

[65] W. Hsiao, I. Katsman, C. Wu, D. Parikh, and K. Grauman. 2019. Fashion++: Minimal edits for outfit improvement. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 5047–5056.

[66] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, and W.-H. Cheng. 2019. Fit-me: Image-based virtual try-on with arbitrary poses. In *Proc. IEEE International Conference on Image Processing (ICIP)*. 4694–4698.

[67] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng. 2019. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 275–283.

[68] Y. Hu, X. Yi, and L. S. Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23th ACM International Conference on Multimedia (MM'15)*. 129–138.

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

[70] J. Huang, W. Xia, and S. Yan. 2014. Deep search with attribute-aware deep network. In *Proceedings of the 22th ACM International Conference on Multimedia (MM'14)*. 731–732.

[71] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. 2020. ARCH: Animatable reconstruction of clothed humans. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3093–3102.

[72] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. 2014. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 7 (2014), 1325–1339.

[73] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. 2017. Image-to-Image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976.

[74] T. Issenhuth, J. Mary, and C. Calauzènes. 2020. Do not mask what you do not need to mask: A parser-free virtual try-on. In *Proc. European Conference on Computer Vision (ECCV)*.

[75] T. Iwata, S. Watanabe, and H. Sawada. 2011. Fashion coordinates recommender system using photographs from fashion magazines. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-11)*. 2262–2267.

[76] H. J. Lee, R. Lee, M. Kang, M. Cho, and G. Park. 2019. LA-VITON: A network for looking-attractive virtual try-on. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 3129–3132.

[77] N. Jetchev and U. Bergmann. 2017. The conditional analogy GAN: Swapping fashion articles on people images. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2287–2292.

[78] W. Ji, X. Li, Y. Zhuang, O. El Farouk Bourahla, Y. Ji, S. Li, and J. Cui. 2018. Semantic locality-aware deformable network for clothing segmentation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*. 764–770.

[79] J. Jia, J. Huang, G. Shen, T. He, Z. Liu, H. Luan, and C. Yan. 2016. Learning to appreciate the aesthetic effects of clothing. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 1216–1222.

[80] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie. 2020. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *Proc. European Conference on Computer Vision (ECCV)*.

[81] S. Jiang and Y. Fu. 2017. Fashion style generator. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*. 3721–3727.

[82] S. Jiang, M. Shao, C. Jia, and Y. Fu. 2016. Consensus style centralizing auto-encoder for weak style classification. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 1223–1229.

[83] S. Jiang, Y. Wu, and Y. Fu. 2016. deep bi-directional cross-triplet embedding for cross-domain clothing retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia (MM'16)*. 52–56.

[84] S. Jiang, Y. Wu, and Y. Fu. 2018. Deep bidirectional cross-triplet embedding for online clothing shopping. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 1, Article 5 (January 2018), 22 pages.

[85] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, Ji. Feng, and S. Yan. 2020. PSGAN: Pose and expression robust spatial-aware GAN for customizable makeup transfer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5194–5202.

[86] Y. Kalantidis, L. Kennedy, and L. Li. 2013. Getting the Look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ACM International Conference on Multimedia Retrieval (ICMR)*. 105–112.

[87] W.-C. Kang, E. Kim, J. Leskovec, C. Rosenberg, and J. McAuley. 2019. Complete the look: Scene-based complementary product recommendation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10532–10541.

[88] M. Kaya, E. Ye şil, M. F. Dodurka, and S. Sıradağ. 2014. Fuzzy forecast combining for apparel demand forecasting. In *Intelligent Fashion Forecasting Systems: Models and Applications (IFFS)*. 123–146.

[89] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. 2014. Hipster Wars: Discovering elements of fashion styles. In *Proc. European Conference on Computer Vision (ECCV)*. 472–488.

[90] A. Kovashka, D. Parikh, and K. Grauman. 2012. WhittleSearch: Image Search with Relative Attribute Feedback. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2973–2980.

[91] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Q. Zhang. 2019. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 3066–3075.

[92] K. Laenen, S. Zoghbi, and M. Moens. 2018. Web search of fashion items with multimodal querying. In *ACM International Conference on Web Search and Data Mining (WSDM)*. 342–350.

[93] Z. Lähner, D. Cremers, and T. Tung. 2018. DeepWrinkles: Accurate and realistic clothing modeling. In *Proc. European Conference on Computer Vision (ECCV)*. 698–715.

[94] C. Lassner, G. Pons-Moll, and P. V Gehler. 2017. A generative model of people in clothing. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 853–862.

[95] A. Laurentini and A. Bottino. 2014. Computer analysis of face beauty: A survey. *Computer Vision and Image Understanding (CVIU)* 125 (2014), 184–199.

[96] S. Lee, S. Oh, C. Jung, and C. Kim. 2019. A global-local emebdding module for fashion landmark detection. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 3153–3156.

[97] C. Li, K. Zhou, and S. Lin. 2015. Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4621–4629.

[98] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng. 2017. Multi-Human parsing in the wild. arXiv:1705.07206

[99] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin. 2018. BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)*. 645–653.

[100] Y. Li, L. Cao, J. Zhu, and J. Luo. 2017. Mining fashion outfit composition using an End-to-End deep learning approach on set data. *IEEE Transactions on Multimedia* 19, 8 (August 2017), 1946–1955.

[101] Y. Li, C. Huang, C. C. Loy, and X. Tang. 2016. Human attribute recognition by deep hierarchical contexts. In *Proc. European Conference on Computer Vision (ECCV)*. 684–700.

[102] Y. Li, L. Song, X. Wu, R. He, and T. Tan. 2018. Anti-Makeup: Learning a Bi-Level adversarial network for makeup-invariant face verification. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*.

[103] Z. Li, Y. Li, W. Tian, Y. Pang, and Y. Liu. 2016. Cross-scenario clothing retrieval and fine-grained style recognition. In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*.

[104] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li. 2018. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*.

[105] X. Liang, K. Gong, X. Shen, and L. Lin. 2019. Look into Person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41 (2019), 871–885.

[106] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan. 2016. Clothes Co-Parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia* 18, 6 (2016), 1175–1186.

[107] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. 2015. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37, 12 (Dec 2015), 2402–2414. https://doi.org/10.1109/TPAMI.2015.2408360

[108] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. 2015. Human parsing with contextualized convolutional neural network. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 1386–1394.

[109] L. Liao, X. He, B. Zhao, C. Ngo, and T. Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)*. 1571–1579.

[110] L. Lin, L. Liang, L. Jin, and W. Chen. 2019. Attribute-Aware convolutional neural networks for facial beauty prediction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*. 847–853.

[111] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*. 740–755.

[112] Z. Lin, H. Xie, P. Kang, Z. Yang, W. Liu, and Q. Li. 2019. Cross-Domain beauty item retrieval via unsupervised embedding learning. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 2543–2547.

[113] K. Liu, T. Chen, and C. Chen. 2016. MVC: A dataset for view-invariant clothing retrieval and attribute prediction. In *ACM International Conference on Multimedia Retrieval (ICMR)*. 313–316.

[114] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, and S. Yan. 2014. Wow! You are so Beautiful Today! *ACM Trans. Multimedia Comput. Commun. Appl.* 11, 1s (2014), 20:1–20:22.

[115] S. Liu, Y. Fan, A. Samal, and Z. Guo. 2016. Advances in computational facial attractiveness methods. *Multimedia Tools and Applications* 75 (2016), 16633–16663.

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

[117] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. 2012. Hi, magic closet, tell me what to wear!. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. 619–628.

[118] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, and S. Yan. 2014. Fashion parsing with video context. In *Proceedings of the 22th ACM International Conference on Multimedia (MM'14)*. 467–476.

[119] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. 2015. Matching-CNN Meets KNN: Quasi-Parametric human parsing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1419–1427.

[120] S. Liu, L. Liu, and S. Yan. 2014. Fashion Analysis: Current techniques and future directions. *IEEE MultiMedia* 21, 2 (2014), 72–79. https://doi.org/10.1109/MMUL.2014.25

[121] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao. 2016. Makeup like a superstar: Deep localized makeup transfer network. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*. 2568–2575.

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

[123] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1096–1104.

[124] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. 2016. Fashion landmark detection in the wild. In *Proc. European Conference on Computer Vision (ECCV)*.

[125] L. Lo, C.-L. Liu, R.-A. Lin, B. Wu, H.-H. Shuai, and W.-H. Cheng. 2019. Dressing for attention: Outfit based fashion popularity prediction. In *Proc. IEEE International Conference on Image Processing (ICIP)*. 3222–3226.

[126] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng. 2020. MER-GCN: Micro-Expression recognition based on relation modeling with graph convolutional networks. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 79–84.

[127] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson. 2014. Fashion 10000: An enriched social image dataset for fashion and clothing. In *Proceedings of the 5th ACM Multimedia Systems Conference (MMSys'14)*. 41–46.

[128] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. 2017. Pose guided person image generation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*. 406–416.

[129] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz. 2018. Disentangled person image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 99–108.

[130] Y. Ma, J. Jia, S. Zhou, J. Fu, Y. Liu, and Z. Tong. 2017. Towards better understanding the clothing fashion styles: A multimodal deep learning approach. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 38–44.

[131] Y. Ma, X. Yang, L. Liao, Y. Cao, and T. Chua. 2019. Who, Where, and What to Wear?: Extracting fashion knowledge from social media. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 257–265.

[132] Z. Ma, J. Dong, Z. Long, Y. Zhang, Y. He, H. Xue, and S. Ji. 2020. Fine-Grained fashion similarity learning by attribute-specific embedding network. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 11741–11748.

[133] U. Mall, K. Matzen, B. Hariharan, N. Snavely, and K. Bala. 2019. GeoStyle: Discovering fashion trends and events. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 411–420.

[134] E. Massip, S. C. Hidayati, W.-H. Cheng, and K.-L. Hua. 2018. Exploiting category-specific information for image popularity prediction in social media. In *Proc. IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 45–46.

[135] K. Matzen, K. Bala, and N. Snavely. 2017. StreetStyle: Exploring World-wide clothing styles from millions of photos. arXiv:1706.01869.

[136] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. 2015. Image-Based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. 43–52.

[137] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert. 2020. Image based virtual Try-On network from unpaired data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5184–5193.

[138] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2011. Multimodal deep learning. *In ICML*. 689–696.

[139] T. V. Nguyen, S. Liu, B. Ni, J. Tan, Y. Rui, and S. Yan. 2012. Sense beauty via face, dressing, and/or voice. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. 239–248.

[140] Y. Ni and F. Fan. 2011. A Two-stage dynamic sales forecasting model for the fashion retail. *Expert Systems with Applications* 38, 3 (2011), 1529–1536.

[141] S. Paris, H. M. Briceño, and F. X. Sillion. 2004. Capture of hair geometry from multiple images. *ACM Transactions on Graphics* 23, 3 (August 2004), 712–719.

[142] J. Park, G. L. Ciampaglia, and E. Ferrara. 2016. Style in the age of instagram: Predicting success within the fashion industry using social media. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*. 64–73.

[143] C. Patel, Z. Liao, and G. Pons-Moll. 2020. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7365–7375.

[144] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics* 36, 4 (July 2017), 73:1–73:15.

[145] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8620–8628.

[146] B. Q. Ferreira, J. P. Costeira, R. G. Sousa, L. Gui, and J. P. Gomes. 2019. Pose guided attention for multi-label fashion image classification. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 3125–3128.

[147] S. Qin, S. Kim, and R. Manduchi. 2017. Automatic skin and hair masking using fully convolutional networks. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*. 103–108.

[148] S. Ren, T. Choi, and N. Liu. 2015. Fashion sales forecasting with a panel data-based particle-filter model. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 45, 3 (2015), 411–421.

[149] Y. Ren, X. Yu, J. Chen, T. H Li, and G. Li. 2020. Deep image spatial transformation for person image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7690–7699.

[150] R. Rothe, R. Timofte, and L. V. Gool. 2016. Some like it Hot - Visual guidance for preference prediction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5553–5561.

[151] C. Rousset and P. Coulon. 2008. Frequential and color analysis for hair mask segmentation. In *Proc. IEEE International Conference on Image Processing (ICIP)*. 2276–2279.

[152] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, Y. Zhao, and T. Huang. 2019. Devil in the Details: Towards accurate single and multiple human parsing. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 4814–4821.

[153] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. 2019. PIFu: Pixel-Aligned implicit function for high-resolution clothed human digitization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2304–2314.

[154] S. Saito, T. Simon, J. Saragih, and H. Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned implicit function for high-resolution 3D human digitization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 81–90.

[155] J. Sanchez-Riera, J.-M. Lin, K.-L. Hua, W.-H. Cheng, and A. W. Tsui. 2017. i-Stylist: Finding the right dress through your social networks. In *International Conference on MultiMedia Modeling (MMM)*. 662–673.

[156] I. Santesteban, M. A. Otaduy, and D. Casas. 2019. Learning-Based animation of clothing for virtual Try-On. *Comput. Graph. Forum* (2019), 355–366.

[157] A. Selle, M. Lentine, and R. Fedkiw. 2008. A mass spring model for hair simulation. *ACM Trans. Graph.* 27, 3 (August 2008), 1–11.

[158] S. Shi, F. Gao, X. Meng, X. Xu, and J. Zhu. 2019. Improving facial attractiveness prediction via Co-attention learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4045–4049.

[159] T. Shi, Y. Yuan, C. Fan, Z. Zou, Z. Shi, and Y. Liu. 2019. Face-to-parameter translation for game character auto-creation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 161–170.

[160] Y. Shih, K. Chang, H. Lin, and M. Sun. 2018. Compatibility family learning for item recommendation and generation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*.

[161] C. Si, W. Wang, L. Wang, and T. Tan. 2018. Multistage adversarial losses for pose-based human image synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 118–126.

[162] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe. 2018. Deformable GANs for Pose-based human image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3408–3416.

[163] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. 2015. Neuroaesthetics in Fashion: Modeling the perception of fashionability. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 869–877.

[164] E. Simo-Serra and H. Ishikawa. 2016. Fashion style in 128 Floats: Joint ranking and classification using weak data for feature extraction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 298–307.

[165] S. Song and T. Mei. 2018. When multimedia meets fashion. *IEEE MultiMedia* 25, 3 (July 2018), 102–108.

[166] S. Song, W. Zhang, J. Liu, and T. Mei. 2019. Unsupervised person image generation with semantic parsing transformation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2357–2366.

[167] X. Song, F. Feng, X. Han, X. Yang, W. Liu, and L. Nie. 2018. Neural compatibility modeling with attentive knowledge distillation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. 5–14.

[168] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma. 2017. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 753–761.

[169] X. Song, X. Han, Y. Li, J. Chen, X. Xu, and L. Nie. 2019. GP-BPR: Personalized compatibility modeling for clothing matching. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 320–328.

[170] G. Sun, X. Wu, and Q. Peng. 2016. Part-based clothing image annotation by visual neighbor retrieval. *Neurocomputing* 213, C (2016), 115–124.

[171] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll. 2020. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *Proc. European Conference on Computer Vision (ECCV)*.

[172] K. Vaccaro, S. Shivakumar, Z. Ding, K. Karahalios, and R. Kumar. 2016. The elements of fashion style. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST'16)*. 777–785.

[173] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth. 2018. Learning type-aware embeddings for fashion compatibility. In *Proc. European Conference on Computer Vision (ECCV)*. 405–421.

[174] A. Veit, S. Belongie, and T. Karaletsos. 2017. Conditional similarity networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1520–1528.

[175] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. 2015. Learning visual clothing style with heterogeneous dyadic Co-Occurrences. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 4642–4650.

[176] S. Vittayakorn, A. C. Berg, and T. L. Berg. 2017. When was That Made?. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*. 715–724.

[177] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi. 2016. Automatic attribute discovery with neural activations. In *Proc. European Conference on Computer Vision (ECCV)*. 252–268.

[178] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. 2015. Runway to realway: Visual analysis of fashion. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*. 951–958.

[179] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proc. European Conference on Computer Vision (ECCV)*. 589–604.

[180] H. Wang, J. F. O'Brien, and R. Ramamoorthi. 2011. Data-driven elastic models for cloth: Modeling and measurement. In *ACM SIGGRAPH' 11*. 1–12.

[181] J. Wang and J. Allebach. 2015. Automatic assessment of online fashion shopping photo aesthetic quality. In *Proc. IEEE International Conference on Image Processing (ICIP)*. 2915–2919.

[182] N. Wang, H. Ai, and S. Lao. 2010. A compositional exemplar-based model for hair segmentation. In *Proc. Asian Conference on Computer Vision (ACCV)*. 171–184.

[183] N. Wang, H. Ai, and F. Tang. 2012. What are good parts for hair shape modeling?. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 662–669.

[184] S. Wang and Y. Fu. 2016. Face behind makeup. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 58–64.

[185] W. Wang, Y. Xu, J. Shen, and S. Zhu. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4271–4280.

[186] W. Wang, W. Zhang, J. Wang, J. Yan, and H. Zha. 2018. Learning sequential correlation for user generated textual content popularity prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*. 1625–1631.

[187] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao. 2020. Learning compositional neural information fusion for human parsing. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 5703–5713.

[188] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y. Jiang. 2016. Matching user photos to online products with robust deep features. In *ACM International Conference on Multimedia Retrieval (ICMR)*. 7–14.

[189] X. Wang, B. Wu, and Y. Zhong. 2019. Outfit compatibility prediction and diagnosis with multi-layered comparison network. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 329–337.

[190] X. Wang and T. Zhang. 2011. Clothes search in consumer photos via color matching and attribute learning. In *Proceedings of the 19th ACM International Conference on Multimedia (MM'11)*. 1353–1356.

[191] Y. Wang, T. Shao, K. Fu, and N. Mitra. 2019. Learning an intrinsic garment space for interactive authoring of garment animation. *ACM Transactions on Graphics* 38, 6 (2019). https://doi.org/10.1145/3355089.3356512

[192] B. Wu, W. Cheng, Y. Zhang, H. Qiushi, L. Jintao, and T. Mei. 2017. Sequential prediction of social media popularity with deep temporal context networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 3062–3068.

[193] B. Wu, W.-H. Cheng, P. Liu, Z. Zeng, and J. Luo. 2019. SMP challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 2667–2671.

[194] B. Wu, T. Mei, W. Cheng, and Y. Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 272–278.

[195] Z. Wu, G. Lin, Q. Tao, and J. Cai. 2019. M2E-Try On Net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 293–301.

[196] F. Xia, P. Wang, X. Chen, and A. L. Yuille. 2017. Joint multi-person pose estimation and semantic part segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6080–6089.

[197] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays. 2018. TextureGAN: Controlling deep image synthesis with texture patches. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8456–8465.

[198] D. Xie, L. Liang, L. Jin, J. Xu, and M. Li. 2015. SCUT-FBP: A benchmark dataset for facial beauty perception. In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*. 1821–1826.

[199] H.-X. Xie, Ling Lo, H.-H. Shuai, and W.-H. Cheng. 2020. AU-Assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*. 2871–2880.

[200] Y. Xiong, K. Zhu, D. Lin, and X. Tang. 2015. Recognize complex events from static images by fusing deep channels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1600–1609.

[201] J. Xu, L. Jin, L. Liang, Z. Feng, D. Xie, and H. Mao. 2017. Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN). In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1657–1661.

[202] Z. Xu, H. Wu, L. Wang, C. Zheng, X. Tong, and Y. Qi. 2014. Dynamic hair capture using spacetime optimization. *ACM Transactions on Graphics* 33, 6 (November 2014), 224:1–224:11.

[203] Y. Yacoob and L. S. Davis. 2006. Detection and analysis of hair. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 28, 7 (2006), 1164–1169.

[204] K. Yamaguchi, T. L. Berg, and L. E. Ortiz. 2014. Chic or Social: Visual popularity analysis in online fashion networks. In *Proceedings of the 22th ACM International Conference on Multimedia (MM'14)*. 773–776.

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

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

[207] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. 2014. Retrieving similar styles to parse clothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37, 5 (2014), 1028–1040.

[208] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi. 2015. Mix and Match: Joint model for clothing and attribute recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 51.1–51.12.

[209] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. 2017. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 172–180.

[210] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. 2020. Towards photo-realistic virtual try-on by adaptively generating ↔ preserving image content. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7850–7859.

[211] S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. Berg, and M. C. Lin. 2017. Detailed garment recovery from a single-view image. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

[212] W. Yang, P. Luo, and L. Lin. 2014. Clothing Co-parsing by joint image segmentation and labeling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3182–3189.

[213] W. Yang, M. Toyoura, and X. Mao. 2012. Hairstyle Suggestion using Statistical Learning. In *International Conference on MultiMedia Modeling (MMM)*. 277–287.

[214] X. Yang, Y. Ma, L. Liao, M. Wang, and T. Chua. 2019. TransNFCM: Translation-based neural fashion compatibility modeling. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*. 403–410.

[215] G. Yildirim, N. Jetchev, R. Vollgraf, and U. Bergmann. 2019. Generating high-resolution fashion model images wearing custom outfits. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*. 3161–3164.

[216] W. Yin, Y. Fu, Y. Ma, Y. Jiang, T. Xiang, and X. Xue. 2017. Learning to generate and edit hairstyles. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 1627–1635.

[217] D. Yoo, N. Kim, S. Park, Anthony S Paek, and In So Kweon. 2016. Pixel-level domain transfer. In *Proc. European Conference on Computer Vision (ECCV)*. 517–532.

[218] A. Yu and K. Grauman. 2014. Fine-grained visual comparisons with local learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 192–199.

[219] C. Yu, Y. Hu, Y. Chen, and B. Zeng. 2019. Personalized fashion design. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 9046–9055.

[220] R. Yu, X. Wang, and X. Xie. 2019. VTNFP: An Image-based Virtual Try-On network with body and clothing feature preservation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 10511–10520.

[221] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, and Y. Liu. 2019. SimulCap: Single-view human performance capture with cloth simulation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5504–5514.

[222] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian. 2017. Trip outfits advisor: Location-oriented clothing recommendation. *IEEE Transactions on Multimedia* 11 (Nov 2017), 2533–2544.

[223] B. Zhao, J. Feng, X. Wu, and S. Yan. 2017. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6156–6164.

[224] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng. 2018. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)*. 792–800.

[225] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. 2015. Scalable person Re-identification: A benchmark. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 1116–1124.

[226] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie. 2019. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*. 266–274.

[227] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. 2018. ModaNet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)*. 1670–1678.

[228] Z. Zheng and C. Kambhamettu. 2017. Multi-level feature learning for face recognition under makeup changes. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 918–923.

[229] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. Han. 2020. Deep Fashion3D: A dataset and benchmark for 3D garment reconstruction from single images. In *Proc. European Conference on Computer Vision (ECCV)*.

[230] J. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2242–2251.

[231] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. C. Loy. 2017. Be your own Prada: Fashion synthesis with structural coherence. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 1680–1688.

[232] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao. 2019. FashionAI: A hierarchical dataset for fashion understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 296–304.