

Received November 29, 2021, accepted December 19, 2021, date of publication December 23, 2021,
date of current version January 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3137893

Deep Learning Approaches for Fashion Knowledge Extraction From Social Media: A Review

MARCO MAMELI^{ID1}, MARINA PAOLANTI^{1,2}, ROCCO PIETRINI^{1,3}, GIULIA PAZZAGLIA¹,
EMANUELE FRONTONI^{1,2}, AND PRIMO ZINGARETTI^{ID1}, (Senior Member, IEEE)

¹Dipartimento di Ingegneria dell'Informazione (DII), Università Politecnica delle Marche, 60131 Ancona, Italy

²Department of Political Sciences, Communication and International Relations, University of Macerata, 62100 Macerata, Italy

³Grottini Lab, 62017 Porto Recanati, Italy

Corresponding author: Marina Paolanti (marina.paolanti@unimc.it)

ABSTRACT Fashion knowledge encourages people to properly dress and faces not only physiological necessity of users, but also the requirement of social practices and activities. It usually includes three jointly related aspects of: occasion, person and clothing. Nowadays, social media platforms allow users to interact with each other online to share opinions and information. The use of social media sites such as Instagram has already spread to almost every fashion brand and been evaluated as business take-off tools. With the heightened use of social media as a means of marketing communication for fashion brands, it has become necessary to empirically analyse and extract fashion knowledge from them. Thus, social brands are investing on them. In this way, they can understand the consumer's preferences. This change is also having a significant impact on social media data analysis. To solve this issue, the Deep learning (DL) methods are proven to be effective solutions due to their automatic learning capability. However, little systematic work currently exists on how researchers have applied DL for analysing fashion knowledge from social media data. Hence, this contribution outlines DL-based techniques for social media data related to fashion domain. In this study, a review of the dataset within the fashion world and the DL methods applied on, it is presented to help out new researchers interested in this subject. In particular, five different tasks will be considered: Object Detection, that includes Clothes Landmark Detection, Clothes Parsing and Product Retrieval, Fashion Classification, Clothes Generation, Automatic Fashion Knowledge Extraction and Clothes Recommendation. Therefore, the purpose of this paper is to underline the multiple applications within the fashion world using deep learning techniques. However, this review does not cover all the methods used: in fact, only Deep Learning methods have been analyzed. This choice was made since, given the huge amount of fashion social media data that has been collected, Deep Learning methods achieve the best performance both in terms of accuracy and time. Limitations point towards unexplored areas for future investigations, serving as useful guidelines for future research directions.

INDEX TERMS Artificial intelligence, machine learning, deep learning, fashion, neural networks, object detection, object parsing, product retrieval, clothes classification, fashion recommendation, fashion datasets, generative adversarial networks, social media.

I. INTRODUCTION

Online Social networks are part of every person's life. More than half of the world's population is connected to the internet and has at least one social platform. According to the report carried out by *We Are Social* of January 2021, in the world there are 7.83 billion people, 66.6% of these have a mobile phone. 4.66 billion people access the internet, an increase of

7.3% compared to January 2020. World internet penetration stands at 59.5%, but the values could be even higher by virtue of problems related to the correct tracking of internet users related to the COVID-19 pandemic. There are 4.20 billion users of social platforms, an increase of 13%. The use of social platforms therefore stands at 53% of the world population.

In particular, social networks have long since changed the way of communicating and perceiving the world: it is therefore no coincidence that fashion, of which communication

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia ^{ID}.

and perception are two fundamental pillars, is an integral part of this revolution. In fact, the fashion industry is one of the most dynamic in society and in this context social media are fundamental communication tools, in particular Facebook (born in 2004), Instagram (born in 2010) and Tik Tok (born in 2018).

Facebook was born in 2004 and, to date, is one of the most used social networks in the world, with over 2 billion active users. To date, many fashion brands are present on Facebook with a company page. The primary goal is to attract new customers and retain existing ones. A strategically managed Facebook page with careful publication of content will make a brand more attractive, involving an increasing number of users.

Instagram was born in 2010 and one of the strengths of this social network is the communicative power of the images that are able to convey the identity of a brand. Tik Tok was born in 2018 and it is a platform where users can express their creativity to the maximum through short videos between 15 and 60 seconds, with background music of all kinds.

The main social reference for the fashion domain is Instagram. However, leading fashion brands have proven the power of social media marketing across multiple channels. Each channel has different features to offer, giving new ways to achieve goals. Facebook is the most used social media platform in the world with more than 2 billion monthly active users. In addition to regular Facebook posts, fashion app marketers can use the platform for live broadcasts.

Instagram has an active global audience of 500 million daily active users, collectively tapping the platforms “Like” button 4.2 billion times every day. Ecommerce brands can also use Instagram’s shopping features, allowing users to purchase items without leaving the app. Instagram offers several ways to connect with audience, including Posts, Reels, Stories, Highlights and IGTV. Many users use Pinterest as a discovery platform to identify and refine their style: 53 percent of users say Pinterest has helped them make a fashion related purchase decision. Twitter has 330 million monthly active users tweeting 500 million times per day. The social media platform can also be used to successfully promote fashion brands, Fashionista.com (2.1 million followers), Zara (1.3 million followers) and Misguided (466.k million followers) have all built sizable audiences by harnessing Twitter’s potential for virality. For example, Fashionista.com have captured the zeitgeist by discussion one of Netflix’s most recent (and fashion-centric) shows, Emily in Paris. Snapchat reported 238 million daily active users in 2020. In 2021, Ralph Lauren collaborated with Snapchat to create virtual reality experience for their users. This enabled Snapchat users to style their personal avatars in Ralph Lauren items, which are also shoppable.

Especially on the Instagram social network, fashion brands have started to invest a large part of their budget, as it allows them to publish very accurate and creative images, similar to professional photographs. With the social network Instagram, the influencer phenomenon has been strengthened. As the

word suggests, the influencer is a famous person who can influence public opinion and constitute an important target to which to direct advertising messages, in order to accelerate their acceptance by a wider audience. Since these characters inspire confidence, fashion houses have an incentive to invest money and resources in this type of strategy. In fact, as reported by the fashion marketing site MuseFind, 92% of consumers consider an influencer campaign more reliable than traditional advertising with models or celebrities. For this reason, since 2016, 65% of luxury brands chose to collaborate with influencers for their advertising campaigns, with amazing results.

Then Instagram from a social network becomes more and more like an e-Commerce showcase. In fact, the fashion and luxury brands are making great efforts to keep up with the times and adapt to change. Everyone is equipping themselves with e-Commerce platforms. The purpose of these brands is therefore to understand the preferences of new consumers, to communicate with them directly and without filters, to be able to customize their offer.

Moreover, researchers have proposed several fashion recommender systems in the literature aiming at choosing the right outfit for different occasions [1]. Companies therefore must try to analyze the information that is spontaneously generated by web users. Big Data analysis now makes it possible to predict future trends even before they explode, providing real-time information not only on the volume of sales, but also on that of online searches. More quickly identifying fabrics, styles and colors for which public interest is growing allows us to satisfy the request in a timely manner and consequently to sell more.

For this reason, the interest in applying artificial intelligence (AI) algorithms to Big Data, and in particular those based on Deep Learning (DL), that is a subset of Machine Learning and mainly in the recent years is growing more and more. Thanks to ML techniques, companies operating in the fashion sector can identify patterns in data and build models that can predict future results. This helps to create a more flexible and faster supply chain and manage inventory in an automated and intelligent way. In addition, algorithms for clothing design have been developed [2]: the aim is to provide the customer with a model capable of generating data similar to those given in input and to give advice on the most relevant products. These algorithms are therefore useful for analyzing consistent datasets and automating the process of recognition and classification of the proposed styles.

Considering the latest achievements in data collection and processing [3], DL is facing the worldwide challenge of, on one hand, reducing the need of manual intervention for huge datasets and, on the other, improving methods for facilitating their interpretation. To close this gap, this review aims to provide a technical overview of the advances and opportunities offered by DL for automatically processing and analysing social media data related to fashion domain.

Existing reviews explore particular approaches for analysing fashion data, generally based on Artificial

Intelligence techniques to solve a specific issue. There are several examples of well-structured systematic reviews focused on this domain [4]. An example aims to study the impact and the significance of AI in the fashion industry in the last decades throughout the supply chain [5], while the most recent [6] has the aim to study the impact and significance of AI in fashion e-commerce. In the context of fashion recommendation system an interesting and recent review is presented by [7]. The authors in detail describe the technical aspects, strengths and weaknesses of the filtering techniques. Moreover they help researchers, and practitioners of machine learning, computer vision, and fashion retailing to understand the characteristics of several fashion recommendation systems. However one limitation of this study is that a review of the datasets that have been used in fashion recommendation was not considered. An aspect that is considered in this review. Moreover, the novelty of this work relies on social media fashion data. In fact, to the best of our knowledge, a complete review on deep learning based approach for deducing insights from social media images is not present in literature. With this work a thorough survey of the state of art related to the use of social media fashion data and their tasks has been presented, with a particular focus on Deep Learning methods. Methods and techniques for each kind of fashion task have been analysed, the main paths have been summarised, and their contributions have been highlighted. The reviewed approaches have been categorised and compared from multiple perspectives, pointing out their advantages and disadvantages. Finally, several interesting examples of the dataset have been presented.

In particular, the purposes, issues, motivations were investigated to set the following research questions (RQ):

- RQ1** To make an overview of the main tasks performed by using fashion data, the question to be answered is: *For what tasks is fashion data used and how has the use of this data developed over time?*
- RQ2** To explore the most used methodologies in recent years to deal with fashion data, the question that has been set is: *Comparing ML and DL methods, does the fashion data influence the choice of using one methodology rather than another?*
- RQ3** To better understand what are the future applications in this area, the following question arises: *What are the future applications that need to be developed and deepened that use fashion data?*
- RQ4** To understand how companies can use information from social media, the question that must be asked is the following: *How has social media changed the marketing strategies of fashion brands?*

This paper is structured as follows. Section II describes the methodology adopted in the choice of the articles identified and selected for the review work. Section III describes the dataset used in general for fashion tasks, in particular for object detection. Section IV shows the deep learning methods used for object detection, classification and for generative clothes task. Section V describes some of the dataset and deep

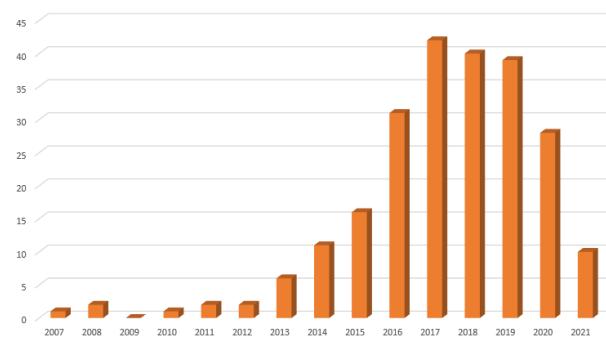


FIGURE 1. Number of papers cited per year.

learning methods used for the Automatic Fashion Knowledge Extraction and Clothes Recommendation of data from social networks. Section VI presents a discussion of the methods taking into exams. Finally Section VIII shows the conclusions of this work.

II. RESEARCH STRATEGY DEFINITION

In literature, there are still no reviews that speak in general of the different research strategies after asking the research questions on the field of deep learning applied to the field of fashion. Guidelines for the review finalisation. These guidelines are motivated by the fact that deep learning approaches for social media data and fashion dataset are quite new. In particular, if we focus on generative adversarial neural networks (GANs) for fashion domain (e.g. virtual try on with GAN) the interesting paper starting in 2017. These lead to an exclusion of paper dated before 2007 for sake of completeness.

A systematic review of the literature was conducted using PRISMA guidelines and electronic databases: ieeExplore,¹ Scopus,² Sciencedirect,³ Citeseerx,⁴ and SpringerLink.⁵ The sequel to a set of keywords was considered. They are chosen in relation to the fashion domain and on the basis of a preliminary screening of the research field. The keywords initially considered in the research were: artificial intelligence, machine learning, deep learning, neural networks, object detection, object parsing, product retrieval, classification, fashion, dataset, generative adversarial networks, social media. To get more accurate results the keywords have been aggregated. In one set of queries, keywords deep learning and fashion was combined with methodology-related others, in other sets deep learning and fashion was combined with application. Each query produced a large amount of articles, which were selected based on relevance and year of publication. Articles found to be inconsistent with the research topic and published before the year 2007 were removed from the list.

¹<https://ieeexplore.ieee.org/Xplore/home.jsp>

²<https://www.scopus.com/>

³<https://www.sciencedirect.com/>

⁴<https://citeseerx.ist.psu.edu/index>

⁵<https://link.springer.com/>

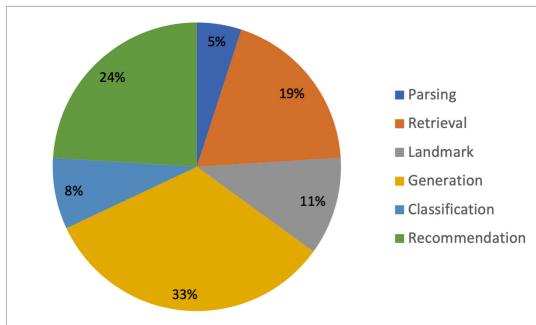


FIGURE 2. Percentage of papers cited per task.

The articles considered for review were published between the years 2007-2021. In total 219 papers were cited, some concerning datasets, others concerning theoretical methodologies, others concerning applications. The number of articles cited per year is shown in Figure 1.

Furthermore, in Figure 2, it is possible to see the percentage of works carried out for each task treated during the review. In other words, it represents the percentage of articles divided into the tasks that have been taken into consideration.

III. FASHION DATASETS

This section is used to give a detailed description of the datasets collected in fashion world. From 2012 to 2020, 51 datasets were built, which are divided by task and by year of publication as shown in Figure 3.



FIGURE 3. Number of dataset for different fashion tasks.

What this graph shows is that the year with the most datasets created was 2019. Looking at the pie chart in Figure 4, it is possible to notice that the most successful task, was Clothes Parsing, with 32% datasets that deal with this tasks.

Finally, the total number of papers is 51: 17 for clothes parsing, 15 for product retrieval; 10 for clothes generation; 4 for clothes landmark detection; 2 for fashion classification and 8 for clothes recommendation. We have to highlight that the total number of categories does not correspond with the number of selected papers, this because two papers concern more than one category.

To summarize all these datasets, Table 1 was built. Here, the datasets from 2012 to 2020 in chronological order are

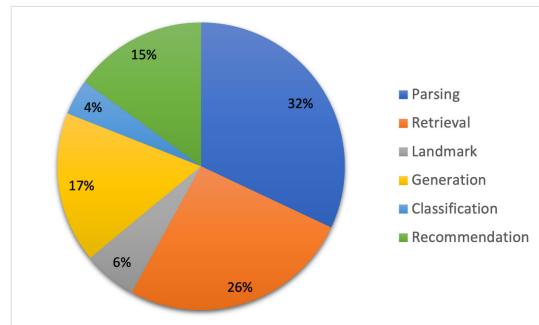


FIGURE 4. Tasks in percentage of papers considered from 2012 to 2020.

reported. This table is divided as follows: the first column represents the name of the dataset and the respective citation; the second column represents the year of publication; the third column shows the purpose for which the dataset was constructed or used; the fourth column presents the source on which the dataset was built.

A. DATASETS FOR OBJECT DETECTION

This section contains the tables that describe the different datasets used for the object detection task:

- Table 2 contains the datasets used for Clothes Landmark Detection. The name of the dataset, the citation, the year of publication, the number of images contained within the dataset and the number of landmark annotations are reported.
- Table 3 contains the datasets used for Clothes Parsing task. The name of the dataset, the citation, the year of publication, the number of images contained within the dataset and the number of classes are reported.
- Table 4 contains the datasets used for Product Retrieval task. The name of the dataset, the citation, the year of publication and the number of images contained within the dataset are reported.

Object detection is one of the best known and most common tasks in the world of deep learning. This section will present the datasets that have been used for this purpose in the fashion world. Object detection can be divided into 3 macro areas:

- *Clothes Landmark Detection*: The purpose of this area is to predict the locations of key points on clothes. These points, as for example, where the collar of a shirt ends or the cuff and hem, are of fundamental importance as they are able to indicate the region of the outfit and delimit it. So, the purpose is to predict which are the locations of the K functional key points defined on the fashion items. Given an image I as input, the aim is to predict the position L of the cloth landmark, where L can be defined as

$$L = \{L_k : k = 1, \dots, K\}, \quad L_k \in \mathbb{R}^2$$

and L_k is the position of every pixel (u, v) in the input image. The datasets used for this purpose are listed in chronological order in Table 2.

TABLE 1. Total number of state-of-the-art datasets, from 2012 to 2020. The first column represents the name of the dataset; the second column represents the year of publication; the third column denotes the number of images within each dataset. The fourth column shows the purpose for which the dataset was constructed or used; the fifth column presents the source on which the dataset was built.

Dataset	year	# of images	purpose	sources
Fashionista [8]	2012	158 235	Clothes Parsing	Chitopia.com
Daily Photos (DP) [9]	2013	2 500	Clothes Parsing	Chitopia.com
Paper Doll [10]	2013	339 797	Clothes Parsing	Fashionista [8], Chitopia.com
CCP [11]	2014	2 098	Clothes Parsing	Shopping websites
CFD [12]	2014	2 682	Clothes Parsing	Chitopia.com
Fashion 10000 [13]	2014	32 398	Product Retrieval	Flickr.com
Deep Search [14]	2014	206 235	Product Retrieval	Taobao.com, Tsmall.com, Amazon.com
UT-Zap50K [15]	2014	50 025	Clothes Generation	Zappos.com
ATR [16]	2015	7 700	Clothes Parsing	Fashionista [8], DP [9], CFD [12]
Chitopia10k [17]	2015	10 000	Clothes Parsing	Chitopia.com
DARN [18]	2015	545 373	Product Retrieval	Shopping websites, corresponding customer review pages
Exact Street2Shop [19]	2015	425 040	Product Retrieval	ModCloth.com, clothing retailers
DeepFashion [20]	2016	289 222	Clothes Landmark Detection - Clothes Generation	Shopping websites, Google images
FLD [21]	2016	123 016	Clothes Landmark Detection	DeepFashion [20]
MVC [22]	2016	161 638	Product Retrieval	Shopping websites
Li et al. [23]	2016	19 896	Product Retrieval	Shopping websites, Flickr.com
LookBook [24]	2016	84 748	Clothes Generation	Bongjashop, Jognunshop, Stylenanda, SmallMan, WonderPlace
Unconstrained Landmark Database [25]	2017	30 000	Clothes Landmark Detection	Fashion blogs, DeepFashion [20]
LIP [26]	2017	50 462	Clothes Parsing	Microsoft COCO [27]
PASCAL-Person-Part [28]	2017	3 533	Clothes Parsing	
MHP v1.0 [29]	2017	4 980	Clothes Parsing	
Video2Shop [30]	2017	112 029	Product Retrieval	Tmall.com, Taobao.com
Dress like a star [31]	2017	7 000 000	Product Retrieval	YouTube.com
Fashion-MNIST [32]	2017	70 000	Fashion Classification	Zalando.com
Fashion Style14 [33]	2017	13 126	Clothes Recommendation	
Polyvore [34]	2017	21 899	Clothes Recommendation	Polyvore.com
CAGAN [35]	2017	15 000	Clothes Generation	Zalando.com
MHP v2.0 [36]	2018	25 403	Clothes Parsing	
CIHP [37]	2018	38 280	Clothes Parsing	Google, Bing
ModaNet [38]	2018	55 176	Clothes Parsing	PaperDoll [10]
Amazon [39]	2018	489 000	Product Retrieval	Amazon.com
Fashion-GEN [40]	2018	293 008	Clothes Generation	
VITON [41]	2018	32 506	Clothes Generation	
DeepFashion2 [42]	2019	491 000	Clothes Parsing - Clothes Landmark Detection - Product Retrieval	DeepFashion [20], Shopping websites
FindFashion [43]	2019	565 041	Product Retrieval	Street2Shop [19], DeepFashion [20]
Polyvore-T [44]	2019	19 835	Clothes Recommendation	Polyvore [34]
iFashion [45]	2019	1 012 947	Clothes Recommendation	Wish.com
Clothing Recommendation Dataset [46]	2019	127 824	Clothes Recommendation	Shopping websites
FashionAI [47]	2019	357 000	Clothes Recommendation	
FashionIQ [48]	2019	77 684	Product Retrieval	
FashionTryOn [49]	2019	28 714	Clothes Generation	Zalando.com
FashionOn [50]	2019	22 566	Clothes Generation	DeepFashion [20], Internet fashion model catwalk
Video Virtual Try-on [51]	2019	791 videos	Clothes Generation	
FashionKE [52]	2019	80 629	Clothes Recommendation	Instagram
FACAD [53]	2020	993 000	Product Retrieval	Google
CBL [54]	2020	57 000	Fashion Classification	Shopping websites
SIZER [55]	2020	2 000	Clothes Parsing (3D)	
UTFPR-SBD3 [56]	2020	4 500	Clothes Parsing	Chitopia.com, Fashionista [8], CFD [12]
Fashionpedia [57]	2020	50 527	Clothes Parsing	Flickr.com, photo websites
Ma et al. [58]	2020	180 000	Product Retrieval	DARN [18], FashionAI [47], DeepFashion [20]
FIT [59]	2020	680 000	Clothes Recommendation	Instagram

TABLE 2. Clothes landmark detection datasets.

Dataset Name	Year of Publication	#images	#landmark annotation
DeepFashion-C [20]	2016	289 222	8
Fashion Landmark Dataset (FLD) [21]	2016	123 016	8
Unconstrained Landmark Database [25]	2017	30 000	6
DeepFashion2 [42]	2019	491 000	

TABLE 3. Clothes parsing datasets.

Dataset Name	Year of Publication	#images	#classes
Fashionista [8]	2012	158 235	56
Daily Photos (DP) [9]	2013	2 500	18
Paper Doll [10]	2013	339 797	56
CCP [11]	2014	2 098	57
CFD [12]	2014	2 682	23
ATR [16]	2015	7 700	18
Chictopia10k [17]	2015	10 000	18
LIP [26]	2017	50 432	20
PASCAL-P-P [28]	2017	3 533	14
MHP v1.0 [29]	2017	4 980	18
MHP v2.0 [36]	2018	25 403	58
CIHP [37]	2018	38 280	19
ModaNet [38]	2018	55 176	13
DeepFashion2 [42]	2019	491 000	13
SIZER [55]	2020	2 000	
UTFPR-SBD3 [56]	2020	4 500	18
Fashionpedia [57]	2020	50 527	

- *Clothes Parsing:* Clothes Parsing is a subsection of semantic segmentation, where the clothing items represent the labels. It can be seen as a labeling problem regions of an image. If an image I that shows a person is taken as input, the aim is to attribute a label of a clothing or null (if the background is considered) item to each pixel. Assuming that uniform appearance regions concern to the same item, the problem can be simplified and reduced to a prediction over a set of superpixels. The datasets used for this purpose are listed in chronological order in Table 3.
- *Product Retrieval:* Given an image that contains fashion styles as input, the purpose of fashion retrieval based on images is to find similar or equal items from an archive of shopping image inside of online sites. Table 4 summarizes the datasets used for the Product Retrieval task.

These aspects will be better explored later in Section IV-A.

Some of the most important datasets used for object detection task will be described in detail.

- *Fashionista (2012):* Fashionista dataset was introduced by Yamaguchi *et al.* in [8]. This dataset consists of 158 235 photographs collected from a social networking website Chictopia.com. They observed 53 different clothing items and adding additional labels for hair, skin, and background, their proposed gives a total of 56 different possible clothing labels. The annotation was made by tags, comments, and links.
- *Paper-Doll (2013):* Paper-Doll dataset was presented by Yamaguchi *et al.* in [10] and it is an extension of Fashionista. This dataset contains 339 797 images,

TABLE 4. Product retrieval datasets.

Dataset Name	Year of Publication	#of images
Fashion 10000 [13]	2014	32 398
Deep Search [14]	2014	206 235
DARN [18]	2015	545 373
Exact Street2Shop [19]	2015	425 040
MVC [22]	2016	161 638
Li <i>et al.</i> [23]	2016	19 896
DeepFashion [20]	2016	52 712
Dress Like a Star [31]	2017	7 000 000
Amazon [39]	2018	489 000
Perfect-500K	2018	500 000
DeepFashion2 [42]	2019	491 000
FashionIQ [48]	2019	77 684
FindFashion [43]	2019	565 041
FACAD [53]	2020	993 000
Ma <i>et al.</i> [58]	2020	180 000

annotated with tags that denote the characteristics of the item present in the image, e.g., brand, clothing type, color, occasion, style, for a total of 56 classes. Since the Fashionista dataset uses Chictopia.com, they automatically excluded all the duplicate images present in Fashionista to create the Paper Doll dataset.

- *Colorful-Fashion (CFD) (2014):* From the site Chictopia.com, the authors Liu *et al.* in [12] collected the images contained in Colorful-Fashion dataset. This dataset includes 2 682 photos annotated with 13 different colors and 23 categories of tags.
 - *Clothing Co-Parsing (CCP) (2014):* CCP is created by Yang *et al.* in [11], and it consists of 2 098 fashion photos with high-resolution. It presents a lot of human/clothing variations, with different styles, accessories, garments, and poses. The images were taken from various online shopping websites. More than 1000 images of CCP are annotated with superpixel-level for a total of 57 tags; the rest of images are with image-level tags annotations.
 - *Fashion 10000 (2014):* Fashion 10000 dataset was created by Loni *et al.* in [13]. It comprises 32 398 images collected from Flickr.com using an automatic approach and annotated with 470 fashion categories.
 - *Chictopia10k (2015):* From the social network Chictopia.com, Liang *et al.* in [17] collected 10 000 real-world human pictures. The images within the dataset are annotated with pixel-level labels. In addition to images with people posing, this dataset also contains images in the wild. The images present different and arbitrary postures, views and backgrounds with a total of 18 classes.
 - *Exact Street2Shop (2015):* Kiapour *et al.* in [19] proposed the Exact Street2Shop dataset, that is divided into two parts:
 - Street Photos: they collected 20 357 style gallery outfit posts from ModCloth.com;
 - Shop Photos: This part of the dataset contains 404 683 shop photos which come from 25 different online clothing retailers.
- Adding these two parts, a dataset that contains 425 040 photos can be obtained. Each image within the dataset

is combined with two variety of connections to clothing items of the shop: the first group of connections are links to objects that perfectly correspond to one of the items in a street photo, while connections within the second group represent object that are only similar to a street item. In total, they extract 39 479 exact matching street-to-shop item pairs.

- **DARN (2015):** DARN dataset, developed by Huang et al in [18], consists of 453 983 images of upper-clothing in high-resolution from several online-shopping websites and 91.390 offline upper-clothing images from corresponding customer review pages, for a total of 545 373 images, annotated with clothing attribute categories.
- **ATR (2015):** ATR dataset was created by Liang et al. in their work [16]. This dataset is composed by 5.867 images of Benchmarks from Fashionista dataset [8], Daily Photos dataset [9] and CFD dataset [12] which represent people with full body view taken from the front or near-front, in which it is possible to see all parts of the body.
- **DeepFashion (2016):** DeepFashion is a large-scale clothes dataset invented by Liu et al. in [20]. It is made up of 800 000 different fashion images that include both store images and consumer photos. For each image, labeling is done through 50 categories, 1 000 descriptive attributes and clothing landmarks. Additionally, it contains over 300 000 cross-pose/cross-domain image pairs. The images are taken from Google Images, Forever21 and Mogujie, two Online Shopping Sites.
- **Fashion Landmark Dataset (FLD) (2016):** Fashion Landmark Dataset, was produced by Liu et al [21] and it contains 123 016 images. The annotations is done using 8 different landmark annotations, i.e. landmark visibility, human body joint, clothing pose variation, clothing bounding box. The sources of this dataset is Deep Fashion [20]
- **Multi-View Clothing (MVC) (2016):** MVC dataset was collected by KH. Liu et al. in [22] by crawling images from a lot of online shopping websites, such as Amazon.com, Zappos.com or Shopbop.com. The MVC dataset consists of 37 499 items and 161 638 clothing images, where most items have at least four views between front, back, left, and right views. They collect the ground truth attributes from the websites and manually select 264 attributes for similarity evaluation. These 264 attributes are organized into a three-layer hierarchy. The first layer enforces the gender of clothes, which contains two attributes, Men and Women. There are eight categories for Men's clothes and nine categories for Women's clothes, where most of them overlap. The third layer contains more detailed attributes.
- **ModaNet (2018):** Modanet Dataset was developed by Zheng et al. in [38] and it consists of 55 176 photos. There are 13 meta categories: these meta-categories are used to group categories that are highly correlated

with each other, reducing ambiguity in the annotation process. This metadata categories are dress, pants, belt, top, skirt, boots, headwear, outer, bag, scarf and tie, sunglasses, shorts, footwear.

- **Crowd Instance-Level Human Parsing (CIHP) (2018):** Gong et al. in [37] proposed CIHP dataset that is composed by 38 280 images. The Sources for the construction was Google and Bing. CIHP contains 19 semantic part labels: Right/Left shoe, Socks, Right/Left leg, Pants, Skirt, Dress, Right/Left arm, Upper-clothes, Gloves, Torso skin, Scarf, Face, Hair, Hat.
- **Fashion Captioning Dataset (FACAD) (2020):** Yang et al. in [53] mainly crawled images that contains detailed information for fashion using the search engine Google Chrome. This dataset consists of over 993 000 images and 130 000 explanations with a lot of categories and attributes. The categories are generated by pick into account the item titles and selecting the last word. After selecting and filtering manually the categories, a total of 472 categories remain. Then, similar categories are merged and only those containing more than 200 items are kept, resulting in 78 categories that are unique. If an element is selected, it can belongs to only one category. The number of the total attributes extracted is over 3 000, and only those attributes that appear in more than 10 items have been retained, resulting in a list of 990 attributes.
- **Fashion IQ (2019):** Fashion IQ contains fashion products images coming from a product review dataset [60] by Guo et al. in [48]. They selected three categories of product items, specifically: Shirts, Dresses and Tops&Tees. For each image, they followed the link to the product website available in the dataset, in order to extract corresponding product information, when available. Leveraging the textual information within the website, the authors pulled out attribute labels that contains fashion information from them. In particular, from the product title, the product summary, and detailed product description, product attributes were extracted. In total, 1 000 attribute labels were extracted, further grouped into 5 attribute: texture, fabric, shape, part, and style.
- **SIZER (2020):** Tiwari et al. in [55] created SIZER, a dataset of clothing size variation of approximately 2 000 scans including 100 subjects wearing 10 garment classes in different sizes, where scans, clothing segmentation, SMPL+G registrations, body shape under clothing, garment class and size labels are available.
- **UTFPR-SBD3 (2020):** In [56], the authors constructed UTFPR-SBD3, intended for clothing segmentation in the context of soft biometrics. The dataset is composed of 4 500 images manually annotated into 18 classes and an addition class for the background. 1 003 of the images come from the CCP dataset [11], 2 679 from the CFPD [12], and 685 from the Fashionista dataset [8]. Furthermore, 133 images were collected

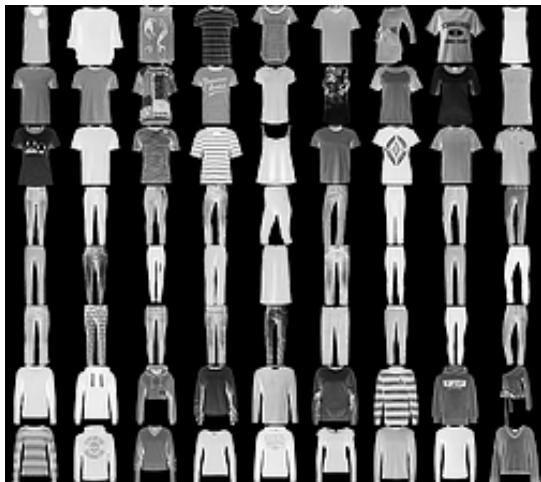


FIGURE 5. Fashion-MNIST dataset.

from the website Chictopia.com: they contains instances of the less frequent classes in the dataset, to ensure that each class has at least 100 instances.

- **Fashionpedia (2020):** From Flickr and other free license photo website, a total of 50 527 images were collected. Then, after filtering the images that contained fashion items, 48 825 images remained and used to build Fashionpedia dataset, collected by Jia *et al.* in [57]. In this dataset, the annotation of the images are done with one or more fundamental garments. Furthermore, each fundamental garment is annotated with its garment parts.

B. DATASETS FOR FASHION CLASSIFICATION

The identification of clothing in image, is called Fashion classification. There are not many datasets that have been created specifically for this task, but certainly one of the most important and most used is **Fashion Mnist**. Created by Xiao *et al.* [32] in 2017, Fashion-MNIST dataset is proposed as a more challenging replacement dataset for the MNIST dataset, that consists of 10-class handwritten digits [61]. The images within this dataset come from the shopping website named Zalando. To construct the dataset, they used miniature images of 70 000 unique products. Those products can contain different gender groups: men, women, kids and neutral. Moreover, white-color products have not been placed inside the dataset since they have not a high contrast to the background. The manually labeled silhouette code of products is used for the labels of class. Each product contains only one silhouette code, for a total of 10 classes (0 = T-Shirt/Top, 1 = Trouser, 2 = Pullover, 3 = Dress, 4 = Coat, 5 = Sandals, 6 = Shirt, 7 = Sneaker, 8 = Bag, 9 = Ankle boots). Examples of images belonging to the dataset are showed in Figure 5 Another dataset used for classification task is CBL. Created by Liu *et al.* in [54] this dataset is composed by 250 000 images manually label extracted from 25 clothing brands; after the labellization phase, 57 000 images with clear logos are kept to form the

CBL Dataset and all of them contains brand and bounding box information.

C. DATASETS FOR CLOTHES GENERATION

Given an image that contains a person, the aim of Clothes Generation is re-wear that person with a different clothing style. It can be done by taking a realistic image containing fashion items and synthesizing it. In this section, the datasets using for this task are presented.

- **UT-Zap50K (2014):** The UT-Zap50K dataset, introduced by Yu and Grauman in [15] contains 50.025 shoes images coming from Zappos.com. There are 4 relative attributes, open, pointy, sporty, and comfort, and for each attribute, there are 3 000 annotated pairs. Shoe images was annotated using metadata: for example, these metadata can be the gender, the type, the materials and the manufacturer.
- **LookBook (2016):** Yoo *et al.* created a dataset called LookBook. This dataset includes two fashion domains. In the first domain, the images represent fashion models, and in the second domain the images represent top products with a clean background. The authors manually associated each product image to the corresponding images of a fashion model that matches the product, so that each pair is exactly associated with the same product. In the end, the dataset consists 84 748 images: 9 732 top product images are linked to 75 016 fashion model images, so that, in average, a product has about 8 fashion model images on average.
- **VITON (2018):** The images within this dataset were collected by Han *et al.* from the shopping website Zalando. They first extracted about 19 000 pairs of images of women and clothes, then removed the noisy images, resulting in 16 253 pairs. These images were then further subdivided into training sets, 14 224 pairs, and into test sets, 2 032 pairs. During the test, the person should wear a different piece of clothing from the original one: then the clothing images present in the test pairs are randomly mixed in the evaluation phase.
- **Fashion-GEN (2018):** The dataset created by Rostamzadeh *et al.* in [40] consists of 293 008 images. All fashion items are photographed from 1 to 6 different angles depending on the category of the item. Each product belongs to a main category and a more fine-grained category (i.e.: subcategory). There are 48 main categories, and 121 fine-grained categories in the dataset.
- **FashionTryOn (2019):** In order to create their virtual try-on dataset, Zheng *et al.* crawled 4 327 clothing items that come from the shopping website Zalando.com, with their corresponding model images. With a preprocessing phase, they removed the images considered noisy, that is, those that show only a part of the human body. Furthermore, they extracted the keypoints of each image using a pose estimator, and removed all the images that had fewer than 10 keypoints. Finally, they create the

FashionTryOn dataset, with a total of 28 714 triplets. Each triplet consists of one image that contains a fashion item and two images that represent the original image that contain a person in a certain pose, and the target person in a different pose.

- *FashionOn (2019)*: Hsieh et. al. created a large-scale dataset that consists of 10 895 in-shop clothes and 10 895 pairs of images of the same person but in two different poses. Moreover, they added the images within the DeepFashion dataset reaching a total of 11 283 in-shop clothes and 11 283 pairs of human images. The resolution of each image is 288×192 . For the training of the network, they created triplets that consists of one in-shop clothing and two images of the same person in different poses.
- *Video VirtualTry-On*: Dong et al. collected a video dataset called Video Virtual Try-on (VVT). They first captured 791 model runway videos with white backgrounds. In addition, they removed the videos and frames that were considered noisy. In each video there are mainly 250 to 300 frames. The total videos were then divided into training set, which contains 661 videos (159,170 frames respectively), and test sets, which contains 130 videos (30 931 frames respectively). Then, 791 images of people and 791 images of clothes were scanned, so that each video could be made by associating it with a new image of person or a new image of clothes. So, for the training of the network, triplets were considered which are composed of one video, one picture of a person and one picture of clothes.

D. DATASETS FOR CLOTHES RECOMMENDATION

With the growths of online shopping platforms and the social network, Clothes Recommendation systems have seen a huge increase. With this kind of systems, the user experience can be improve and it can bring great profit to shopping platforms. This type of service that is offered to the customer also aims to select and display a series of articles that are online and compatible with the choices already made and seen by the customer. Some datasets used for this purpose are listed below.

- *Fashion Style14 (2017)*: Takagi et al. in [33] proposed a dataset for prediction of fashion styles formed by 13.126 images, each one corresponding to one of 14 modern fashion styles (conservative, dressy, ethnic, fairy, feminine, casual, retro, rock, etc.).
- *Polyvore (2017)*: Han et al. in [34] collected their own dataset from Polyvore, a popular fashion website, containing 21 889 outfits with 8 categories. Each sample consists of one product image and the corresponding text description.
- *Polyvore-T (2019)*: Wang et al. in [44], using Polyvore [34] dataset as basis, build a type-labeled fashion outfit dataset. They identified type information for each item by grouping 381 categories into 5 types: top, bottom, shoes, bag and accessories.

- *iFashion (2019)*: All images in iFashion are provided by Wish. Guo et al. in [45] collected more than a million, precisely 1 012 947, fashion images with multilabel and fine-grained attributes. The attributes of the dataset are: Category (with 105 classes), Color (with 21 classes), Gender (with 3 classes), Material (with 34 classes), Neckline (with 11 classes), Pattern (with 28 classes), Sleeve (with 5 classes) and Style (with 21 classes).
- *Clothes Recommendation Dataset (2019)*: Liu et al. in [46] created the Clothes Recommendation Dataset searching many brands of clothing websites including H&M, Forever21, Superdry etc., then they crawled the clothing images and record the product information as their label at the same time. In this way, 127.824 images with 7 different brands were reserved and for every image, category, color, material, pair and price are specified.
- *FashionAI (2019)*: Zou et al. [47] introduced FashionAI, an high quality fashion dataset. It takes into account 6 categories of women's clothing and 41 sub-categories. This dataset has a hierarchical structure: in fact the categories can be considered the radii and the sub-categories the leaf nodes and each of these has both a dimension and a value. Therefore, the total number of annotations within the dataset is not calculated by adding all attribute values, but by making the product of the number of the attribute values in each attribute dimension. With this process, they obtained 24 different key points and 245 attribute values in 68 attribute dimensions.
- *FashionKE (2019)*: Ma et. al extracted millions posts from the social network Instagram. Both automated and manual filtering are performed sequentially to ensure data quality. Finally, they contributed a large and high quality dataset named FashionKE consisting of 80 629 eligible images.
- *FIT (2020)*: Ma et. al provided a dataset based on social media Instagram, called Fashion Instagram Trends (FIT). Specifically, they extracted millions of posts uploaded by users around the world, running automatic and manual filters on these posts. First, they detected the person's body and face using the pre-trained object detection model. Images that do not include the face or body or those in which the face and body were of abnormal size were discarded and filtered. In this way they collected about 680.000 images.

IV. DEEP LEARNING METHODS

To understand and estimate trends affecting the world of fashion, several tasks must be overcome. These are summarized in Figure 6 and explained in detail in Section IV-A, Section IV-B, Section IV-C, Section V-A and Section V-B.

Given the huge amount of data that has been collected within the datasets concerning Fashion, only Deep Learning methods have been analyzed. In fact, these methods are the best performing both in terms of efficiency and time.

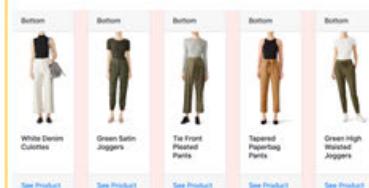
TASK	INPUT	OUTPUT										
LANDMARK DETECTION DATASET  GENERALIZATION 												
CLOTHES PARSING DATASET  GENERALIZATION 		 <div style="border: 1px solid black; padding: 5px; margin-left: 10px;"> <ul style="list-style-type: none"> ■ null ■ bag ■ coat ■ hair ■ jeans ■ shoes ■ skin ■ sunglasses ■ t-shirt </div>										
PRODUCT RETRIEVAL DATASET  GENERALIZATION 												
FASHION CLASSIFICATION DATASET  GENERALIZATION 	   	 <p>Trousers (1)</p>  <p>Ankle boot (9)</p>  <p>Pullover (2)</p>  <p>T-shirt/top(0)</p>										
CLOTHES GENERATION DATASET  GENERALIZATION 												
CLOTHES RECOMMENDATION DATASET  GENERALIZATION 		 <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center; padding: 5px;">  White Denim Culottes </td> <td style="text-align: center; padding: 5px;">  Green Satin Joggers </td> <td style="text-align: center; padding: 5px;">  Tie Front Plaid Pants </td> <td style="text-align: center; padding: 5px;">  Tapered Paperbag Pants </td> <td style="text-align: center; padding: 5px;">  Green High-Waisted Joggers </td> </tr> <tr> <td style="text-align: center; padding: 5px;"> See Product </td> <td style="text-align: center; padding: 5px;"> See Product </td> <td style="text-align: center; padding: 5px;"> See Product </td> <td style="text-align: center; padding: 5px;"> See Product </td> <td style="text-align: center; padding: 5px;"> See Product </td> </tr> </table>	 White Denim Culottes	 Green Satin Joggers	 Tie Front Plaid Pants	 Tapered Paperbag Pants	 Green High-Waisted Joggers	See Product				
 White Denim Culottes	 Green Satin Joggers	 Tie Front Plaid Pants	 Tapered Paperbag Pants	 Green High-Waisted Joggers								
See Product	See Product	See Product	See Product	See Product								

FIGURE 6. Fashion-related social media pictures tasks. A score from 1 to 5 is given for the number of datasets and methods developed for a specific task. Furthermore, it is possible to find an example of the image given in input, and the image in output, after being processed through deep learning methods.

A. OBJECT DETECTION

In the field of computer vision, one of the main problems is object detection. As explained by Zou *et al.* in [62], object detection is an important problem which consists on identifying instances of objects within an image and classifying them as belonging to a certain class. The object detection determines the location and size of object detected.

The models for object detection task can be divided in two macro categories: two-stage detectors and one-stage detectors. In the first case, these models divide the task of identifying objects into several phases, following a “coarse-to-fine” policy. In the second case, the process of these models tries to complete the detection in a single step with the use of a single network. Below, some of the most famous methods for the purpose of object detection will be reported:

- R-CNN (Two-stage-Detector). Region Based Convolutional Neural Networks (R-CNNs) follow a relatively simple process. In fact, they begin by extracting a set of object candidate boxes using a selective search [63]. Therefore, for each proposal, a fixed-size image is cut out and fed to a trained CNN to extract its fundamental characteristics. In the end, linear SVM classifiers are used to decide if the object is present in any specific region and to recognize the categories of the objects found. With selective search, the subdivision of the areas to be proposed takes place in a hierarchical manner, so as to capture the presence of objects in various poses and sizes. The R-CNN models have an important contraindication: the large number of features to be classified, resulting from the overlapping of the many proposed areas (more than 2,000 areas for each image). This leads the model to have to process a large amount of data, compromising its performance.
- SPP-Net (Two-stage-Detector). In 2015 He *et al.* have proposed Spatial Pyramid Pooling Networks (SPP-Net) [65]. Before SPP-Net, CNN’s models required fixed-size input. This entailed a loss of accuracy in the detection of images of different sizes and proportions from those set, which had to be resized, with possible loss of detail or deformation of the image, or even cut. The reason is that, by their nature, CNNs are composed, in the last levels, of fully connected layers, which work on an input of a predetermined size. The innovation of SPP-net lies precisely in having introduced, between the convolutional levels and those fully connected, a level of Spatial Pyramid Pooling, which allows you to put together the features highlighted by the convolutional levels and to return an output of a predetermined size. Using SPP-Net for object detection, the set of proposed areas is calculated starting from the features extracted from the first convolutional levels. Subsequently, representations of predetermined dimensions are generated

on the proposed regions. These representations can therefore be fed to the classifiers.

Dong *et al.* [66] combine VGG-19 network with the SPP-Net to recognize the clothing image style.

- Fast R-CNN (Two-stage-Detector). Fast R-CNN, proposed by Girshick [67] allows to train a recognizer and a bounding box designer within a single model simultaneously. The loss function of Fast R-CNN takes into account the error made in each phase of forward propagation. Fast R-CNN takes as input an image and a set of object proposals, i.e. areas that are supposed to contain objects within the image; the network processes the image through a succession of convolutional layers and max pooling layers to extract its features and produce a convolutional feature map. At this point, a pooling layer, called RoI Pooling Layer (Region of Interest), extracts a vector of a predetermined size from the newly obtained map and processes it through two entirely connected layers. Each vector thus obtained propagates in two directions: in both cases it passes through a series of entirely connected levels; in the first case, the output then passes through a softmax level to estimate the probabilities, for each of the K classes of recognizable objects, that in the area (vector) there is an object of the k-th class; in the second, the levels produce four real numbers for each of the K classes of objects. These values encode, for each recognizable class, the center and the dimensions of the corresponding bounding box.

For example, Liu *et al.* show preliminary results using Fast R-CNN trained on their own created DeepFashion dataset [20].

- Faster R-CNN (Two-stage-Detector). In 2016 Ren *et al.*, Shortly after the publication of Fast R-CNN, proposed the Faster R-CNN recognizer [68]. The main innovation brought about by Faster R-CNN is the introduction of the Region Proposal Network (RPN) which implements a very efficient region proposal system, almost without computational costs when compared to previous models. The RPN consists of a series of convolutional levels applied on the feature maps obtained from the initial convolutional levels of the Fast R-CNN network. To generate the proposed areas, S. Ren *et al.* they opted to integrate a small network at the end of the convolutional levels of Fast R-CNN. This network takes as input a feature map of dimension $n \times n$ and is composed of three convolutional levels. A first shared level of dimension $n \times n$ and two entirely connected “twin” levels of dimension 1×1 : once the hyperparameter k is fixed, which denotes the maximum number of proposals to be advanced for each location, one of the two twin levels outputs the coordinates of k “bounding box”; the other twin level returns, for each proposed bounding box, the probabilities that there is an object in it or not.

For example, Kuang *et al.* [43] in their work, trained a Faster R-CNN detector over DeepFashion dataset to do

- the detection on the bounding boxes, and then manually correct them.
- **Mask R-CNN:** Mask R-CNN is an extension to Faster R-CNN. Mask R-CNN adds the mask branch to generate the object mask, and proposes an ROI Align layer that removes the harsh quantization of ROI Pool, and properly aligns the extracted features with the input. Mask R-CNN avoids any quantization of the ROI boundaries or bins. Bilinear interpolation is used by ROI Align to compute the exact values of the input features at four regularly sampled locations in each ROI bin, and aggregates the results (using max or average) [69].
Mask R-CNN is used for example into the work of Yang *et al.* [70]. They present CLDM, a clothing landmark detector based on Mask R-CNN. CLDM detects the functional regions of cloth items, which makes the features extracted from clothes more discriminative.
 - **R-FCN (Two-stage-Detector):** this is a framework called Region-based Fully Convolutional Network (R-FCN) for object detection. It consists of shared, fully convolutional architectures as in FCN. Unlike Faster R-CNN, for each category, the last convolutional layer of R-FCN, proposed by Dai *et al.* [71] produces a total of k^2 position-sensitive score maps with a fixed grid of $k \times k$. Then, a position sensitive ROI pooling layer is added to join the responses from these score maps. Finally, in each ROI, k^2 position-sensitive scores are averaged to produce a vector and softmax responses between categories are calculated. Another convolutional layer is added to obtain class-independent bounding boxes.
 - **FPN (Two-stage-Detector).** Feature Pyramid Networks (FPNs). Before FPN, created by Lin *et al.* in 2017 [72], most of the deep learning-based decoders performed recognition only on features obtained from the last layers of the network. The starting point is the feature map of lower resolution and higher semantic value, i.e. from the one obtained in the last convolutional level (i.e. the tip of the pyramid). To this is concatenated the feature map of the next level of the pyramid, which has a higher resolution but less semantic value. We proceed in this way until we arrive at the feature map which forms the base of the pyramid. The predictions are made on each of the feature maps thus obtained. The result is a pyramid of features, rich in semantics at every level, quickly built from an image, without this having to be resized. For example, Martinsson and Mogren [73] proposed a fully convolutional neural network based on feature pyramid networks to approach the problem of semantically segmenting fashion images into different categories of clothing.
 - **YOLO (One-stage-Detector).** YOLO [74] network was proposed by J. Redmon *et al.* It was the first one-stage detector. The name YOLO stands for “You Only Look Once”. The idea behind this type of network is the following: apply a single model to the entire

image. In fact, YOLO divides the image into regions, predicts the bounding boxes and, for each of them, determines the probabilities of belonging to a certain class, all using a single network. Later J. Redmon made a series of improvements to the network, releasing a second [75], a third [76] and a fourth version proposed by Bochkovskiy *et al.* in [77] recently. These new versions have better recognition accuracy, while maintaining a very high execution speed.

For example, Huang *et al.* [78] used Yolo V2 to present a real-time tracking system from surveillance videos to detect and track the various clothing categories.

- **SSD (One-stage-Detector).** Single-Shot Detector (SSD) [79], was developed by W. Liu *et al.* SSD’s main contribution was the change of perspective towards the generation of bounding boxes. From this set, SSD predicts a deviation for each of these default bounding boxes: in fact it starts from a set of default bounding boxes. For each bounding box, translated by the predicted deviation, the model then carries out the classification. Thanks to the use of different filters, chosen according to the aspect ratio of the image (i.e. width/height), and thanks to the simultaneous application of these filters to “multiple” feature maps, each obtained in a different point of the convolutional levels, SSD also achieves excellent accuracy in predicting object classes.

For example Gabale and Subramanian [80] used SSD to construct a developed version, called CDSSD, to facilitates unsupervised training of the underlying network architecture, with the aim of extract fashion trends from social media.

As already mentioned in section III-A, in the world of fashion when we talk about object detection, we refer to three different tasks: Clothes Landmark Detection, Clothes Parsing and Product Retrieval.

1) CLOTHES LANDMARK DETECTION

Detecting fashion landmarks from an image is a fundamental and practical task, whose goal is to predict the location of useful keypoints defined on fashion items, such as the corners of the neckline, hemline, and cuff.

Extensive research has been devoted to detecting fashion landmarks and has achieved excellent performance.

The first to introduce neural networks for this task was Liu *et al.* in his work [20]. Clothes Landmark detection is seen here as a regression task and they created the FashionNet network for direct regression of landmark coordinates. Liu *et al.* [21] design pseudo-labels to improve the invariability of fashion landmarks. Yan *et al.* [25] combine selective dilated convolution and recurrent spatial transformers to detect clothing landmarks in unconstrained scenes. In all the above methods, the benchmarks are estimated separately for each landmark point, and therefore there is a possibility of detecting ambiguous and inconsistent landmark points from the structure. Inspired by the attentional

mechanism, Wang *et al.* [81] proposed an attentive grammar network with high human knowledge to globally predict landmarks's positions. At the same time, they point out that the fashion landmark regression is a problem with an highly non-linear level and it is very difficult to learn directly. Therefore, they learn to predict a confidence map of the position distribution for each landmark. Chen *et al.* [82] also adopted this method for mode landmark recognition: They proposed a Clothes Landmark Detection network based on Feature Pyramid Network and designed the Dual Attention Feature Enhancement (DAFE) module to improve the feature representations while recovering the size of the feature maps. Li *et al.* [83] inspired by visual attention mechanism [84] and non-local block [85], proposed Spatial-Aware Non-Local (SANL) block, which encodes prior knowledge taking into account a spatial attention map. The more current method proposed by Yu *et al.* [86] define a complicated fashion layout-graph and propose to model the structural layout relationships among landmarks. However, they propagate the information according to a fixed layout-graph and cannot deal with the diverse deformation or occlusion. Recently, Chen *et al.* [87], proposed a novel framework, called Adaptive Graph Reasoning Network (AGRNet), for Clothes Landmark Detection. It introduces graph-based reasoning to adaptive impose structural layout constraints among landmarks on the deep representations. The best results for both FLD Dataset and DeepFashion Dataset are provided by Kai *et al.* [88] with their MDDNet Network: in fact, MDDNet achieves the best NE score in average of 0.0267 on FLD Dataset and of 0.0251 on DeepFashion Dataset compared with other fashion networks.

The latest works developed for this task are those of Kim *et al.* [89] and Song *et al.* [90]. The first is an innovative method based on a one-stage detector that aims to reduce the high computational costs required by large-scale datasets. This network, which is an adaptation of the EfficientDet, developed by Google Brain, can perform two tasks very quickly: the first is that of detecting multiple clothes within the image, while the other is that of identifying landmarks. Through this adaptation, the authors achieved an accuracy of 0.686 mAP in bounding box detection, and 0.450 mAP in landmark identification: the procedure was very fast, there being a very rapid inference time of 42 ms in each single GPU. In this way the authors have tried to solve the problem that arises when large datasets are chosen, that is the balance that must exist between accuracy and speed.

The second work, on the other hand, was proposed to be able to solve the problem of occlusions that can be found in the images. In particular, the authors developed a new Loss function, called Position Constraint Loss (PCLoss) which uses the position relationships between the various landmarks to understand which of these are wrong, regularizing their position using a regularization term for each landmark.

The results of all the methods are reported in Table 5 and Table 6.

TABLE 5. Performance of state-of-the-art methods for fashion landmark detection using FLD dataset. The best performance are highlighted in bold.

Method	L.collar	R.collar	L.sleeve	R.sleeve	L.waist	R.waist	L.hem	R.hem	avg.
FashionNet [20]	0.0784	0.0803	0.0975	0.0923	0.0874	0.0821	0.0802	0.0893	0.0859
DFA [21]	0.048	0.048	0.091	0.089			0.071	0.072	0.068
DLAN [25]	0.0531	0.0547	0.0705	0.0735	0.0752	0.0748	0.0693	0.0675	0.0672
AFGN [81]	0.0463	0.0471	0.0627	0.0614	0.0635	0.0692	0.0635	0.0527	0.0583
DAFE [82]	0.0366	0.0369	0.0587	0.0573	0.0485	0.0485	0.0504	0.0497	0.0482
SANL [83]	0.0296	0.0298	0.0489	0.0471	0.0402	0.0413	0.0546	0.058	0.0437
LGR [86]	0.0423	0.0152	0.0502	0.0735	0.0195	0.0512	0.0452	0.0393	0.0419
AGR [87]	0.0257	0.0263	0.0429	0.0431	0.0347	0.0343	0.0458	0.0463	0.0374
CPN [91]	0.0497	0.048	0.051	0.0511	0.0351	0.0344	0.0518	0.0527	0.0473
CPN* [91]	0.044	0.0443	0.0491	0.0453	0.0325	0.0321	0.0414	0.0411	0.0414
MDDNet [88]	0.0194	0.0196	0.0372	0.0357	0.0255	0.0254	0.0259	0.0258	0.0267

TABLE 6. Performance of state-of-the-art methods for clothes landmark detection using DeepFashionC dataset. The best performance are highlighted in bold.

Method	Left collar	Right collar	Left sleeve	Right sleeve	Left waist	Right waist	Left hem	Right hem	avg.
FashionNet [20]	0.0854	0.0902	0.0973	0.0935	0.0854	0.0845	0.0812	0.0823	0.0872
DFA [21]	0.0628	0.0637	0.0658	0.0621	0.0726	0.0702	0.0658	0.0663	0.0660
DLAN [25]	0.0570	0.0611	0.0672	0.0647	0.0703	0.0694	0.0624	0.0627	0.0643
AFGN [81]	0.0415	0.0404	0.0496	0.0449	0.0502	0.0523	0.0537	0.0551	0.0484
DAFE [82]	0.0295	0.0297	0.0363	0.0631	0.0311	0.0313	0.0394	0.0402	0.0342
SANL [83]	0.0277	0.0282	0.0391	0.0394	0.0297	0.0299	0.0395	0.0401	0.0342
LGR [86]	0.0270	0.0116	0.0286	0.0347	0.0307	0.0435	0.0160	0.0162	0.0336
AGR [87]	0.0256	0.0251	0.0318	0.0324	0.0271	0.0286	0.0328	0.0341	0.0297
CPN [91]	0.0392	0.0406	0.0416	0.0419	0.0268	0.0266	0.0472	0.0493	0.0408
CPN* [91]	0.0323	0.0325	0.0386	0.0346	0.0256	0.0265	0.0373	0.0355	0.0337
MDDNet [88]	0.0182	0.0186	0.0311	0.0307	0.0227	0.0223	0.0273	0.0273	0.0251

Table 5 presents the Performance of state-of-the-art methods for Clothes Landmark Detection using FLD [21] Dataset in terms of normalized error(NE). Table 6 shows the Performance of state-of-the-art methods for Clothes Landmark Detection using DeepFashion [20] Dataset in terms of normalized error(NE).

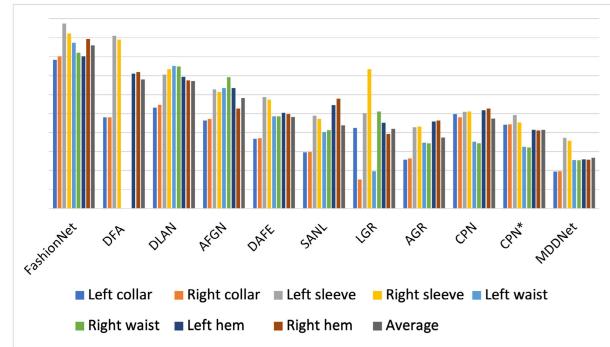


FIGURE 7. Normalized error for clothes landmark detection in FLD dataset.

2) CLOTHES PARSING

The purpose of object parsing is to understand the contents that are inside an image in a detailed way: this is done by segmenting the image into regions that have a different semantic meaning. In particular, fashion parsing and human parsing with clothing classes aims to resolve the problem of finding significant regions within images that contain people with certain clothes on. Similar to the semantic segmentation task, object and label diversity is a challenge not closed for human parsing. And, unlike classic semantic segmentation tasks, such as the parsing of a scene, the purpose of human

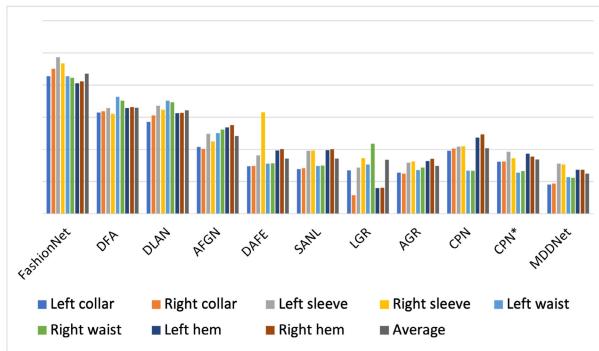


FIGURE 8. Normalized error for clothes landmark detection in deep fashion dataset.

parsing is both to understand the different parts of the person in the input image, and to assign the right label to each clothes that the person wears. Unlike semantic segmentation, human parsing also requires that the methods used to solve this task, can withstand large variation in occlusion, pose, lighting and viewpoints. Therefore, it is not advisable to apply semantic segmentation frameworks directly to human parsing. Moreover, the hand-developed algorithms for human parsing are also not very powerful as they are not robust and inflexible in adaptation.

Yamaguchi *et al.* [8] were the first to tackle the task of Clothes Parsing. In their work, In their work, they considered the problems of clothing parsing and pose estimation, and refined them by considering the relationship between the two. However, this method was mainly based on solving a problem in which the images that were analyzed had been labeled through tags that had been provided by the user and that indicated the articles of clothing depicted. To overcome this limitation, Yamaguchi *et al.* [10] proposed garment parsing using a retrieval-based approach. Given an image as input, the first step were to retrieve similar images from a dataset; the second step was find the closest parsings that were then applied to the final result via dense matching.

Liu *et al.* [92] proposed a quasi-parametric parsing framework called Matching Convolutional Neural Network (M-CNN), which is able to fully utilise the monitoring information from the annotation training data and extend it in the meantime for new added labels.

Inspired by the performance in traditional classification and recognition tasks, Liang *et al.* [16] used Deep Convolutional Neural Network to construct an end-to-end relationship between the input image, that consist of an human image, and the outputs.

Liang *et al.* [93] solved this problem by incorporating the LGLSTM layers into CNNs instead of learning features only from local convolution kernels, as in [17]: this step was done to taking into account both long and short distance spatial dependencies. Moreover, it is possible to store the previous contextual interactions from neighboring locations and the complete image in previous LG-LSTM layers, by adopting hidden and memory cells in LSTMs,. Furthermore, they

TABLE 7. Performance of state-of-the-art methods for clothes parsing using Fashionista dataset. The best performance are highlighted in bold. Blank spaces indicate that the results are not available.

Method	Accuracy	F.G. Acc	Avg Precision	Avg Recall	Avg F1
Yamaguchi et al [8]	87.87	58.85	51.04	48.05	42.87
PaperDoll [10]	89.98	65.66	54.87	54.16	46.80
ATR [16]	92.33	76.54	73.93	66.49	69.30
Co-CNN [17]	96.08	84.71	82.98	77.78	79.37
Co-CNN (more) [17]	97.06	89.15	87.83	81.73	83.78
LG-LSTM [93]	96.85	87.71	87.05	82.14	83.67
LG-LSTM (more) [93]	97.66	91.35	89.54	85.54	86.94
Graph LSTM	97.93	92.98	88.24	87.13	87.57
Graph LSTM (more)	98.14	93.75	90.15	89.46	89.75
Finer-Net [94]	93.37	81.45	86.78	80.98	
Finer-Net+pose [94]	95.38	85.72	87.66	81.75	

introduced Graph LSTM structure to capture long-distance dependencies on the superpixels. As can be seen from tables 7 and 8, the latter method is the best in accuracy and average F1-score. Table 7 proposes Performance of state-of-the-art methods for Clothes Parsing using Fashionista Dataset in terms of Accuracy, F.G.Accuracy, Average Precision, Average Recall and Average F1-Score in percentage. Table 8 exhibits Performance of state-of-the-art methods for Clothes Parsing using ATR Dataset in terms of Accuracy, F.G.Accuracy, Average Precision, Average Recall and Average F1-Score in percentage.

Ye *et al.* [94] introduced a new network called FinerNet, which first segments the human foreground region. The following stage then takes as input the original input image and the results of the last stage as input to attribute finer labels to each pixel. Moreover, by effectively using human posture features, the network can achieve better segmentation results. In fact, FinerNet performs better than the other state-of-the-art methods in FG.Accuracy, Average Precision and Average Recall in the ATR dataset.

The various state of the art methods used for Clothes Parsing have been reported below. In particular, the results were reported on five different types of datasets: Fashionista, ATR, CCP, CFPD and CIHP. Table 9 presents Performance of state-of-the-art methods for Clothes Parsing using CFPD Dataset in terms of Accuracy and IoU in percentage. Table 10 shows Performance of state-of-the-art methods for Clothes Parsing using CCP Dataset in terms of Accuracy, F.G.Accuracy, Average Precision, Average Recall and Average F1-Score in percentage. Table 11 proposes Performance of state-of-the-art methods for Clothes Parsing using CIHP Dataset in terms of Pixel Accuracy, Average Accuracy and Average IoU in percentage;

3) PRODUCT RETRIEVAL

Given the rapid development of e-commerce sites, which has resulted in an increase in online shopping, many researches have dealt with the task of product retrieval based on images

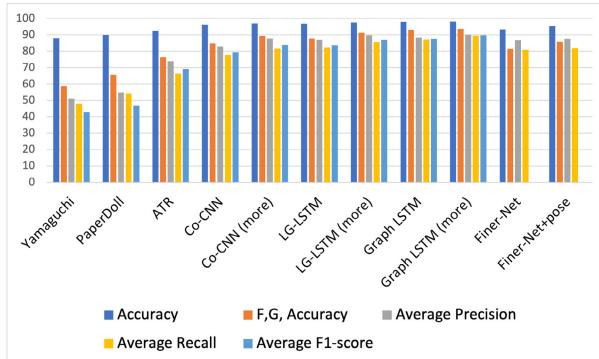


FIGURE 9. Evaluation of clothes parsing in Fashionista dataset.

TABLE 8. Performance of state-of-the-art methods for clothes parsing using ATR dataset. The best performance are highlighted in bold. Blank spaces indicate that the results are not available.

Method	Accuracy	F.G. Acc	Avg Precision	Avg Recall	Avg F1
Yamaguchi et al. [8]	84.38	55.59	37.54	51.05	41.80
PaperDoll [10]	88.96	62.18	52.75	49.43	44.76
M-CNN [92]	89.57	73.98	64.56	65.17	62.81
ATR [16]	91.11	71.04	71.69	60.25	64.38
Deeplab-v2 [95]	94.42	82.93	69.24	78.48	73.53
Co-CNN [17]	95.23	80.90	81.55	74.42	76.95
Co-CNN+ [17]	96.02	83.57	84.95	77.66	80.14
LG-LSTM [93]	96.18	84.79	84.64	79.43	80.97
LG-LSTM+ [93]	96.85	87.35	85.94	82.79	84.12
CRFasRNN [96]	96.34	85.10	84.00	80.70	82.08
Graph LSTM [93]	97.60	91.42	84.74	83.28	83.76
Graph LSTM+ [93]	97.99	93.06	88.81	87.80	88.20
CP-Net [97]	96.46	90.38	83.27	80.03	81.00
FinerNet [94]	94.44	95.08	94.98	95.14	
FinerNet+ [94]	94.45	98.10	95.24	94.87	
PSPNet [98]	95.20	80.23	79.66	73.79	75.84
DeepLabV3+ [99]	95.96	83.04	80.41	78.79	79.49
TGPNNet [100]	96.45	87.91	83.36	80.22	81.76
PGECKNet [101]	97.03	89.01	86.61	84.31	85.23
Tao et al. [102]			85.4	86.9	86.1

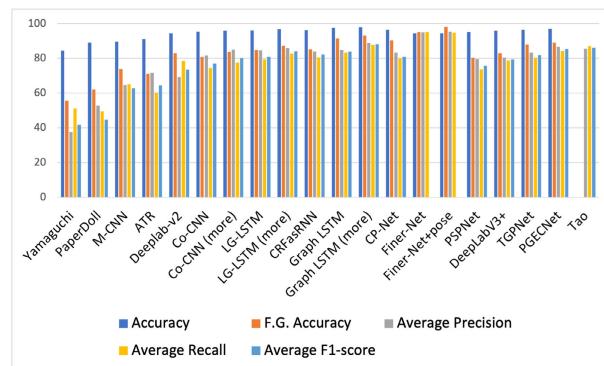


FIGURE 10. Evaluation of clothes parsing in ATR dataset.

or videos. This type of study manages to make consumers and the computer interact: given an input image, in fact, the consumer is allowed to be able to provide additional information on the desired attributes. Although this is a very

TABLE 9. Performance of state-of-the-art methods for clothes parsing using CFPD dataset. The best performance are highlighted in bold. Blank spaces indicate that the results are not available.

Method	Accuracy	IoU
Outfit Encoder [103]	92.3	54.7
PaperDoll [10]	87.1	
SSL [26]	88.5	49.1
DeepLabV2 (ResNet) [95]	89.9	48.3
DeepLabV2 (VGG) [95]	89.2	47.2
FCN-8s [104]	91.6	51.2
Khurana et al. [105]	93.5	58.7

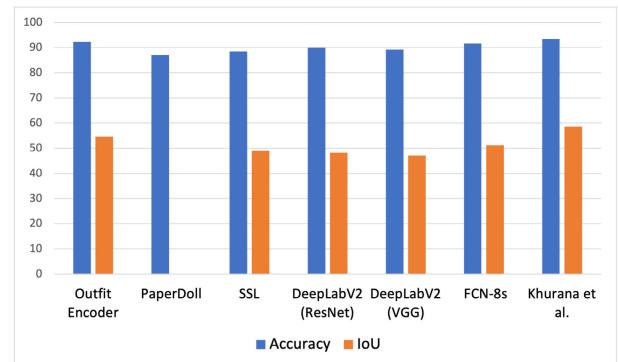


FIGURE 11. Evaluation of clothes parsing in CFPD dataset.

TABLE 10. Performance of state-of-the-art methods for clothes parsing using CCP dataset. The best performance are highlighted in bold. Blank spaces indicate that the results are not available.

Method	Accuracy	F.g. Acc	AVG. Precision	AVG. Recall	AvG F1	AVG IoU
DDN [106]	87.68	60.63	69.91	31.13	62.31	48.96
PCF [8]	87.24	54.76	63.47	58.94	57.03	45.65
Paper	91.13	70.72	71.53	78.54	73.54	60.04
Doll [10]						
CRF [107]	92.16	76.06	77.85	77.23	77.50	64.80

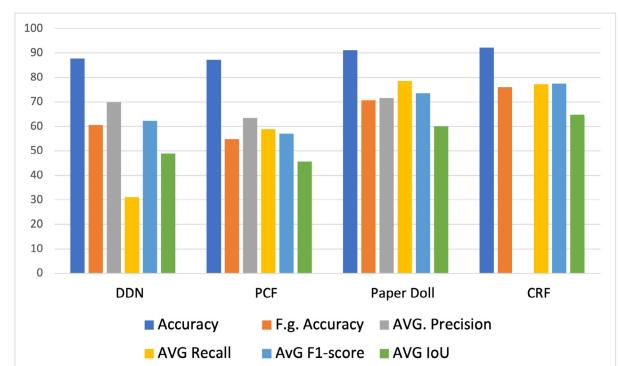


FIGURE 12. Evaluation of clothes parsing in CCP dataset. Blue = Accuracy; Orange = F.G. Accuracy; Gray = Average Precision; Yellow = Average Recall; Light Blue = Average IoU; Green = Average F1-Score.

recent task, the first work is in fact that of Wang *et al.* [115] in 2019, many systems have already been developed.

Wang *et al.* [115], started from the sketch-based-image performance retrieval (SBIR) method, and developed it by adding a re-ranking approach based on multi-clustering. Furthermore, they propose an unsupervised method using blind feedback, in order to make the re-ranking approach

TABLE 11. Performance of state-of-the-art methods for clothes parsing using CIHP dataset. The best performance are highlighted in bold. Blank spaces indicate that the results are not available.

Method	Pixel Accuracy	AVG Accuracy	Avg IoU
DeepLab (ResNet-101) [108]	84.09	55.62	44.8
MMAN [109]			46.81
MuLA [110]	88.5	60.5	49.3
PSPNet [98]	86.23	61.33	50.56
SPUNet	85.86	66.3	51.21
JPPNet [111]	86.39	62.32	51.37
CPUNet	86.45	64.6	52.77
CE2P(wflip) [112]	87.37	63.2	53.1
BraidNet [113]	87.6	66.09	54.42
DPUNet [114]	87.65	69.49	56.21

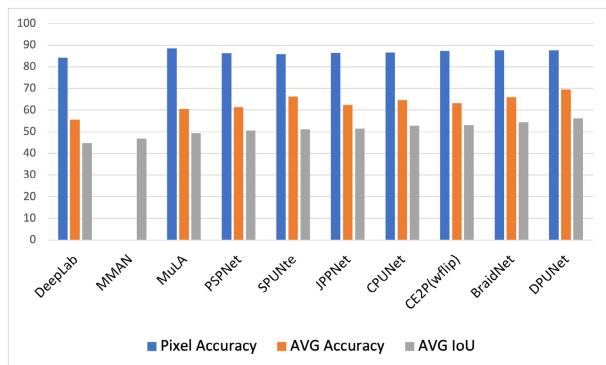


FIGURE 13. Evaluation of clothes parsing in CIHP dataset.

adaptive to different types of image datasets and invisible to users.

The paper proposed by Peng and Chi [116], uses the Domain Adaptation with Scene Graph (DASG) approach: the purpose of this method is which transfer knowledge from the source domain to improve cross-media retrieval in the target domain.

In the study conducted by Nie *et al.* [117], the authors proposed an end-to-end deep hashing method called deep multiscale fusion hashing (DMFH) to perform the cross-modal retrieval task. In particular, they built different network branches for two modalities and then used multiscale merging models for each branch: this was done to merge multiscale semantics which can then be used later to explore semantic relevance.

A novel deep hashing method, proposed by Wang *et al.* [118], is based on pairwise similarity-preserving quantization constraint, referred to as Deep Semantic Reconstruction Hashing (DSRH). In order to learn compact binary codes, they developed a high-level semantic affinity within each data pair.

Some works, such as those that can be found in [18], [20], [119]–[121], improve the performance for the task of product retrieval in fashion word by including supplementary semantic information. Instead other works, such as [122]–[125], concentrate the attention on training a mode retrieval model with losses that were specially designed. There have also been works that have tried to optimize the

representation of characteristics instead, such as [122], [126], [127]. The work developed by Ji *et al.* [128] employed the attention mechanisms in Fashion Retrieval focusing on some significant regions of the image.

FashionNet, proposed by Liu *et al.* in [20], to perform this type of task, includes attribute and landmark information.

The method proposed by Su *et al.* [129] is the best among the other methods mentioned before: the novelty is that it integrates the attribute and landmark information with a bilinear attention pooling module.

The most recent works that have been developed in the field of fashion retrieval are three. The first is that of Sharma *et al.* [130]. The difference with the previous methods is the fact that they used two different datasets: the source dataset and the target dataset. In this way, they built a cross-domain retrieval model, trained on the source dataset, and tested to a new unlabeled dataset. Thus the entire model is unsupervised.

Instead D’Innocente *et al.* [131] proposed a method in which an image and the position of the points of interest that identify the attributes required within the image are passed in input. To achieve this representation, points of interest are mapped into a coordinate system using bilinear interpolations. The generated feature map is then passed through a convolutional layer. The model is then driven by a loss function called localized triplet loss [132], which searches for similar images, considering the similarity between similar points of interest. Similarly, Dong *et al.* [133] have proposed a network that is made up of two branches: the first branch takes the whole image as input, while the second takes as input only the part of the image that is of interest. This crop is obtained using a specific location method. The joined network was called Attribute-Specific Embedding Network (ASEN).

Tables 12 and 13 show the results of state-of-the-art methods with respect to Top-20 Accuracy. Given a query image, top-20 accuracy is calculating using the Euclidean distances between the query image and all images in the gallery set. In this way, top-20 accuracy is the ranking in an ascending order of the distances. If the ground-truth gallery image is found in this ranking, the retrieval will be considered as a success. Table 12 is more detailed as it shows precisely the top 20 accuracy per attribute, i.e. Dress, Leggings, Outwear, Pants, Skirts and Tops. In both cases the best performances are those of the AHBN method [129]. Table 13 presents the Performance of state-of-the-art methods for Product Retrieval using DeepFashion with Consumer To Shop Benchmark (Table (a)) and with In-Shop Benchmark (Table (b)).

B. FASHION CLASSIFICATION

ML and DL techniques bring great benefits to image recognition and classification in the fashion environment. In fact, they can help to improve the user experience [147], which is a fundamental factor for the calculation of the Key Performance Indicator (KPI), which can be measured

TABLE 12. Performance of state-of-the-art methods for product retrieval using exact Street2Shop dataset. The best performance are highlighted in bold. Blank spaces indicate that the results are not available.

Method	Dress	Leggins	Outwear	Pants	Skirts	Tops
WTBI [19]	0.371	0.221	0.21	0.292	0.546	0.381
Impdrop+	0.621				0.709	0.523
GoogLeNet [134]						
Xiong et al [135]	0.583		0.509		0.736	0.47
Jiang et al [136]	0.212	0.233	0.224	0.322	0.103	0.174
R.Contrastive+ [123]	0.592	0.201	0.207	0.213	0.498	0.471
Attribute [123]						
GRNet [43]	0.642		0.386	0.485	0.725	0.583
AHBN [129]	0.712	0.469	0.523	0.561	0.753	0.639

TABLE 13. Performance of state-of-the-art methods for product retrieval using DeepFashion with consumer to shop benchmark (table (a)) and with in-shop benchmark (table (b)). The best performance are highlighted in bold.

(a) (b)

Method	Top-20 Accuracy	Method	Top-20 Accuracy
CtxYVGG [128]	0.479	Studio2Shop [139]	0.818
Liu et al. [21]	0.510	DREML [140]	0.958
Verma et al. [137]	0.253	VAM [134]	0.923
R.Contrastive [123]	0.230	Weakly [119]	0.781
AMNet [138]	0.338	Zhao et al. [141]	0.958
FashionNet [20]	0.188	Verma et al. [137]	0.784
GRNet [43]	0.644	BIER [142]	0.952
AHBN [129]	0.603	HDC [143]	0.890

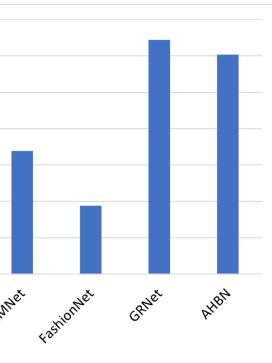


FIGURE 14. Evaluation of product retrieval in deep fashion (CTB) dataset, in term of top20-accuracy.

through factors such as the time spent by the user in front of the computer, the purchase volume and average checkout value.

Deep Learning methods, and in particular Convolutional Neural Networks, can help the user to have a more pleasant experience on the site, being able to make a quicker and more convenient search of the products. As a consequence, there will be an increase in KPIs, in the business profits and in the efficiency of the product management system. An online store that is multi-brand has to group the products and establish the rules necessary for unification and quality

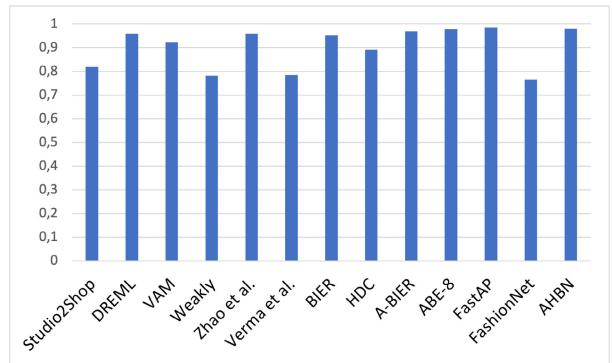


FIGURE 15. Evaluation of product retrieval in deep fashion (CIB) dataset, in term of top20-accuracy.

standard. When a brand shop proposes some products to the multi-brand online store, the manager reviews the incoming products and decides whether to approve or reject them. This methodology meets two different problems. The first problem is that paying the individual people who carry out the supervised learning process becomes a very expensive process. The second problem is that the time frame for carrying out this type of human reviews of different products in different stores is very long. One way to reduce costs and times, and consequently increase the performance and quality of the results, is the use of automatic systems based on CNN.

Considering the importance of clothing in society, there are many applications for Fashion Classification. An example is the prediction of clothing details in an image, that can help find similar clothing items in a dataset from e-commerce sites. Analogously, Fashion Classification based on user preferences can be used to provide recommendations to the user.

Some problems and issues must be considered in the Fashion Classification, to make these applications effective. In particular, the difficulties caused by the clothing property must be considered: [148]–[151]:

- 1) Same clothing can be considered different depending on the point of view, and different clothing can be considered the same (the lower part of a dress that is particularly short can be classified as a skirt);
- 2) Clothing can be easily deformed by stretching or folding;
- 3) A picture of clothing can change; for example, the images can only contain the type of fabric, or models wearing a dress with that same fabric.;
- 4) The images can be very different from each other, in the sense that they can have many different conditions, including different angles and lighting, cluttered backgrounds, and partially hidden by other objects or people;
- 5) Some classes of clothing have almost identical features and can be confused with each other. For example, the pants and tights classes are two classes that are very similar to each other and very difficult to distinguish;

- 6) some clothing classes are very difficult to identify. For example, this may be due to their small size, such as accessories.

Therefore, algorithms that achieve high classification performance for multi-class fashions are needed. For these reasons, DL methods and specially CNN are the most commonly used applications for this task.

1) CONVOLUTIONAL NEURAL NETWORKS FOR CLASSIFICATION TASK

The first CNN architecture to be known and enhanced is AlexNet [152]. This network classified the images within ImageNet dataset: in this dataset there are 15 million of images and 22 000 categories. AlexNet had a total of 8 layers: 5 convolutional layers, some of this followed by dropout layers or max-pooling layers, and 3 fully connected layers. As activation function, the Rectified Linear Unit function (ReLU) is used, which improved the performance in terms of speed over the Hyperbolic Tangent (tanh) function and the sigmoid function. The data augmentation is performed using patch extractions, translation of the image and horizontal mirroring. The best results of this network is an error rate of 16.4% for the top-5 test.

Another network developed and used for the classification task is ZFNet [153]. This network improves the results of AlexNet by reaching an error rate on top-5 test of 11.7%. There are one main difference between ZFNet and AlexNet: AlexNet uses a filter of size 11×11 in the first layer, instead ZFNet utilizes a smaller filter size (7×7) with reduced stride value. This choice was made for a main reason, that is to be able to keep much more information about the input volume inside the layers. Another contribution of ZFNet is the possibility of visualization of the weights and filters into the architecture: in fact, the authors in fact, using a deconvolutional network, have developed a new visualization technique: this will allow to find dissimilar feature activations and the relationship with the input. Also this network, as previously done by AlexNet, uses ReLU as the activation function, and the categorical cross-entropy as loss function.

In 2014, VGG network was introduced by Simonyan and Zisserman in [154] with 7.31% error rate. This network is composed by 16 (VGG16) or 19 (VGG19) convolutional fully connected layers: the filters have size 3×3 and the pooling layers have size 2×2 . Using a smaller filter many times is convenient as this reduces the number of parameters, but does not decrease performance. Moreover, at the end of each pooling layer, the number of filters is doubled. In this way the spatial dimension decreases, while the depth increases. The data augmentation is performed in this network through scale jittering, and ReLU function is used as activation function.

To improve the performance of VGGNet, Szegedy *et al.* [155] built GoogLeNet, with an error rate of 6.7%. This network introduced for the first time the concept of Inception Module, i.e. a parallel layer structure. This module consists of parallel connections with filters of

different sizes (1×1 , 3×3 , 5×5): filters of different sizes are used to be able to process the input image in scales with different dimensions. Max pooling layers, with a sized of 33×33 , are merged to each parallel connection. The output of these layers are then concatenated for the module output. Furthermore, a bottleneck layer with size 1×1 is applied: this was done to decrease the number of channels and the number of weights for each filter. At the end, VGGNet is composed by a total of 22 layers: at the beginning there are 3 convolution layers; then 9 inception layers and each of these is followed by 2 convolution layers, and 1 layer fully-connected.

One of the most important network for the classification task is ResNet, created by He *et al.* [156]. It reached an error rate of 3.57%. This network has 152 layers, and thanks to the residual learning, it can go up to the layers deeper without degrading the output.

In addition, several methods with different configurations have been developed. Furthermore, the networks can be integrated with other methods. the outcome is a hybrid method.

The most recent work in the field of classification is the one conducted from Kuang *et al.* [157]. The authors have built a hierarchical system that manages to classify clothes within large datasets. In particular, using multiple deep CNN branches, they performed the task of classification. In addition, to be able to improve performance, they added a hierarchical method, and finally applied it to the recommendation task.

2) PERFORMANCE COMPARISON ON FASHION-MNIST DATASET

A lot of CNN architectures, such as LeNet [61], Alex Net [158], Google Net [155], VGGNet [154] and ResNet [156], some of these already mentioned in the previous paragraph, have been used in image classification.

In addition to these other types of work have been carried out for this task. Convolutional Neural Network are trained to classify images of Fashion MNIST dataset. McKenna [159] proposed a model to add and compare the sigmoid features, ELUs and ReLUs of the missing benchmarks in the Fashion MNIST dataset. First, the missing multilayer non-convolutional feed-forward neural networks in Fashion-MNIST as a benchmark. Then, testing the effectiveness of the contemporary activation features (compared to ELU, ReLU and Sigmoid).

Xiao *et al.* [32] created a dataset where the Fashion MNIST images are converted into the format corresponding to the MNIST dataset, which is easier to perform. Greeshma and Sreekumar [160] suggested a system to classify the fashion items in the Fashion MNIST dataset using HOG features and a multi-class SVM as classifier of the network.

Table 14 shows the performance comparison between some of the methods present at the state of the art for Fashion Classification, using Fashion-MNIST Dataset. The best performance on this dataset is provided by LeNet-5 Network [161]. This network has the following architecture:

TABLE 14. Performance of state-of-the-art methods for fashion classification using fashion-MNIST dataset. The best performance are highlighted in bold.

Methods	Accuracy (%)
3-layer Neural Network	87.23
SVC with rbf kernel	89.70
Evolutionary Deep Learning Framework	90.60
CNN - SVM activation function	90.72
CNN - Softmax activation function	91.86
CNN - Batch normalization	92.22
CNN - Batch Normalization & Residual skip	92.54
Decision Tree Classifier	79.80
ExtraTreeClassifier	77.50
GaussianNB	88.00
KNeighborsClassifier	85.40
LinearSVC	83.60
LogisticRegression	84.20
RandomForestClassifier	87.30
SGDClassifier	81.90
SVC	89.70
CNN	91.61
CNN-LeNet-5	98.80
Multilayer perceptron	78.33

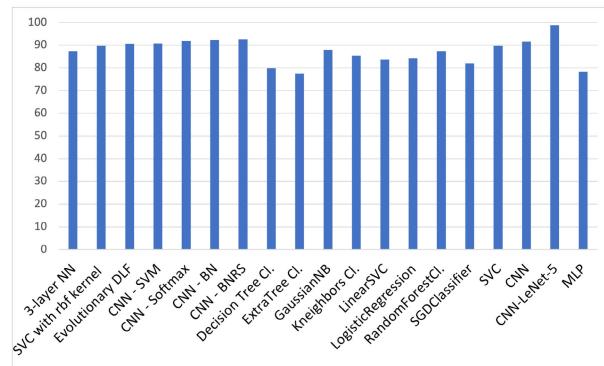


FIGURE 16. Evaluation of fashion classification in FashionMnist dataset.

3 convolutional layers, 2 subsampling layers and 2 fully connected layers. At first, a convolutional layer takes in input an image with 32×32 size and in gray scale. Then there are six 5×5 convolutional filters with a stride of 1. The second convolutional layer apply a filter of size 2×2 with a stride of 2, after the input is passed through a layer of pooling. These kind of layers are then connect to the first, the second and the fifth convolutional layers. A fully connected layer is added, the last layer is connected to a softmax that will give the image classification output. This network reach an accuracy of 98, 80%.

C. CLOTHES GENERATION WITH GANS

In recent years, one of the most developed topic in the Deep Learning world, was Generative Adversarial Networks (GANs), created by Goodfellow in 2014 [162]. Their importance is due to the fact that they have proven to be excellent in many areas, especially in the generation of images [163], [164] and in image processing [165], [166]. They have also shown interesting results in generating novel images, e.g. faces [164], indoor scenes [167] and fine-grained objects [24], [168].

For the purpose of this paper, the main argument is focused on Clothes generation with GANs [41], [169], [170].

The task consists in taking two images as input: the first contains the image of the clothes to try on, while the other contains the image of a person who will be in a certain pose and who is already wearing clothes. This is a very difficult task especially because the pose of the person can bring various problems, for example some parts of the body may be hidden or even the pose does not allow you to see some of them. Existing methods perform this task using three different networks, each of which is used for a specific purpose. The first network will have the task of carrying out a similar transformation to be able to align the desired clothes in the desired person; the second network will have the task of dressing the person; the last network will have the accomplishment of carrying out a post-processing phase to try to make the final image as realistic as possible.

In this context, Yoo *et al.* [24] produced a dressed person determined by clothes image and vice versa, without considering the person's pose.

Another work, proposed by Lassner *et al.* [139] it is based on the production of a generative network of dressed people.

FashionGAN, created by Zhu *et al.* [169], uses textual descriptions to succeed in replacing a dress that a person is already wearing with another.

The first to use the generative network to resolve the task of text-to-image synthesis, capable of generating low-resolution images were Reed *et al.* [168]. StackGAN++ [171] has the purpose to produce images as realistic as possible, using tree methods with multiple discriminators and generators. Another model, with a structure similar to StackGAN++, is AttnGAN [172]. It is composed by additional attention modules and deep similarity model. Emir *et al.* developed a new adversarial network called e-AttnGAN [173] which includes an integrated attention module that, during the image generation process, incorporates word and sentence context characteristics. This image generation process is done by Feature-wise Linear Modulation (FiLM) layers [174] that can control the visual features without using a supervised approach.

Viewing the successes of GANs, conditional GANs were developed by [175], [176]: they incorporate a specific conditional constraint in such a way that it acquires knowledge from specific condition given to the network, in order to produce realistic fake images. This type of GANs include, as conditional input, a binary mask: it is done by connecting it with the input image, as in the work proposed by Ma *et al.* [177], or with a latent noise vector, as in the work developed by Park *et al.* [178].

Within fashion environment, generative networks have been used to exchange the clothes of a person in a source image with some other clothes on a target image. This task is known as Virtual Try On Networks (VTON) [41], [179], [180]. These methods generally have a the same structure that consists of three main components executed by different networks:

- a network that has the task of alignment of the target cloth to the source image by learning an affine transformation [181];
- a network that has the task of stitching and swapping, that is usually a GAN;
- a refinement network, that changes from method to method.

Pandey and Savakis [182] introduced the network poly-GAN which allows to build the output starting not only from two but using more inputs and it can be used to perform different tasks. This is a very important network as it is the first example of a network that performs all three tasks that have been described above.

Jiang *et al.* [183] proposed two different framework: the first performs fashion style generator task, called FashionG, for single-style generation; the second performs a spatially constrained FashionG, called SC-FashionG, for mix-and-match style generation. Both this network are end-to-end forward networks which comprise a fashion style generator and a patch-global style discriminator. The inputs are composed by cropped images of online shopping apparel products and full images of street fashion photographs.

The work proposed by Liu *et al.* [184] investigates clothing match rules based on semantic attributes according to the GAN model. Specifically, an attribute GAN was proposed to automatically generate clothing match pairs. The core of Attribute-GAN constitutes training a generator, supervised by an adversarial trained collocation discriminator and attribute discriminator.

1) VIRTUAL TRY ON

To perform the virtual try on task, which is very competitive especially in the last few years, many works have been carried out.

Han *et al.* [41] presented a network called VITON that uses a coarse-to-fine for seamlessly transfer a desired garment to the right area of a person within an image using only 2D information. In this process, a rough fitting result is first generated and the mask for the garment is predicted. To make the result more accurate they then introduced a refinement network, using the mask and the coarse result. However, this kind of framework does not work well when there is a large deformation. To solve this problem, Wang *et al.* [179] introduced a novel architecture called CP-VTON, that means Characteristic-Preserving VTON. Using a geometric matching module, it can better manage the spatial deformation by aligning the input clothing with the body shape.

A development of these works are FashionGAN [169] and M2E-TON [185]. FashionGan creates the image based on the text description: with as input a fashion image and a sentence that describe a dress other than that worn by the person, this generative network manages to dress the person in the manner described by the sentence. A GAN creates a segmentation map according to the description, and another GAN, driven by the segmentation map, generates the output

image. On the other hand, M2E-TON creates the image based on the model's image: it manages to dress one person with the clothing of another, whose image is passed on as input. Also these two people can have different poses.

Fit-Me [186] is the first work that performs the virtual test allowing arbitrary poses. In fact, the authors designed an architecture from coarse to fine both for the transformation of the pose and for the virtual test.

Hsieh *et al.* in [50], created FashionOn, which applies semantic segmentation and refines the face part and clothing region, aiming to achieve a more realistic output image and solve the human limb occlusion problem in CP-VTON.

ClothFlow [187], proposed by Han *et al.*, focuses on clothing regions for more natural results.

Unlike the works described so far, which work on images, FW-GAN, presented by Dong *et al.* [51], is a model that learns to generate a video of a moving person based on a person image, the desired clothes image, and a sequence of target poses. FW-GAN wants to synthesize the coherent and natural video through a manipulation of poses and clothes.

Recently, Fincato *et al.* proposed a new solution for this task, called VITON GT, where GT means Geometric Transformation, where a multi-stage geometric transformation is performed to reduce distortions and artifacts in the generated images. Within this model there are two different parts:

- a two-stage geometric transformation module, where an affine transformation and a warping transformation are performed to shift the desired clothes in the target person and to generate a warped version of the desired clothes;
- a transformation-guided try-on module, where the results are generated.

Furthermore, they added an adversarial training into the second module, to make the output image more realistic, and they To increase the realism of generated images, they integrated adversarial training in the second stage of their architecture and they designed a finetuning scheme to improve the final image quality even more.

2) POSE GUIDED GENERATIVE MODEL

Another very interesting task, which has developed especially over the last few years, is being able to change a person's pose.

The first work dedicated to this type of problem is the one done by Ma *et al.* in [177]. They adopted a divide-and-conquer method, separating the task into two different steps: the first step consists on acquire knowledge of the body structure of the person, and it is performed by a variant of U-Net [188] that combine the desired pose to the person image; the second step consists on learning the particular details of the physical appearance to improve the results using a modified Deep Convolutional GAN (DCGAN) model. The model learns to fill in more appearance details via adversarial training and creates sharper images. Unlike the previous methods that use GANs to generate directly an image, in this method, GANs are used to create a difference map that

connect the image within the target person and the image that comes from the output of the first step.

Balakrishnan *et al.* [189] constructed a network that takes as input two different images: a source image with a source 2D pose, and a desired 2D target pose, and creates an output image. This network first segments the source image into a background layer and multiple foreground layers corresponding to different body parts, allowing it to spatially move the body parts to target locations. The moved body parts are then modified and fused to synthesize a new foreground image, while the background is separately filled with appropriate texture to address gaps caused by disocclusions. Finally, the network composites the foreground and background to produce an output image.

Dong *et al.* [190] proposed a Soft-Gated Warping-GAN to be able to solve the problems that are created when the geometric transformations create the poses, in particular the problems of spatial misalignment. This method consists in two different steps.

- To create a part segmentation map, there is a pose-guided parser: when the target pose is known, it can better generate the image with an high-level structure limitation, describing the spatial layouts.
- To provides accurate representation within any segmentation part, a Warping-GAN is used: this type of network can learn geometric connections that exist between the original image and the pose that comes from the predicted segmentation map.

To solve the task of shape-guided image generation, conditioned on the output of a variational autoencoder for appearance, Esser *et al.* [191] presented a conditional U-Net [188].

Lassner *et al.* [192] proposed a generative model of people called ClothNet. This network takes 3D information form an image that contains a body model, and it is data-driven. In their work, they presented two version of this network: a simpler model, called ClothNet-full, that can randomly generate images that contains people from a learned latent space, and a conditional model, called ClothNet-body, that produces random people with a pose similar to the target pose but with different garments.

Ma *et al.* [193] created a structure that can learn manipulable embedding features that came from three different factors: foreground, background and pose. The structure is composed by two principal steps: in the first step, there is a network that can encode the multi-branched reconstructions, to generate the image; in the second step, an adversarial network does the sampling task using mapping functions.

In contrast, the GAN-based method proposed by Siarohin *et al.* [194] is end-to-end trained by expressly considering pose related spatial deformations. In particular, basing on the structural deformations that are presented in the conditioning variables, they proposed crumple zone skip connections which move local information. These layers are employed in a U-Net [188] based generator. Pumarola *et al.* [195] proposed a fully unsupervised GAN

framework that, considering a photo of a person, automatically generates images of that person with a novel camera views and distinct body poses. They proposed a GAN architecture that combines the pose conditional adversarial networks [196], Cycle-GANs [197] and the loss functions used in image style transfer with the purpose to produce novel images of high perceptual quality [198].

Zhu *et al.* [199] created a model that transfer the pose of the person in the target image during the encoding phase. It is done by an attention based progressive system [200]. Both the above methods extract the image features during the encoding step and merge the extracted representations before the decoding phase.

Another work is that conducted by Yildirim *et al.* [201]. Here, StyleGAN [170] is applied to the task of image generation. A constant vector is taken as content input and a combination of pose and garments information is taken as the style input.

Huang *et al.* [202] introduced an image generation method using an end-to-end system that consists of an appearance encoder, that learns the appearance representation of the person within the image, and an Appearance-aware Pose Stylizer (APS), where the image is progressively created from a small to a large scale, to improve the final image quality by adding as much detail as possible.

TABLE 15. Performance of state-of-the-art methods for clothes synthesis using and fashion-GEN dataset. The best performance are highlighted in bold.

Method	IS	R-precision (%)
StackGAN++ [171]	5.46 ± 0.13	17.5
AttnGAN [172]	7.94 ± 0.13	68.28
e-AttnGAN [173]	8.97 ± 0.15	72.00

3) PERFORMANCE OF THE NETWORKS

The performance of the networks that have been discussed so far are shown below. The datasets used for the purpose of the Clothes Synthesis were Fashion-GEN, Deep Fashion, VITON and Zap-Seq & DeepFashion-Seq. collected from UT-Zap50K and DeepFashion through crowdsourcing via Amazon Mechanical Turk (AMT) [203].

- **Fashion-GEN dataset**
Table 15 proposes Performance of state-of-the-art methods for Clothes Synthesis using and Fashion-GEN Dataset. The evaluation is computed in terms of Inception Score metric (IS) and R-precision.
- **Deep Fashion dataset**
Table 16 exhibits the Performance of state-of-the-art methods for Clothes Synthesis using and DeepFashion Dataset. The evaluation is computed in terms of Inception Score metric (IS), R-precision and classification accuracy.
- **VITON dataset**
Table 17 presents the Performance of state-of-the-art methods for Clothes Synthesis using VITON Dataset.

TABLE 16. Performance of state-of-the-art methods for clothes synthesis using and DeepFashion dataset. The best performance are highlighted in bold.

Method	IS	R-precision (%)	Avg. Acc. (%)	Clss.
StackGAN++ [171]	1.74 ± 0.02	12.3	37.08	
AttnGAN [172]	4.12 ± 0.06	70.73	56.18	
e-AttnGAN [173]	4.77 ± 0.10	76.21	58.39	

TABLE 17. Performance of state-of-the-art methods for clothes synthesis using VITON dataset. The best performance are highlighted in bold.

Method	SSIM	IS
CP-VTON	0.689	2.605
Poly-GAN Stage 2	0.717	2.819
Poly-GAN Stage 3	0.737	2.655
Poly-GAN Stage 4	0.725	2.790
VITON-GT	0.886	2.760

TABLE 18. Performance of state-of-the-art methods for clothes synthesis using Zap-50k-seq (a) and DeepFashion-Seq (b) dataset. The results with the best performance are highlighted in bold.

(a)		
Method	SSIM	IS
StackGAN	0.437	7.88
AttnGAN	0.527	9.79
TAGAN	0.512	9.83
SeqAttnGAN	0.651	9.58

(b)		
Method	SSIM	IS
StackGAN	0.316	6.24
AttnGAN	0.405	8.28
TAGAN	0.428	8.26
SeqAttnGAN	0.498	8.41

The evaluation is computed in terms of Structural Similarity Index metric (SSIM) and Inception Score metric (IS).

- ZAP-Seq AND DeepFashion-Seq datasets
- Table 18 shows the Performance of state-of-the-art methods for Clothes Synthesis using Zap-50k-seq (a) and DeepFashion-Seq (b) Dataset. The evaluation is computed in terms of Structural Similarity Index metric (SSIM) and Inception Score metric (IS).

V. DEEP LEARNING FOR SOCIAL MEDIA ANALYSIS

Extracting fashion knowledge from general users within social networks, such as Instagram, is a very important source of data as the images published on social media generally have ideas for different occasions or person identity information. However this is a very difficult and competitive task to perform. In fact, many images that are present throughout the social networks present descents that are difficult to understand, and it is not easy in these contexts to go and extract fashion concepts, especially for the fact that the datasets used for the tasks described above, they were based on images that contained a white background. Furthermore, the images present in social networks do not have a label already inserted, or at least not sufficient, and these data are fundamental for the construction of fashion knowledge. A possible solution could be to manually label the dataset; however, this is a solution that requires a large amount of time and carries a relatively high cost.

The datasets found in the literature mainly contain images that come from online shopping sites and that only have specific attributes relating to that specific item, and of course this data is not enough to be able to identify the type of occasion or the identity of people within an image.

A. AUTOMATIC FASHION KNOWLEDGE EXTRACTION

In recent years, Automatic Fashion Knowledge Extraction has been the focus of many studies in the field of computer vision, Machine Learning and Deep Learning. Some of these are listed in the following list.

- YAGO (Yet Another Great Ontology), is a knowledge base developed in Saarbrücken, Germany, by researchers from the Max Planck Institute for Computer Science in 2006 [204] and has undergone continuous improvements and extensions over the years, which continue to this day. The German experience in question combines broad coverage with high accuracy [205] and is therefore significant in the context of the semantic Web. YAGO represents all facts in the form of unary relations or binary: entity classes and pairs of entities linked by specifications relations. The data model can be seen as a graph in which entities and classes are the nodes and in which the relations are oriented arcs. The knowledge is organized in subject-property-object RDF format [206]: two adjacent nodes and the arc that connects them make up a triple [204]. The core of YAGO is based on Wikipedia, one of the most complete digital encyclopedias available [205];
- Freebase, [207], is a system in which all knowledge in the world is made public. In terms of design it can be compared to Wikipedia and the Semantic Web. Inside there are about 125 million tuples, about 4 000 categories and about 7 000 characteristics;
- Wikidata [208] is also a public knowledge gathering system that can be accessed and modified by any person, or computer, who enters it. It provides centralized access to structured data management to Wikimedia projects including Wikipedia, Wikivoyage, Wikisource and others.
- DBpedia [209] is a knowledge extraction system based on wikipedia. This process is done through Semantic Web and Linked Data technologies. Data extraction is done in 11 different languages and is made up of 400 million facts describing 3.7 millions of events

The main problem of all the dataset described above is that they were curated by only textual resource, ignoring the rich information that is in the visual data.

Subsequently, research studies have focused on extracting knowledge from visual rather than textual data such as

- NEIL [210] is a program that runs every day for 24 hours that was created to extract internet knowledge at the level of visual data, such as images. This mechanism is based on semi-supervised algorithms that are able to find relationships between the images and the texts written under each of them.

- Visual Genome [211], a dataset that is able to modeling the relationships between objects present within an image. They gathered dense annotations of objects, attributes, and relationships inside each image to learn these models. Especially, this dataset includes over 100K images where each image has a mean of 21 objects, 18 attributes, and 18 pairwise relations between objects. The objects, attributes, relationships, and noun phrases are canonicalized in region representation and questions answer pairs to WordNet synsets. Mutually, the notes denote the most dense and highest dataset of image representation, objects, attributes, relationships, and question answers.
- Video Visual Relation Detection (VidVRD) [212], perform visual relation detection in videos instead of still images (ImgVRD). It consists of tracklet proposal, short-term relation prediction and greedy relational association. Moreover, it is a dataset for VidVRD evaluation, which contains 1 000 videos with manually labeled visual relations.

Although most researches had the aim to derive knowledge considering the visual data and the sentences related to the visual data both textual and visual data, but there are not many works that do these types of extraction in the fashion world.

Following the works done in [52], the problem of extract fashion knowledge directly from social media can be formulate in the following manner: an user fashion knowledge can be defined as a triplet of the form $\mathcal{K} = \{\mathcal{P}, \mathcal{C}, \mathcal{O}\}$ where

- \mathcal{P} is Person, with a set of attributes (age, gender, height, weight, etc.)
- \mathcal{C} is Clothing, with a set of clothing categories and attributes (dress, long, short, material, etc.)
- \mathcal{O} is Occasion (wedding, dating, etc.) and it is accompanied with its own metadata (weather, time, location etc.).

A set of post \mathcal{X} within any social network is then referred to as a triplet of objects $\mathcal{X} = \{\mathcal{V}, \mathcal{T}, \mathcal{M}\}$, where \mathcal{V} is a set of images, \mathcal{T} is a set of texts, \mathcal{M} is a set of metadata. The aim is therefore to create an automatic knowledge extraction system that is able to provide all three useful information to describe a post about the world of fashion \mathcal{K} .

In this context, Ma *et al.* [52], have built FashionKE: this is a large dataset that consists of 80 629 images and each of these has the three characteristics written above (person, clothes, occasion) clearly identifiable. The extraction process takes place through three different steps. The first step consists in the pre-training of an object detection system, with a pre-trained ResNet, which serves to identify whether a person is contained within an image and the vector of each clothing region; the second step consists in filtering all those images that do not contain people, and those that do not have a too deformed body and face; in the last step, they manually control all the images and eliminate those that cannot show any occasion. At this point they used a bidirectional Long-Short-Term Memory (Bi-LSTM) network to be able to learn the regions that contained a dress within

the image and the occasion in which it was worn. The final hidden representation for each clothing region is the union of the hidden vectors in both directions.

The most recent work of Ma *et al.* [59], proposes a novel dataset proceeding from the popular social network Instagram, called **Fashion Instagram Trends** (FIT). Specifically, they extracted millions of posts contained in the social network from all over the world. The collected data were automated and manual filtering, in a similar way to the work of [52], [213]. The first step of this work is the same as the previous one [52], i.e. the detection of body [76] and face [214]. Also the second step is almost the same: the images with partial or deformed body or face, are eliminated. Another step, unlike the previous work, is the deleting images that contained people who did not belong to the account where the images were downloaded. At the end, a total of 680 000 images were obtained. Moreover they proposed a novel knowledge system based on LSTM model, called KERN, which also manages to take into account the time series in fashion trends.

In 2021, Parekh *et al.* [215], extracted attributes from images on an Indian e-commerce site. They also proposed a framework that uses attention mechanisms to carry out multi-task training, and also tries to balance the datasets as well.

The table 19 considers the algorithms of automatic fashion knowledge extraction for the methods described in this section.

TABLE 19. List of fashion categories in relation to the algorithms for automatic fashion knowledge extraction (AFKE).

Method	Algorithm
Ma et al [52]	ResNet+Bi-LSTM
Ma et al. [59]	ResNet+KERN

B. CLOTHES RECOMMENDATION

Clothes Recommendation cannot be general, since the preferences of user are naturally subjective. In fact, they depend on the age, occupation, culture, place of living, and so on. From this perspective, the personalisation is fundamental since it guarantees that the Clothes Recommendation agrees with the personal taste of users and includes their likes and dislikes from several perspectives. in the last few years, personalised Clothes Recommendation has received a great attention [216]–[219]. However, the problem is that, many of these methods are not able to deal with the cold-start [220] for new users. There is a few number of papers that focus on this issue.

In the work developed by Bracher *et al.* [221], the novelty is on a latent space representation, known as FashionDNA, created for fashion items.

In [222], the authors combined the categories and styles with the clothes using visual representations. These two approaches wanted to overcome the problem of cold starting when new elements were inserted for the fashion recommendation task, but they completely neglected the characteristics of the users. In the works carried out by Piazza *et al.* [222]

and Sun *et al.* [223], there is the aim of overcoming the cold-start problem by using the information provided by users.

The approach set by Verma *et al.* [224] to overcome this problem is instead practical: to understand what the preferences of a new user are, they used a limited set of images of preferences by exploiting forecasting in the fashion world.

Within each outfit it is possible to extract different types of concepts, both in terms of style and product design [225] which are of a high-level. Some works [226] also show that even if they are considered low-level concepts, then they can always be transformed into high-level concepts. For this purpose, several approaches based on computer-vision have been investigated for representing clothing items from images.

Algorithms based on deep neural networks have reached new heights in terms of performance, also thanks to the possibility of having rich fashion databases such as DeepFashion [20] or ModaNet [38].

Liu *et al.* [20] used a branched CNN architecture, Fashion-Net. This network, by being able to predict both attributes and landmarks simultaneously, can learn the characteristics of clothing.

Ma *et al.* [52] used learning system that integrates Bi-LSTMs with a CNN backbone, contextualizing it in fashion domain.

StyleNet was introduced by Yan *et al.* [227]. This is a network that creates clothing representation using a multi-task representation learning, and it incorporates different fashion concepts.

In recent years, multi-task learning provided significant results in vary applications. For this reason, Verma *et al.* [224] decided to exploit the capabilities of multi-task learning in the prediction of fashion concepts, modeling the dependencies between clothing categories and attributes. In their work, They also proposed a dataset consisting of 2 893 images in a high quality resolution sourced from Instagram and Pinterest.

Many other datasets have been built for the outfit recommendation task from Social Networks. Zheng *et al.* [228] created a dataset with images that comes from Lookbook.nu, a social media where users freely post their photos representing their own outfit. In total they selected 2 293 profiles, that had no more than 7 000 follower, and for each of these they took the 100 most recent photos, also downloading the photo caption and the corresponding hashtag.

Verma *et al.* [229] constructed occasion-oriented fashion knowledge dataset that consists of images downloaded from Instagram and Pinterest, and manually annotated as described in [224].

Also Lin *et al.* [230] formulated the fashion outfit recommendation problem as a Multiple-Instance Learning (MIL) problem, developing a new network called OutfitNet. The process of this network is divided in two phase. The first step is learning about the compatibility between fashion items and is done through a Fashion Item Relevancy network (FIR) and furthermore generates a relevance embedding of fashion

items; the second phase is learning user preferences through a Outfit Preference Network based on visual information.

Jo *et al.* [231] proposed a system that recommends fashion designs that match target scenarios or natural landscapes. When a user inputs a query that describes the target scenery, a set of candidate images related to the query is collected in a keyword-based matching manner. Then, the collected image set is used to automatically generate new fashion images using a cross-domain generative adversarial network.

The work of Jo *et al.* [232] focuses on developing intelligent modern methods for Sketch-Product and personalized voting, consisting of a Sketch-Product mode retrieval model that overcomes the limitations of a text-based search approach. The sketch-product mode retrieval model works as follows: A user sketches a fashion product he or she wants, which is extrapolated to the level of a product image using a GAN; the GAN derives the attributes of the extrapolated product image as vector values and examines the vector space for comparable images. The vector-based user preference model Clothes Recommendation works as follows: A profile obtained by professional filtering based on a DNN is pre-trained and set as the base weighting value of the recommendation model, and customized fashion trends are recommended as the weighting values for the individual are learned over time based on the preferred fashion profiling.

Tangseng and Okatani [233] have proposed a system that is not only able to assess and quantify the goodness/badness of an outfit, but also provide a rationale for the prediction. This system receives images of several items that make up an outfit as inputs and then calculates a score that quantifies the goodness/badness of the outfit. For this purpose, each image is represented as a combination of characteristics that are easily interpreted by the user: in this way the score can be better understood by identifying the most influential characteristics. Given an outfit as a set of items, the system extracts the edge image and the main colors of each item. The edge image is forward propagated through a pre-trained CNN, then the output and main colors are forward propagated through a series of concatenation and fully linked layers using ReLU to obtain the score. The system also calculates the gradient of the score with respect to the representation of each element by backpropagation. The gradients are multiplied by the corresponding features to obtain the Item Feature Influence Value (IFIV).

Li *et al.* [234] presented a Clothes Recommendation approach that models the compatibility among diverse fashion products. The method is based on a category-aware metric learning framework that embeds the fashion articles so that the cross-category compatibility notions among the items are learned while maintaining the intra-category variety among them. Their system mainly consists of three parts: a pre-trained CNN for visual feature extraction; a category complementary relation embedding space for modeling category-aware compatibility; a multiple relation-specific projection spaces for preserving the intra-class diversity.

TABLE 20. List of fashion categories in relation to the algorithms for clothes recommendation.

Method	Algorithm
Liu et al. [20]	CNN
Ma et al. [52]	CNN+Bi-LSTMs
Lin et al. [230]	OutfiNet
Jo et al. [231]	GAN
Jo et al. [232]	GAN+DNN
Tanseng et al. [233]	CNN
Li et al. [234]	CNN
Li et al. [235]	HFGN
Liu et al. [236]	neural graph filtering
Yu et al. [237]	DNN
Zhang et al. [239]	A3-FKG

Li *et al.* [235] use the hierarchical graph network to describe the relationships between the users, the proposed outfits and the items within the images: the new framework Hierarchical Fashion Graph Network (HFGN). They assigned a ID embedding to each user/outfit and they used the visual features to represent each item. To update the outfit representation and refine the user's representations, they used the information sharing rule that can also manage to trace the historical outfit. Moreover, a joint learning scheme was proposed to perform compatibility matching and outfit recommendation simultaneously.

Liu *et al.* [236] proposed neural graph filtering framework that allows to enable the flexible and diverse fashion collocation. The innovation is that it can accept inputs and outputs with different lengths and it can recommend various styled fashion collocations. Lastly, it also manages very well the datasets that are unbalanced.

The work of Yu *et al.* [237] is an extension of their earlier work [238], in which they demonstrate that the aesthetic part is very important in modeling and predicting users' preferences, especially for some fashion related domains, and it is important to model the aesthetic information. Through a deep neural network, they managed to extract the aesthetic characteristics from the images and subsequently incorporated them into the recommendation system. Then, the aesthetic features were given as input to the basic tensor model: this phase was done to apprehend the temporal preferences. In addition, they investigated aesthetic features in negative sampling to obtain further benefits in recommendation tasks.

The work conducted by Zhan *et al.* [239] has the purpose of predicting user preferences through a Attentive Attribute-Aware Fashion Knowledge Graph called A3-FKG, which is used to establish a relationship between multiple outfits considering outfit-level and product-level attributes. Moreover, it was developed a mechanism composed by two attention layers which is used to understand the preferences of each user, the first with the task of capturing the user's fine-grained preferences, and the second with the task of The first attention layer consists of the user-specific relations-aware, which captures the user's fine-grained preferences with different focus on relations for learning the outfit representation; the second focuses on the target-aware.

The table 20 considers the algorithms of clothes recommendation used for different applications.

VI. DISCUSSIONS AND OPEN QUESTIONS

We close this paper by returning briefly to the questions raised at the beginning, which remain largely open.

For what tasks is fashion data used and how has the use of this data developed over time? Initially, the research in the world of fashion aimed only at the classification of images of people with certain items of clothing that could be annotated or not. We then moved on to the search for landmark points to facilitate the detection of clothes. With the advent of GAN, there has been a particular focus on Clothes Synthesis whose task is to dress a model in a certain arbitrary pose. At the same time with the continuous growth of the power of social networks, research on fashion has also focused on the extraction of knowledge within images that came from social networks, such as Instagram.

Comparing machine and deep learning methods, does the fashion data influence the choice of using one methodology rather than another? As for the most used methods in the fashion world, we can say that deep learning approaches prevail over machine learning ones. In fact, as we can also see in the tables present in the various sections, convolutional neural networks are the basis of most of the methods developed. The datasets that exist in the fashion world greatly influence the choice between deep learning and machine learning methods. In fact they are very competitive and very difficult to face. For this reason, deep learning methods are more common.

What are the future applications that need to be developed and deepened that use fashion data? Improve the way the customer can discover a product. For example, develop methods that allow you to buy online simply by taking or inserting a photo: the site should return that article, or at least an article similar to the one entered.

Do research on Clothes Recommendation. Fashion brands have the necessity to predict in a better way the preferences of the customer by collecting and investigating shopper behavior, customer profile and customer feedback. Based on this information in conjunction with deep learning techniques allows fashion retailers to provide a customized choice of clothing for customers.

In this context to have a large scale annotation of data is very important. Having a huge amount of data concerning fashion and beauty, the generation of a precise annotation to decrease the cost and maintaining the quality is a challenging problem. Hence, a major effort in the growth of cost-effective annotations approach on fashion and beauty connected to data is required to deal with the problem.

How has social media changed the marketing strategies of fashion brands? Social networks have had and continue to have a very strong influence in the fashion world, completely rewriting the rules of marketing. It is on social channels that trends in the fashion sector, consumer behavior are studied and it is always on social media that new influencers

and Brand Ambassadors come out every day. If before the TV channel and the physical store were the best way to promote a brand, today they are obsolete and social networks, e-commerce and influencers prevail. The real protagonist in this area is the consumer, who participates at 360° from the product creation phase to the launch and promotion phase. That's why it is necessary to rethink marketing strategies and adapt them to social networks. Therefore, It is necessary to create a close and deep relationship of trust with the user. To build customer loyalty and enhance the connection with the Brand, companies can use different digital strategies that build a new customer experience:

- Mobile commerce: through social networks and apps, the company can create direct contact with the customer wherever they are;
- Brand Ambassador: represents the image of the Brand. He knows the mission, vision, values and goals and uses them to involve his followers and guide them in the purchasing process;
- Big data: Big data Analysis is used to perfect Customer care, help predict trends, develop new market strategies, all with much more precise results and in minimum times;

VII. LIMITATIONS AND LESSON LEARNT

Exploiting Deep Learning for the interpretation of fashion social media data is challenging, since we deal with the management of heterogeneous data, the different scales of representation, and the purpose of data processing. However, the aforementioned challenges related to application can be categorised as follows:

- *Lack of Available Dataset:* Regardless of the topic and/or the kind of data in the training phase (given the assumption that DL models can be arranged to fit a specific task), there is a lack of available datasets in the literature to be used as benchmarks. It is well known that DL are data-driven techniques that perform better as the number of input samples increases. Attempts to solve this problem have involved the generation of synthetic datasets. Recently, generative models have proven to be effective for this task. Generative adversarial networks (GANs) are an appealing DL approach developed in 2014 by Goodfellow [162].
- *Domain Dependent Models:* Regarding its respective fashion compartment, when there is no all-in-one solution for every task, each AI-based model should be chosen according to the task one is attempting to solve. In other words, as AI improves, the need has emerged to understand how to make such models effective, choosing them according to the kind of data for which they have been designed. Integrating the knowledge of domain expert into AI models increases the reliability and the robustness of algorithms, making decisions more accurate. Moreover, the knowledge acquired for one task can be used to solve related ones thanks to transfer learning strategies.

• *Hardware Limitations:* despite the growing computational capabilities of better-performing CPUs and the advances in distributed and parallel high performance computing (HPC), the computational costs of the above-mentioned tasks remain high. We are not still at a stage where the ratio between time/gained and resources/spent is in balance, making the use of DL-based methods unhelpful at times compared with time-consuming but more affordable manual solutions.

VIII. CONCLUDING REMARKS

Valued at over 3 trillion dollars, the global fashion industry contributes to a healthy 2% of the global Gross Domestic Product (GDP). For this reason, fashion companies are increasingly trying to invest in the world of artificial intelligence to be able to satisfy the customer 100%. In particular, social media have long since changed the way of perceiving the world of fashion by the costumers: in this context social networks are fundamental communication tools, in particular Facebook and Instagram. Above all, the Instagram social network has become of fundamental importance for companies as the influencer sponsoring products is paid by companies to influence consumer preferences.

For this reason, this review aims to summarize the datasets that have been collected and the methods that have been used in deep learning in the fashion sector, and in particular in social networks. Methods and techniques for each kind of fashion task have been analysed, the main paths have been summarised, and their contributions have been highlighted. This review offers rich information and improves the understanding of the research issues related to the use of AI with social media fashion data. Furthermore, it is informative on how and if DL techniques and methods could help the development of applications in various fields. This work thus paves the way for further research in the domain. Future research directions include the improvement of the algorithms to use other comprehensive features, thereby achieving better performance.

REFERENCES

- [1] A. Adewumi, A. Taiwo, S. Misra, R. Maskeliunas, R. Damasevicius, R. Ahuja, and F. Ayeni, "A unified framework for outfit design and advice," in *Data Management, Analytics and Innovation*. Springer, 2020, pp. 31–41.
- [2] T. H. Nobile, A. Noris, N. Kalbaska, and L. Cantoni, "A review of digital fashion research: Before and beyond communication and marketing," *Int. J. Fashion Des., Technol. Educ.*, pp. 1–9, May 2021.
- [3] M. Paolanti and E. Frontoni, "Multidisciplinary pattern recognition applications: A review," *Comput. Sci. Rev.*, vol. 37, Aug. 2020, Art. no. 100276.
- [4] X. Gu, F. Gao, M. Tan, and P. Peng, "Fashion analysis and understanding with artificial intelligence," *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, Art. no. 102276.
- [5] C. Giri, S. Jain, X. Zeng, and P. Bruniaux, "A detailed review of artificial intelligence applied in the fashion and apparel industry," *IEEE Access*, vol. 7, pp. 95376–95396, 2019.
- [6] L. Q. Lomas, A. G. Elordi, A. A. Escondrillas, and D. L. De Ipina Gonzalez De Artaza, "A systematic literature review of artificial intelligence in fashion retail B2C," in *Proc. 6th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Sep. 2021, pp. 1–6.

- [7] S. Chakraborty, M. Hoque, N. Rahman Jeem, M. C. Biswas, D. Bardhan, and E. Lobaton, "Fashion recommendation systems, models and methods: A review," *Informatics*, vol. 8, no. 3, p. 49, 2021.
- [8] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3570–3577.
- [9] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with part-segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3408–3415.
- [10] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3519–3526.
- [11] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3182–3189.
- [12] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, Jan. 2013.
- [13] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson, "Fashion 10000: An enriched social image dataset for fashion and clothing," in *Proc. 5th ACM Multimedia Syst. Conf. (MMSys)*, 2014, pp. 41–46.
- [14] J. Huang, W. Xia, and S. Yan, "Deep search with attribute-aware deep network," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 731–732.
- [15] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 192–199.
- [16] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.
- [17] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1386–1394.
- [18] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1062–1070.
- [19] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3343–3351.
- [20] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
- [21] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 229–245.
- [22] K.-H. Liu, T.-Y. Chen, and C.-S. Chen, "MVC: A dataset for view-invariant clothing retrieval and attribute prediction," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 313–316.
- [23] Z. Li, Y. Li, W. Tian, Y. Pang, and Y. Liu, "Cross-scenario clothing retrieval and fine-grained style recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2912–2917.
- [24] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, "Pixel-level domain transfer," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 517–532.
- [25] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Unconstrained fashion landmark detection via hierarchical recurrent transformer networks," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 172–180.
- [26] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 932–940.
- [27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.
- [28] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6769–6778.
- [29] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng, "Multiple-human parsing in the wild," 2017, *arXiv:1705.07206*.
- [30] Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua, "Video2Shop: Exact matching clothes in videos to online shopping images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4048–4056.
- [31] N. Garcia and G. Vogiatzis, "Dress like a star: Retrieving fashion products from videos," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2293–2299.
- [32] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
- [33] M. Takagi, E. Simo-Serra, S. Iizuka, and H. Ishikawa, "What makes a style: Experimental analysis of fashion prediction," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2247–2253.
- [34] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional LSTMs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1078–1086.
- [35] N. Jetchev and U. Bergmann, "The conditional analogy GAN: Swapping fashion articles on people images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2287–2292.
- [36] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 792–800.
- [37] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 770–785.
- [38] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu, "ModaNet: A large-scale street fashion dataset with polygon annotations," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1670–1678.
- [39] L. Liao, X. He, B. Zhao, C.-W. Ngo, and T.-S. Chua, "Interpretable multimodal retrieval for fashion products," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1571–1579.
- [40] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, "Fashion-gen: The generative fashion dataset and challenge," 2018, *arXiv:1806.08317*.
- [41] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.
- [42] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5337–5345.
- [43] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang, "Fashion retrieval via graph reasoning networks on a similarity pyramid," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3066–3075.
- [44] X. Wang, B. Wu, and Y. Zhong, "Outfit compatibility prediction and diagnosis with multi-layered comparison network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 329–337.
- [45] S. Guo, W. Huang, X. Zhang, P. Srikantha, Y. Cui, Y. Li, H. Adam, M. R. Scott, and S. Belongie, "The iMaterialist fashion attribute dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019.
- [46] K.-H. Liu, F. Wang, and T.-J. Liu, "A clothing recommendation dataset for online shopping," in *Proc. IEEE Int. Conf. Consum. Electron. Taiwan (ICCE-TW)*, May 2019, pp. 1–2.
- [47] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao, "FashionAI: A hierarchical dataset for fashion understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019.
- [48] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "FashionIQ: A new dataset towards retrieving images by natural language feedback," 2019, *arXiv:1905.12794*.
- [49] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie, "Virtually trying on new clothing with arbitrary poses," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 266–274.
- [50] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 275–283.
- [51] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "FWGAN: Flow-navigated warping GAN for video virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1161–1170.
- [52] Y. Ma, X. Yang, L. Liao, Y. Cao, and T.-S. Chua, "Who, where, and what to wear?: Extracting fashion knowledge from social media," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 257–265.
- [53] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," 2020, *arXiv:2008.02693*.

- [54] K.-H. Liu, T.-J. Liu, and F. Wang, "CBL: A clothing brand logo dataset and a new method for clothing brand recognition," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 655–659.
- [55] G. Tiwari, B. Lal Bhatnagar, T. Tung, and G. Pons-Moll, "SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing," 2020, *arXiv:2007.11610*.
- [56] A. De Souza Inacio and H. S. Lopes, "EPYNET: Efficient pyramidal network for clothing segmentation," *IEEE Access*, vol. 8, pp. 187882–187892, 2020.
- [57] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie, "Fashionpedia: Ontology, segmentation, and an attribute localization dataset," 2020, *arXiv:2004.12276*.
- [58] Z. Ma, J. Dong, Z. Long, Y. Zhang, Y. He, H. Xue, and S. Ji, "Fine-grained fashion similarity learning by attribute-specific embedding network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11741–11748.
- [59] Y. Ma, Y. Ding, X. Yang, L. Liao, W. K. Wong, and T.-S. Chua, "Knowledge enhanced neural fashion trend forecasting," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 82–90.
- [60] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 507–517.
- [61] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [62] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.
- [63] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1879–1886.
- [64] B. Lao and K. Jagadeesh, "Convolutional neural networks for fashion classification and object detection," in *Proc. Comput. Vis. (CCCV)*, 2015, pp. 120–129.
- [65] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
- [66] C.-Y. Dong, Y.-Q. Shi, and R. Tao, "Convolutional neural networks for clothing image style recognition," *DEStech Trans. Comput. Sci. Eng.*, 2018.
- [67] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [68] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.
- [69] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
- [70] T. Yang, Y. Shi, and H. Huang, "CLDM: A clothing landmark detector based on mask R-CNN," in *Proc. 9th Int. Conf. Softw. Comput. Appl.*, Feb. 2020, pp. 1–5.
- [71] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*.
- [72] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
- [73] J. Martinsson and O. Mogren, "Semantic segmentation of fashion images using feature pyramid networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Oct. 2019.
- [74] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [75] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
- [76] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [77] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [78] J. Huang, X. Wu, J. Zhu, and R. He, "Real-time clothing detection with convolutional neural network," in *Recent Developments in Intelligent Computing, Communication and Devices*. Springer, 2019, pp. 233–239.
- [79] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.
- [80] V. Gabale and A. P. Subramanian, "How to extract fashion trends from social media? A robust object detector with support for unsupervised learning," 2018, *arXiv:1806.10787*.
- [81] W. Wang, W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4271–4280.
- [82] M. Chen, Y. Qin, L. Qi, and Y. Sun, "Improving fashion landmark detection by dual attention feature enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019.
- [83] Y. Li, S. Tang, Y. Ye, and J. Ma, "Spatial-aware non-local attention for fashion landmark detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 820–825.
- [84] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
- [85] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [86] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao, and L. Lin, "Layout-graph reasoning for fashion landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2937–2945.
- [87] M. Chen, H. Ying, Y. Qin, L. Qi, Z. Gan, and Y. Sun, *Adaptive Graph Reasoning Network for Fashion Landmark Detection* (Frontiers in Artificial Intelligence and Applications), vol. 325. 2020, pp. 2672–2679.
- [88] Z. Kai, J. Feng, R. Sutcliffe, W. Xiaoyu, and B. Qirong, "Multi-depth dilated network for fashion landmark detection," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 60–65.
- [89] H. J. Kim, D. H. Lee, A. Niaz, C. Y. Kim, A. A. Memon, and K. N. Choi, "Multiple-clothing detection and fashion landmark estimation using a single-stage detector," *IEEE Access*, vol. 9, pp. 11694–11704, 2021.
- [90] M. Song, H. Liu, W. Shi, and X. Li, "PCLoss: Fashion landmark estimation with position constraint loss," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 108028.
- [91] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.
- [92] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-CNN meets KNN: Quasi-parametric human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1419–1427.
- [93] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3185–3193.
- [94] J. Ye, Z. Feng, Y. Jing, and M. Song, "Finer-Net: Cascaded human parsing with hierarchical granularity," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.
- [95] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
- [96] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.
- [97] Z. Su, J. Guo, G. Zhang, X. Luo, R. Wang, and F. Zhou, "Conditional progressive network for clothing parsing," *IET Image Process.*, vol. 13, no. 4, pp. 556–565, Mar. 2018.
- [98] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
- [99] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
- [100] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He, "Trusted guidance pyramid network for human parsing," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 654–662.
- [101] S. Zhang, G.-J. Qi, X. Cao, Z. Song, and J. Zhou, "Human parsing with pyramidal gather-excite context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1016–1030, Mar. 2021.
- [102] E. Huang, Z. Su, and F. Zhou, "Tao: A trilateral awareness operation for human parsing," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

- [103] P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," 2017, *arXiv:1703.01386*.
- [104] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [105] T. Khurana, K. Mahajan, C. Arora, and A. Rai, "Exploiting texture cues for clothing parsing in fashion images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2102–2106.
- [106] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2648–2655.
- [107] W. Fan, Z. Qiyang, Y. Baolin, and X. Tao, "Parsing fashion image into mid-level semantic parts based on chain-conditional random fields," *IET Image Process.*, vol. 10, no. 6, pp. 456–463, Jun. 2016.
- [108] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
- [109] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.
- [110] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 502–517.
- [111] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2018.
- [112] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 4814–4821.
- [113] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "BraidNet: Braiding semantics and details for accurate human parsing," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 338–346.
- [114] R. Zhao, Y. Xue, J. Cai, and Z. Gao, "Parsing human image by fusing semantic and spatial features: A deep learning approach," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102306.
- [115] L. Wang, X. Qian, X. Zhang, and X. Hou, "Sketch-based image retrieval with multi-clustering re-ranking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4929–4943, Dec. 2020.
- [116] Y. Peng and J. Chi, "Unsupervised cross-media retrieval using domain adaptation with scene graph," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4368–4379, Nov. 2020.
- [117] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 401–410, Jan. 2021.
- [118] Y. Wang, X. Ou, J. Liang, and Z. Sun, "Deep semantic reconstruction hashing for similarity retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 387–400, Jan. 2021.
- [119] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2268–2274.
- [120] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Trip outfits advisor: Location-oriented clothing recommendation," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2533–2544, Nov. 2017.
- [121] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1946–1955, Aug. 2017.
- [122] Y. Song, Y. Li, B. Wu, C.-Y. Chen, X. Zhang, and H. Adam, "Learning unified embedding for apparel recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2243–2246.
- [123] Y.-G. Jiang, M. Li, X. Wang, W. Liu, and X.-S. Hua, "DeepProduct: Mobile product search with portable deep features," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2, pp. 1–18, May 2018.
- [124] S. Jiang, Y. Wu, and Y. Fu, "Deep bi-directional cross-triplet embedding for cross-domain clothing retrieval," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 52–56.
- [125] X. Gu, Y. Wong, L. Shou, P. Peng, G. Chen, and M. S. Kankanhalli, "Multi-modal and multi-domain embedding learning for fashion retrieval and analysis," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1524–1537, Jun. 2018.
- [126] E. Simo-Serra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 298–307.
- [127] W.-L. Hsiao and K. Grauman, "Learning the latent 'look': Unsupervised discovery of a style-coherent embedding from fashion images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4213–4222.
- [128] X. Ji, W. Wang, M. Zhang, and Y. Yang, "Cross-domain image retrieval with attention modeling," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1654–1662.
- [129] H. Su, P. Wang, L. Liu, H. Li, Z. Li, and Y. Zhang, "Where to look and how to describe: Fashion image retrieval with an attentional heterogeneous bilinear network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3254–3265, Aug. 2021.
- [130] V. Sharma, N. Murray, D. Larlus, M. S. Sarfraz, R. Stiefelhagen, and G. Csurka, "Unsupervised meta-domain adaptation for fashion retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1348–1357.
- [131] A. D'Innocente, N. Garg, Y. Zhang, L. Bazzani, and M. Donoser, "Localized triplet loss for fine-grained fashion image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3910–3915.
- [132] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.
- [133] J. Dong, Z. Ma, X. Mao, X. Yang, Y. He, R. Hong, and S. Ji, "Fine-grained fashion similarity prediction by attribute-specific embedding learning," 2021, *arXiv:2104.02429*.
- [134] Z. Wang, Y. Gu, Y. Zhang, J. Zhou, and X. Gu, "Clothing retrieval with visual attention model," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
- [135] Y. Xiong, N. Liu, Z. Xu, and Y. Zhang, "A parameter partial-sharing CNN architecture for cross-domain clothing retrieval," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.
- [136] S. Jiang, Y. Wu, and Y. Fu, "Deep bidirectional cross-triplet embedding for online clothing shopping," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–22, Jan. 2018.
- [137] S. Verma, S. Anand, C. Arora, and A. Rai, "Diversity in fashion recommendation using semantic parsing," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 500–504.
- [138] B. Zhao, J. Feng, X. Wu, and S. Yan, "Memory-augmented attribute manipulation networks for interactive fashion search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1520–1528.
- [139] J. Lasserre, K. Rasch, and R. Vollgraf, "Studio2Shop: From studio photo shoots to fashion articles," 2018, *arXiv:1807.00556*.
- [140] H. Xuan, R. Souvenir, and R. Pless, "Deep randomized ensembles for metric learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 723–734.
- [141] Y. Zhao, Z. Jin, G.-J. Qi, H. Lu, and X.-S. Hua, "An adversarial approach to hard triplet generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 501–517.
- [142] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "BIER—Boosting independent embeddings robustly," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5189–5198.
- [143] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 814–823.
- [144] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Deep metric learning with BIER: Boosting independent embeddings robustly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 276–290, Feb. 2020.
- [145] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon, "Attention-based ensemble for deep metric learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 736–751.
- [146] C. F. Higham, R. Murray-Smith, M. J. Padgett, and M. P. Edgar, "Deep learning for real-time single-pixel video," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, 2018.
- [147] A. Iliukovich-Strakovskaya, A. Dral, and E. Dral, "Using pre-trained models for fine-grained image classification in fashion field," in *Proc. 1st Int. Workshop Fashion KDD*, 2016, pp. 31–40.
- [148] S. G. Eshwar, A. V. Rishikesh, N. A. Charan, and V. Umadevi, "Apparel classification using convolutional neural networks," in *Proc. Int. Conf. ICT Bus. Ind. Government (ICTBIG)*, 2016, pp. 1–5.
- [149] K. Hara, V. Jagadeesh, and R. Piramuthu, "Fashion apparel detection: The role of deep convolutional neural network and pose-dependent priors," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

- [150] X. Wang and T. Zhang, "Clothes search in consumer photos via color matching and attribute learning," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1353–1356.
- [151] R. Patki and S. Suresha, "Apparel classification using CNNs," *Tech. Rep.*, 2016.
- [152] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.
- [153] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 818–833.
- [154] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [155] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [156] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [157] Z. Kuang, X. Zhang, J. Yu, Z. Li, and J. Fan, "Deep embedding of concept ontology for hierarchical fashion recognition," *Neurocomputing*, vol. 425, pp. 191–206, Feb. 2021.
- [158] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.
- [159] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
- [160] K. Greeshma and K. Sreekumar, "Fashion-MNIST classification based on HOG feature descriptor using SVM," *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, no. 5, pp. 960–962, 2019.
- [161] M. Kayed, A. Anter, and H. Mohamed, "Classification of garments from fashion MNIST dataset using CNN LeNet-5 architecture," in *Proc. Int. Conf. Innov. Trends Commun. Comput. Eng. (ITCE)*, Feb. 2020, pp. 238–243.
- [162] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.
- [163] A. Dosovitskiy, J. T. Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1538–1546.
- [164] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
- [165] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 597–613.
- [166] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," 2016, *arXiv:1611.06355*.
- [167] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 318–335.
- [168] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*.
- [169] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be your own prada: Fashion synthesis with structural coherence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1680–1688.
- [170] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
- [171] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2018.
- [172] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
- [173] K. E. Ak, J. H. Lim, J. Y. Tham, and A. Kassim, "Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019.
- [174] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," 2017, *arXiv:1709.07871*.
- [175] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
- [176] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2172–2180.
- [177] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 406–416.
- [178] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.
- [179] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 589–604.
- [180] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, "Towards multi-pose guided virtual try-on network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9026–9035.
- [181] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6148–6157.
- [182] N. Pandey and A. Savakis, "Poly-GAN: Multi-conditioned GAN for fashion synthesis," *Neurocomputing*, vol. 414, pp. 356–364, Nov. 2020.
- [183] S. Jiang, J. Li, and Y. Fu, "Deep learning for fashion style generation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 26, 2021, doi: 10.1109/TNNLS.2021.3057892.
- [184] L. Liu, H. Zhang, Y. Ji, and Q. M. J. Wu, "Toward AI fashion design: An attribute-GAN model for clothing match," *Neurocomputing*, vol. 341, pp. 156–167, May 2019.
- [185] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2E-try on net: Fashion from model to everyone," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 293–301.
- [186] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, and W.-H. Cheng, "Fit-me: Image-based virtual try-on with arbitrary poses," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4694–4698.
- [187] X. Han, W. Huang, X. Hu, and M. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10471–10480.
- [188] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.
- [189] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8340–8348.
- [190] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," 2018, *arXiv:1810.11610*.
- [191] P. Esser and E. Sutter, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8857–8866.
- [192] C. Lassner, G. Pons-Moll, and P. V. Gehler, "A generative model of people in clothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 853–862.
- [193] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.
- [194] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.
- [195] A. Pumarola, A. Agudo, A. Sanfelix, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8620–8628.
- [196] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," 2016, *arXiv:1610.02454*.
- [197] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [198] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

- [199] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2347–2356.
- [200] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
- [201] G. Yildirim, N. Jetchev, R. Vollgraf, and U. Bergmann, "Generating high-resolution fashion model images wearing custom outfits," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019.
- [202] S. Huang, H. Xiong, Z.-Q. Cheng, Q. Wang, X. Zhou, B. Wen, J. Huan, and D. Dou, "Generating person images with appearance-aware pose stylizer," 2020, *arXiv:2007.09077*.
- [203] M. Buhmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?" Tech. Rep., 2016.
- [204] M. Fabian, K. Gjergji, and W. E. I. K. U. M. Gerhard, "YAGO: A core of semantic knowledge unifying wordnet and Wikipedia," in *Proc. 16th Int. World Wide Web Conf. (WWW)*, 2007, pp. 697–706.
- [205] F. M. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A large ontology from Wikipedia and WordNet," *J. Web Semantics*, vol. 6, no. 3, pp. 203–217, 2008.
- [206] Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum, "Timely YAGO: Harvesting, querying, and visualizing temporal knowledge from Wikipedia," in *Proc. 13th Int. Conf. Extending Database Technol. (EDBT)*, 2010, pp. 697–700.
- [207] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.
- [208] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.
- [209] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.
- [210] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: Extracting visual knowledge from web data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1409–1416.
- [211] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, and M. S. Bernstein, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
- [212] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, "Video visual relation detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1300–1308.
- [213] Y. Ma, L. Liao, and T.-S. Chua, "Automatic fashion knowledge extraction from social media," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2223–2224.
- [214] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
- [215] V. Parekh, K. Shaik, S. Biswas, and M. Chelliah, "Fine-grained visual attribute extraction from fashion wear," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3973–3977.
- [216] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 765–774.
- [217] Y. Hu, X. Yi, and L. S. Davis, "Collaborative fashion recommendation: A functional tensor factorization approach," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 129–138.
- [218] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, and B. Zhao, "POG: Personalized outfit generation for fashion recommendation at Alibaba iFashion," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2662–2670.
- [219] R. Yin, K. Li, J. Lu, and G. Zhang, "Enhancing fashion recommendation with visual compatibility relationship," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 3434–3440.
- [220] B. Likà, K. Kolomvatsos, and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2065–2073, Mar. 2014.
- [221] C. Bracher, S. Heinz, and R. Vollgraf, "Fashion DNA: Merging content and sales data for recommendation and article mapping," 2016, *arXiv:1609.02489*.
- [222] A. Piazza, P. Kröckel, and F. Bodendorf, "Emotions and fashion recommendations: Evaluating the predictive power of affective information for the prediction of fashion product preferences in cold-start scenarios," in *Proc. Int. Conf. Web Intell.*, Aug. 2017, pp. 1234–1240.
- [223] M. Sun, F. Li, J. Lee, K. Zhou, G. Lebanon, and H. Zha, "Learning multiple-question decision trees for cold-start recommendation," in *Proc. 6th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2013, pp. 445–454.
- [224] D. Verma, K. Gulati, and R. R. Shah, "Addressing the cold-start problem in outfit recommendation using visual preference modelling," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2020, pp. 251–256.
- [225] R. Sorger and J. Udale, *The Fundamentals of Fashion Design*. London, U.K.: Bloomsbury Publishing, 2017.
- [226] K. Vaccaro, S. Shivakumar, Z. Ding, K. Karahalios, and R. Kumar, "The elements of fashion style," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016, pp. 777–785.
- [227] C. Yan, L. Zhou, and Y. Wan, "A multi-task learning model for better representation of clothing images," *IEEE Access*, vol. 7, pp. 34499–34507, 2019.
- [228] H. Zheng, K. Wu, J.-H. Park, W. Zhu, and J. Luo, "Personalized fashion recommendation from personal social media data: An item-to-set metric learning approach," 2020, *arXiv:2005.12439*.
- [229] D. Verma, K. Gulati, V. Goel, and R. R. Shah, "Fashionist: Personalising outfit recommendation for cold-start scenarios," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4527–4529.
- [230] Y. Lin, M. Moosaei, and H. Yang, "OutfitNet: Fashion outfit recommendation with attention-based multiple instance learning," in *Proc. Web Conf.*, Apr. 2020, pp. 77–87.
- [231] S.-Y. Jo, S.-H. Jang, H.-E. Cho, and J.-W. Jeong, "Scenery-based fashion recommendation with cross-domain generative adversarial networks," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2019, pp. 1–4.
- [232] J. Jo, S. Lee, C. Lee, D. Lee, and H. Lim, "Development of fashion product retrieval and recommendations model based on deep learning," *Electronics*, vol. 9, no. 3, p. 508, Mar. 2020.
- [233] P. Tangseng and T. Okatani, "Toward explainable fashion recommendation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2153–2162.
- [234] Y. Li, Y. Luo, and Z. Huang, "Fashion recommendation with multi-relational representation learning," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Springer, 2020, pp. 3–15.
- [235] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, "Hierarchical fashion graph network for personalized outfit recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 159–168.
- [236] X. Liu, Y. Sun, Z. Liu, and D. Lin, "Learning diverse fashion collocation by neural graph filtering," 2020, *arXiv:2003.04888*.
- [237] W. Yu, X. He, J. Pei, X. Chen, L. Xiong, J. Liu, and Z. Qin, "Visually aware recommendation with aesthetic features," *VLDB J.*, pp. 1–19, Feb. 2021.
- [238] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 649–658.
- [239] H. Zhan, J. Lin, K. E. Ak, B. Shi, L.-Y. Duan, and A. C. Kot, "A³-FKG: Attentive attribute-aware fashion knowledge graph for outfit preference prediction," *IEEE Trans. Multimedia*, early access, Feb. 16, 2021, doi: 10.1109/TMM.2021.3059514.



MARCO MAMELI received the master's degree in computer engineering from the University of Salento, in 2018. He is currently pursuing the Ph.D. degree with the Department of Information Engineering (DII), Università Politecnica delle Marche. His master's thesis entitled Mathematical Modeling of Ventilator-Patient Interaction With Focus on the Estimation of Lung Ventilator Parameters. His research interest includes automatic generation of augmented and virtual reality contents through artificial intelligence algorithms.



MARINA PAOLANTI is currently an Assistant Professor with the Department of Political Sciences, Communication and International Relations, University of Macerata, and an Adjunct Professor with the Department of Information Engineering (DII), Università Politecnica delle Marche. During her Ph.D., she worked with GfK Verein, Nuremberg, Germany, for visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks. Her research focuses on artificial intelligence and computer vision, with particular focus to specialized machine learning algorithms and deep learning architectures.



ROCCO PIETRINI received the Ph.D. degree in information engineering from Università Politecnica delle Marche, in 2020, working on deep learning and image processing for human behavior analysis. He spent part of his Ph.D. at the University of Central Florida, Orlando, USA, working on semantic segmentation for action detection. He is currently a Postdoctoral Researcher with the Department of Information Engineering (DII), Università Politecnica delle Marche, and a Research and Development Engineer with Grottini Lab Company, working worldwide in human behavior analysis in retail environment.



GIULIA PAZZAGLIA received the master's degree in mathematics and applications from the Università degli Studi di Camerino, Italy, in 2019. She is currently pursuing the Ph.D. degree with the Department of Information Engineering (DII), Università Politecnica delle Marche. Her master's thesis entitled Stochastic Gradient Method for Artificial Neural Networks. Her research interest includes shopper's behavior understanding in intelligent retail environment.



EMMANUELE FRONTONI is currently a Full Professor of computer science with the University of Macerata and the Co-Director of the VRAI Laboratory, Department of Information Engineering (DII), Università Politecnica delle Marche. He coordinated and participated in several industrial research and development projects in collaboration with ICT and mechatronics companies in the field of ambient assisted living. His research interests include computer vision and artificial intelligence with applications in robotics, video analysis, and human behavior analysis, and the automatic classification of images. He is also involved in several e-health projects in the field of data interoperability, cloud-based technologies, and big data analysis. He is a member of the European Association for Artificial Intelligence, the European AI Alliance, and the International Association for Pattern Recognition.



PRIMO ZINGARETTI (Senior Member, IEEE) is currently a Full Professor of computer science with Università Politecnica delle Marche, Italy. He has authored over 150 scientific research articles in English. His main research interests include artificial intelligence, robotics, intelligent mechatronic systems, computer vision, pattern recognition, image understanding and retrieval, information systems, and e-government. Robotics vision and geographic information systems have been the main application areas, with great attention directed to the technological transfer of research results. He is a member of ASME and GIRPR-IAPR, and a Co-Founder of AI*IA.

• • •