International Journal of Multimedia Information Retrieval (2023) 12:24
https://doi.org/10.1007/s13735-023-00294-5

REGULAR PAPER

An interactive attribute-preserving fashion recommendation with 3D image-based virtual try-on

Ahmad Alzu'bi[1] · Lojin Bani Younis[1] · Alia Madain[1]

Received: 31 March 2023 / Revised: 11 June 2023 / Accepted: 10 July 2023 / Published online: 9 August 2023

Abstract
Online shopping experiences should be simplified by incorporating essential features such as virtual try-on clothing and recommending new items based on the customer's preferences. This is necessary given the rapid growth of the fashion industry and the expansion of shopping technologies. In this paper, we propose a new approach integrating fashion image retrieval and recommendation with a 3D virtual try-on network. We aim to build an interactive attributes-preserving model that allows users to choose the favourite garments and virtually try them on after uploading a frontal image of the whole body. Several deep learning architectures are used to extract and learn the key attributes of garment image, by which each formulated image is subjected to effective human body segmentation and pose estimation procedures. Then, a 3D VTON network is used to generate a 3D image of the user wearing a specific garment, after which the fashion retrieval system recommends and ranks more relevant items. Extensive experiments on multi-domain fashion dataset demonstrate that the proposed framework outperforms the state-of-the-art methods in terms of fashion retrieval and attribute relevancy, achieving a top@30 accuracy of 80.02%, an NDCG@30 of 80.26%, and a top@30 mAP of 87.71%. Additionally, the generated generic image descriptors require very little memory space, enabling rapid online learning and retrieval of large-scale 3D images.

Keywords Fashion recommendation · Virtual try-on · Image retrieval · Deep learning · 3D imaging

1 Introduction

People are increasingly turning to online apparel shopping as a form of fashion shopping. According to several studies [1, 2], e-commerce merchants such as Amazon, eBay, and Shopstyle, as well as social networking sites such as Pinterest, Snapchat, Instagram, and Facebook, have emerged as the most popular media for fashion recommendation. According to Statista [3], the women's apparel revenue is expected to exceed $0.99tn in 2023. Also, it is expected that the revenue will experience an annual growth rate of 11.45% (Compound Annual Growth Rate (CAGR) 2022–2025), resulting in a market volume of $1.37tn by 2025. With the digitization of the fashion industry, the system evaluates customer fashion desires and seeks to meet them as rapidly as possible. As a result, the fashion industry's digital technology is catching the attention of consumers from a wide variety of demographics, who have a wider choice of options due to the reduced production cycle [4]. Furthermore, with over 10 million consumers using smart devices, these devices play a significant role as a new e-commerce platform [5].

Online shopping differs greatly from offline shopping, vast range of choices available to consumers. Thus, companies are integrating artificial intelligence (AI) solutions throughout their supply chains to enhance creativity, improve customer service quality, assist designers, and increase overall efficiency [6]. Fashion retrieval refers to the process of locating fashion items that match a user's search criteria, such as a specific color, style, or brand. It entails analyzing images of fashion items with computer vision techniques to identify patterns and features that correspond to the image query [7]. Fashion retrieval systems can be used in a variety of settings, including online fashion marketplaces, fashion blogs, and virtual try-ons. For instance, if a user searches for a particular kind of dress on an online marketplace, the

✉ Ahmad Alzu'bi
agalzubi@just.edu.jo

Lojin Bani Younis
lhbaniyounis19@cit.just.edu.jo

Alia Madain
asmadain@just.edu.jo

[1] Department of Computer Science, Jordan University of Science and Technology, Irbid 22110, Jordan

retrieval system will examine their search parameters and provide a list of garments that fit those requirements. Visual search is a popular fashion retrieval technique that involves analyzing images of fashion items to identify patterns, colors, and other features [8–10]. This technique can be used to find items that are visually similar to a user's query or to identify specific details such as the cut or fabric of a garment. In this sense, fashion retrieval can be used in fashion recommendation systems (FRS) to improve the accuracy and relevance of the recommendations provided to users. Recommendation systems can produce more individualized and pertinent recommendations by utilizing fashion retrieval techniques, such as visual and semantic search, to better understand a user's tastes and behavior.

However, the inability of customers to try on products before buying them is a significant problem for e-commerce and online retailers. As a result, numerous studies tackled this challenge and investigated a number of solutions, including virtual try-on (VTON) systems that let clients digitally try on clothing on images of their bodies. The ability to view how an item of clothing or an accessory might look on them in a more convenient situation is one of the main benefits of 3D virtual try-ons, which can increase customers' confidence in making a purchase and save time and money, making internet shopping more trustworthy.

Pose estimation pinpoints the location and orientation of an object or a person in an image or a video. It has been utilized in a wide range of vision applications, including robots, virtual reality, human-computer interaction, and surveillance [11]. Pose estimation can be also applied to virtual try-ons. This can be accomplished by determining a person's 3D pose in an image or a video, and then accurately positioning and orienting the virtual clothing on that person's body. This method may be used in virtual fitting rooms, online purchasing, and fashion design [12].

Previous works were only concerned either with VTON or fashion recommendation. To the best of our knowledge, this is one of the first works to consider combining VTON and fashion recommendation into a single pipeline with an efficient 3D image retrieval system. In this paper, the proposed fashion recommendation system is based on the use of discriminating characteristics and attributes extracted from 2D garment images and learnt by deep learning models, by which a 3D garment representation is virtually provided according to the user's preferences. However, this retrieval and ranking system requires a sufficient diverse collection of training images to facilitate the learning procedure and maintain the quality of such virtual online shopping. Thus, we used the Shopping100k fashion recommendation dataset [13, 14], in which specific categories are carefully selected and processed to maintain them compatible with the virtual try-on system. In particular, the categories of long coat, jacket, shirt, T-shirt, and jumper were used. Most importantly, the VTON component in our fashion recommendation system, which includes a body pose estimation and 3D virtual fitting, benefits from the deep models pretrained on open-domain image collections. This enabled the recommendation system to learn and preserve the characteristics of fashion images using specific-domain data, i.e., fashion recommendation, and general-domain data, i.e., virtual reality or pose estimation.

The main contribution of this work can be summarized as follows:

1. A deep learning pipeline combining VTON with fashion image retrieval is proposed to provide an interactive personalized fashion recommendation system with 3D virtual try-ons. Based on a set of 2D image attributes, it models and learns the 3D item-body pairs provided by users. In addition, the image pairs for the virtual try-on system are generated dynamically based on user selection, making it more personalized with the option to create multiple pairs for each image.
2. A modified version of the attribute-driven disentangled encoder (ADDE) algorithm is used to process the fashion images. The attributes and possible attribute values for each image are then pre-defined in the utilized dataset, allowing the fully connected deep learning networks to embed and map the image features to attribute-specific subspaces. Therefore, a transfer learning procedure using several CNN-based models is applied to improve the discrimination capability of the retrieval algorithm.
3. A set of compact features are formulated to represent the generic semantic descriptor of item-body 3D image pair, which is a crucial requirement for large-scale image retrieval tasks including fashion recommendation. The low-dimensional image vector maintains fast online training and low space on actual storage. In addition to the standard metrics, we compute the mean average precision for each fashion attribute for the first time as a part of the evaluation of the quality of the fashion attributes and the performance of the recommendation system.

The rest of this paper is organized as follows: Sect. 2 discusses the fashion recommendation and virtual try-ons and reviews the most related works that addressed these two components; our methodology to build the virtual fashion recommendation framework is illustrated in Sect. 3; Sect. 4 presents the experimental setups and the performance results of the proposed model; and Sect. 5 concludes this paper and summarizes the main findings.

2 Literature review

This section discusses relevant previous works that dealt with fashion recommendation, virtual try-ons, and the combination of both.

2.1 Fashion recommendation systems

Many works have introduced fashion matching systems to retrieve and rank fashion items. Sarkar et al. [15] introduced a framework called OutfitTransformer, which is designed to address compatibility prediction and complementary item retrieval in the context of fashion outfits. It utilizes task-specific tokens and employs the self-attention mechanism to learn outfit-level representations that capture the compatibility relations between all items in the outfit.

Song et al. [16] proposed a content-based neural scheme based on the Bayesian personalized ranking (BPR) framework to model compatibility between fashion items. They also proposed a neural compatibility modeling scheme with attentive knowledge distillation in [17]. To enhance the BPR method, He et al. [18] proposed the Adversarial Personalized Ranking (APR).

Another aspect of fashion retrieval and recommendation is the user's history. Wu et al. Quadrana et al. [19] proposed a session-based recommendation system that is based on RNNs. Smirnova and Vasil [20] modified the session-based recommendation systems by adding context information to the input and output layers of conditional RNNs. Abugabah et al. [21] adopted efficientNet-B7 to propose a fashion recommendation system that is based on user preferences and historical user interactions. It takes into account the image region-level features as well and learns better representations by estimating the weights of items.

There are some works addressed the FRS based on joint attributes, for instance, Liu et al. [22] proposed FashionNet, a branched CNN architecture. They learn the characteristics of clothing by jointly predicting attributes and landmarks. Morelli et al. [23] used CNNs with a modified triplet loss function for fashion retrieval by the integration of hard negatives.

Li et al. [24] proposed an aspect-based fashion recommendation model with an attention mechanism that predicts customer ratings on fashion products by extracting latent aspect features of users and products using two parallel paths of CNNs: LSTMs and attention mechanisms. Attention-based memorability estimation network (AMNet) [25] provided an attribute manipulation fusion module and a memory block with internal memory and a neural controller. FashionSearchNet [26] used attribute activation maps that have been trained in a weakly supervised manner to manipulate attributes of the fashion items.

Baldrati et al. [27] presented an interactive system that improves e-shop search engines by combining visual and textual features using a combiner network trained with contrastive learning. Shimizu et al. [28] introduced the concept of "fashion intelligence" and proposed a system based on visual-semantic embedding for learning and interpreting fashion. Their method enables the embedding of abstract tag information alongside outfit images in a shared projective space to allow for effective searching of outfit images using fashion-specific abstract words.

Divitiis et al. [29] proposed a MANN (memory augmented neural network) architecture that considers the co-occurrence of clothing attributes like shape and color to create outfits. They used disentangled representations of fashion items and store them in external memory modules, which are utilized during the recommendation process.

2.2 Virtual try-ons

Image-based virtual try-on (VITON) [30] is one of the best-known approaches in virtual try-on. It depends only on plain RGB images and uses the thin-plate spline (TPS) [31]-based warping method to create new images of the same person wearing new outfits that fit perfectly to the corresponding region of the human body by warping the clothes.

Characteristic-preserving image-based virtual try-on network (CP-VTON) [32] is a refined version of VITON; it learns the parameters of TPS using a neural network-based geometric matching module (GGM) instead of using image descriptors, which makes it obtain better results of image details. Looking-attractive virtual try-on (LA-VITON) [33] is derived from CP-VTON, and it also uses the GGM, and two-stage transformation strategy depending on perspective transformation and TPS transformation. However, these approaches focus on the clothes and ignore the pose and the overall body information details, leading to blurry regions in the images. To address this problem, virtual try-on network with feature preservation (VTNFP) [34] extracts the high-level features from the body parts using a self-attention mechanism and concatenates them with the bottom clothes, it uses the semantic segmentation as the main input to create a body segmentation map to the clothed person, yet their method still produces blurry parts in the images because it does not take the semantic layout in the reference image into account.

To solve this issue, adaptive content generating and preserving network (ACGPN) [35] was introduced that uses a mask generation mechanism in semantic segmentation to generate masks for body parts and warped clothes, and it also proposes the clothes warping module (CWM) which is a second-order constraint on TPS parameters.

CF-VTON [36] is a novel multi-pose virtual try-on network that overcomes the unnatural garment alignment and difficulty in preserving the person's identity that happens due to weak mapping relationships between different feature crosses. Their approach predicts an "after-try-on" semantic map to guide garment alignment and try-on synthesis. Then, an improved garment alignment network (GANet) is used to optimize the alignment and correct unnatural distortions. A coarse result is synthesized using the try-on synthesis network (TSN), and finally, the output is refined to reconstruct the virtual try-on result with rich facial identity and garment details.

Another framework that aims to accurately capture the "after-try-on" semantics is RTVTON [37], leading to improved try-on quality and adaptability across various garment types. By focusing on representing the final appearance of garments after being worn, it enhances the realism and effectiveness of virtual try-on experiences.

Zhao et al. [38] proposed a new approach to reconstruct 3D try-on meshes using only the target clothing and a person image as inputs. This approach is called M3D-VTON, which is mainly based on three modules: the monocular prediction module (MPM), the depth refinement module (DRM), and the texture fusion module (TFM). The MPM constructs an initial depth map for the human body and aligning 2D cloths to the body using a cloth warping strategy. Then, the DRM produces more details of the body and face features by refining the constructed depth map. Finally, the TFM fuses the outputs of the previous modules to obtain the best results by producing a colored point cloud and reconstructing the final textured 3D virtual try-on mesh.

Since 2D VTON is less reliable and less realistic, the previous works focused on producing 3D VTON systems. Once available, 3D body and garment models based on commercial computer graphics techniques can allow versatile and natural alterations. However, given today's technological context, creating 3D models for each individual is prohibitively expensive. Recently, parametric statistical 3D human body models such as the skinned multi-person linear model (SMPL) [39] and SMPL-X [40], as well as the unified deformation model [41], have been presented. 3D human pose and shape estimation [40, 42, 43] research for 3D human body reconstruction is also ongoing.

Using 3D SMPL models [39], Zanfir et al. [44] presented the appearance transfer between human images. Except for VTON, these are best suited for 3D character animation or real-time capturing because fully clothed reconstruction does not provide separate geometric features for the human body and garments.

Recently, experimental studies on 3D clothing model reconstruction have been ongoing. Because of the vast variety of clothing and fashion, reconstructing 3D garment models for all categories is prohibitively expensive. Multi-Garment Net [45] creates 3D garment models for 3D VTON from 3D scans of people. They use 3D garment templates for five different garment types: shirt, t-shirt, coat, short-pants, and long pants [45]. Pix2Surf [46] learns to generate 3d clothing from images for 3D VTON using Multi-Garment Net (MGN) [45] garment meshes.

ULNeF [47] is a neural model that utilizes layered neural fields to represent collision-free garment surfaces. It incorporates a neural untangling projection operator that directly works on the layered neural fields, rather than explicit surface representations.

2.3 Fashion recommendation with VTON

FashionFit [12] is an architecture that allows users to combine outfits provided by shops and visualize them on their own using neural body fit with a recommendation based on their choices. In our work, both virtual try-ons and fashion recommendation are combined, by which a 3D image-based recommendation system is mainly based on the attributes of garments extracted and learnt on a benchmarking dataset of still-images, allowing faster image search and lower memory storage.

3 Methodology

3.1 The proposed framework

In this paper, an interactive fashion recommendation pipeline is introduced that integrates fashion retrieval, attribute-driven disentangled encoder (ADDE), and monocular-to-3d virtual try-on network (M3D-VTON). The use of an effective 3D pose estimation algorithm is crucial for our recommendation system to enhance the quality of garment-body pair representation, providing better online shopping experience.

As shown in Fig. 1, the pipeline of the recommendation system includes several steps to achieve the final result as follows. A user is asked to upload an image of their whole front body; then, they choose a garment they want to try from a collection of fashion images available in the dataset including the following categories: long coats, jackets, shirt blouses, T-shirt tops, and jumpers. Then, several data preprocessing techniques are followed to prepare the images in this work including image resizing, thresholding and segmentation. Human pose estimation is a crucial step in the proposed framework to obtain the human body joints and the pose used later for the 3D fitting process. Then, the 3D virtual try-on system is used to fit the chosen garment image on the user's provided image. Then, the system recommends more garments relevant to the user choice, which are ranked according to the similarity scores reported with all the fashion images in the whole collection. Furthermore, a new evaluation metric is calculated, which is the mean average precision for every attribute present in the fashion dataset. This met-

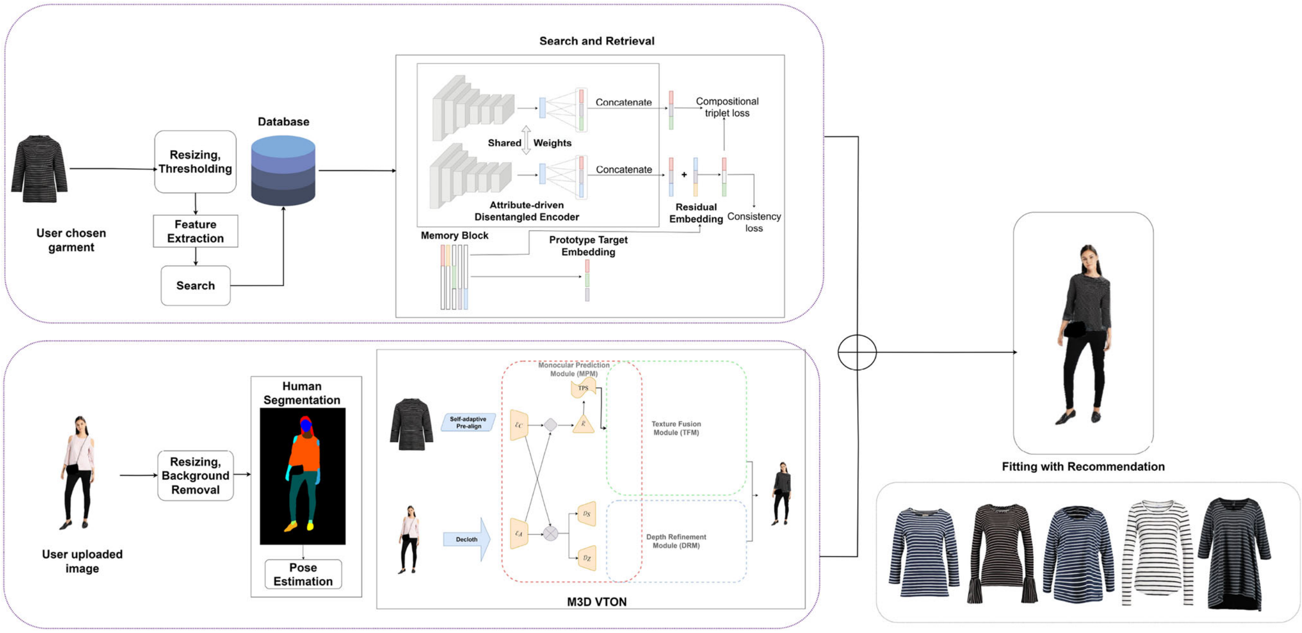

Fig. 1 The pipeline of the proposed fashion recommendation system

ric is introduced to assess the performance of the retrieval system with respect to the retrieved attributes.

3.2 Dataset preparation

This work is based on user-supplied custom input images, in which the user uploads a body image to be processed. However, for the clothing images, the Shopping100k [13, 14], a large-scale dataset of various clothing items with multiple attributes is chosen for training and evaluating the fashion recommendation system. The Shopping100k dataset includes only clothing items with a simple background collected from several e-commerce providers for fashion research, making image searches more detailed for users because the existing datasets of fashion-related research community feature a person posing, making high-resolution analysis difficult. Also, the posing model in these datasets can include a variety of outfits, which might impair product concentration and produce occlusion problems.

The Shopping100k dataset consists of 100,586 images of different kinds of clothing items, and 12 generic attributes and 151 labels are used to describe each image, with labels being better suited for attribute tweaking, thus, conducting fashion searches. Each image has at least 5-6 attributes, with varying numbers of labels. The dataset is divided into a training set of 80,586 images and a testing set of 20,000 images. Table 1 shows the statistics of attribute-label in Shopping100k.

In this work, specific categories are selected from the dataset to be compatible with the virtual try-on system since it only considers tops due to the insufficient training data available to align and fit tops and bottoms. These categories are long coat, jacket, shirt, T-shirt, and jumper. Figure 2 shows sample fashion images with attributes.

Table 1 Attribute-label statistics in Shopping100k dataset

Attributes	Sub-Categories
Category	16
Collar	17
Color	19
Fabric	14
Fastening	10
Fit	15
Gender	2
Neckline	11
Pattern	16
Pocket	7
Sleeve	9
Sport	15
12	151

3.3 Image preprocessing

To obtain the 3D fitting with outfit recommendation results, several steps and preprocessing techniques are applied as follows:

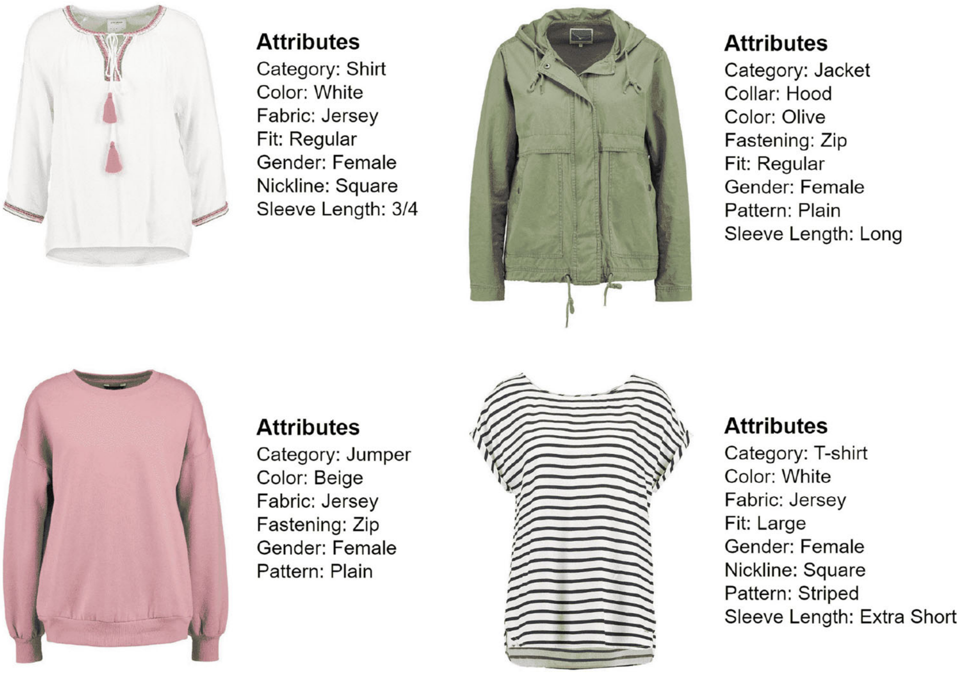

Fig. 2 Sample images of the Shopping100k dataset

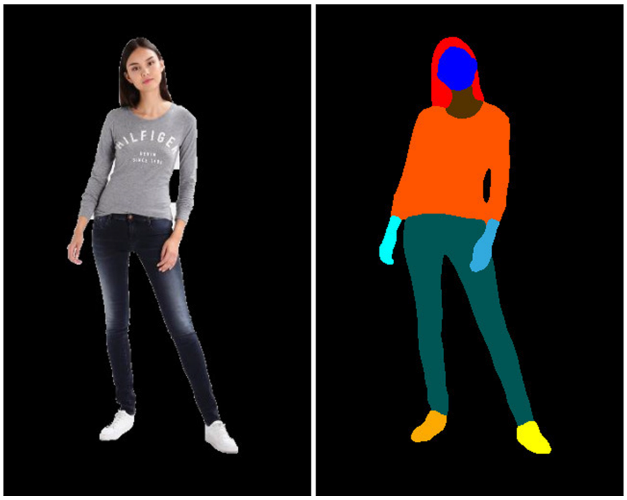

Fig. 3 A user-submitted image [38] and its corresponding segments

1. *Resizing:* To maintain the M3D-VTON compatibility, both user's and cloth's images are resized to be with a resolution of 320*512 since it was trained with a dataset of the same size.
2. *Human Body Segmentation:* The segmentation algorithm provided by Zhao [48] is applied to the submitted image, enabling the model to precisely fit outfits on the body without overlapping or mixing on existing clothes. It creates a border between each body segment to simplify mapping global similarities between garments and body. Figure 3 (left) shows a sample of the whole frontal body image uploaded by the user, and it demonstrates the corresponding segmented image (right).

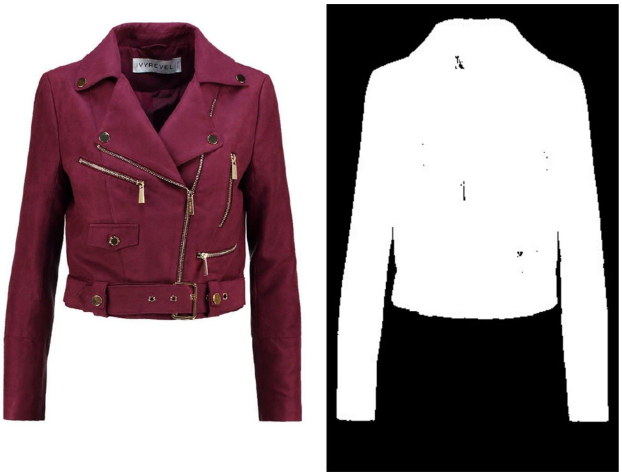

Fig. 4 A sample of the user's garment choice (left) with its corresponding image after thresholding

3. *Thresholding:* This process follows the selection of fashion image by the user. A mask of the cloth image is obtained by converting it from RGB to gray scale then to binary, i.e., black and white. The inverted binary thresholding technique is utilized to keep only the regions of interest in the image. Figure 4 shows an example of the user-chosen image (left) and the result of image thresholding applied to the garment image (right).

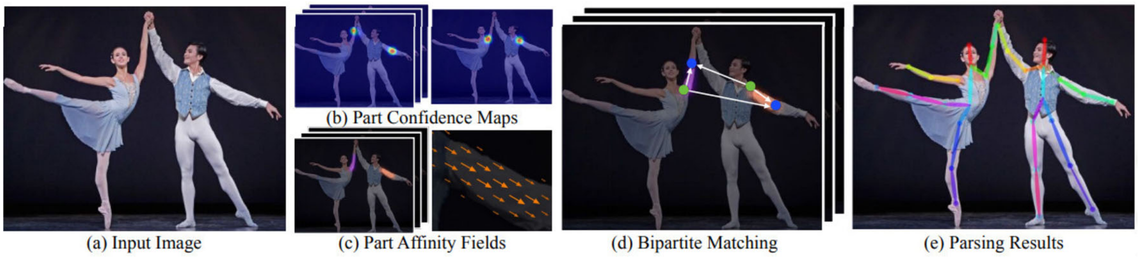

Fig. 5 The generic framework of OpenPose [49]

3.4 Pose estimation

Human pose estimation is used to recognize and classify human body joints. It essentially gathers a set of coordinates for each joint, e.g., arm, head and torso, that serves as a crucial point for defining a person's posture. A pair is a relationship between these points. Not all points can form a pair since the relationship between them must be significant. Pose estimation's first objective is to represent the human body before processing it for task-specific applications.

In this step, the multi-person OpenPose [49] is used to detect the human body, foot, hand, and facial key points on single images. It is compatible with a variety of platforms, including Ubuntu, Windows, Mac OS-X, and embedded systems. It also supports a variety of hardware, including CUDA GPUs, OpenCL GPUs, and CPU-only devices.

OpenPose is divided into three parts: body/foot detection, hand detection, and face detection. The combined body/foot key-point detector is the core part. It can use the original body-only models, which were trained on COCO [50] and MPII [51] datasets. Depending on the output of the body detector, the general positions of specific body components, such as the ears, eyes, nose, and neck, can be used to compute face bounding box proposals. Similarly, the arm key points are produced with the hand-bounding box proposals.

The OpenPose algorithm, as demonstrated in Fig. 5, is incorporated into our framework to extract 2D positions of key anatomical points for each person present in an image. This algorithm is particularly beneficial as it takes in a color image of size $w \times h$ as input and outputs the 2D positions of anatomical key points for each person in the image. A feedforward network initially predicts a collection of 2D confidence maps of body part placements, and a set of 2D vector fields of Part Affinity Fields (PAFs), which represent the degree of correlation between parts. $S = (S_1, S_2, ..., S_J)$ contains J confidence maps, one for each portion, where $S_j \in \mathbb{R}^{w \times h}$, $j \in 1...J$. $L = (L_1, L_2, ..., L_C)$ has C vector fields, one for each limb, where $L_c \in \mathbb{R}^{w \times h \times 2}$, $c \in 1...C$. In L_c, each image position encodes a 2D vector which can provide a deeper understanding of the relationships between different body parts. Finally, the algorithm uses a greedy inference process to parse the confidence maps and PAFs to determine the 2D key points for all persons in the image [49]. This is beneficial for the fashion recommendation system as it provides detailed information about the posture and body language of individuals in an image.

3.5 3D virtual try-on

Monocular-to-3D virtual try-on network (M3D-VTON) [38] is one of the successful virtual try-on approaches. The procedure transfer learning is employed in this step, as the pretrained model of the M3D-VTON is used and applied directly to the user-provided image paired with a garment image of their choice. M3D-VTON reconstructs the 3D try-on meshes using only the target clothing and a person image as inputs.

The M3D-VTON is divided into three modules: (a) monocular prediction module (MPM) to obtain the cloth-agnostic person representation A, deforming the in-shop clothing C to the warped clothing C^w through a self-adaptive pre-alignment followed by a TPS transformation, predicting a person segmentation S, and estimating an initial double-depth map D^i; (b) depth refinement module (DRM) which refines the initial depth map and provides more local details, such as garment folds and face structure, by introducing a novel depth gradient constraint, given the input of the double-depth map D^i, the warped clothes C^w, the preserved human portion I^p, and their shadow information I^g; and (c) texture fusion module (TFM) which merges the warped garments and the conserved texture information, and the results I^t are rendered under the guidance of MPM's semantic layout. Once I^t and the refined depth map D^r are spatially aligned, resulting in an RGB-D representation, colored point clouds can be directly extracted and triangulated to obtain the 3D dressed human O wearing the target garments while retaining its identity.

The purpose of using virtual try-on is to create the image pairs as input to our system. However, the image pairs in the M3D-VTON are provided automatically in the MPV3D

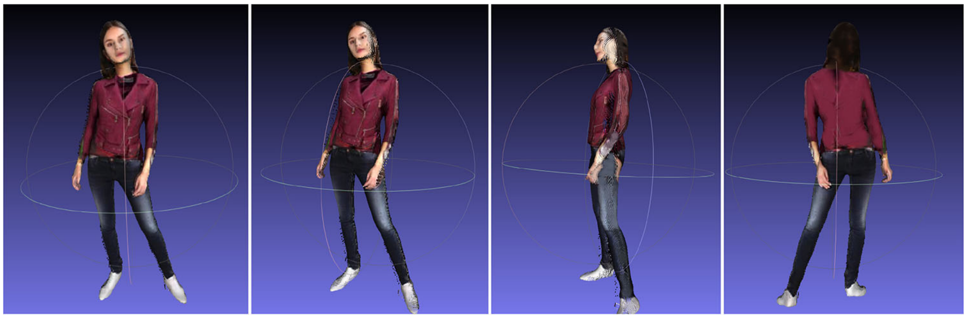

Fig. 6 A set of 3D views of a human-garment pair

dataset used to train and test the model. In contrast, our approach creates the image pairs based on the user's selection of the desired garment, making it more personalized recommendation system and allowing for the creation of multiple pairs for each image. In this step, the user-selected garment image is fitted on the user-uploaded image then represented in a 3D point cloud representation, as shown in Fig. 6. We investigate four backbone deep CNN-based models that has previously been trained and proven its efficiency in different domains. AlexNet [52], Resnet50 [53], Resnet101 [53], and MobileNet [54] are used to extract and learn image features with their weights.

3.6 Fashion retrieval and recommendation

3.6.1 The procedure of semantic retrieval

We used the implementation of retrieval model presented by [55] as it disentangles semantic components in distinct subspaces, which is necessary to support the virtual try-on system combined with to produce an interactive virtual try-on recommendation system. Therefore, visual characteristics are viewed by the image retrieval process as a supervisory signal that can be utilized to direct the learning of disentangled representations. In our work, AlexNet, ResNet and MobileNet, are used to encode the representation f_n of an image I_n, this step is discussed in detail below.

Firstly, a modified version of the attribute-driven disentangled encoder (ADDE) algorithm is used to analyze the dataset of fashion images. The attributes and possible attribute values for each image are pre-defined in the dataset.

A fully connected two-layer network maps the feature of image n to attribute-specific subspaces. This mapping is denoted as $\Phi(f_n)$, and the attribute-specific embedding is represented by r_n. A classification layer, composed of a fully connected layer with a softmax function, is used to predict the attribute values for each image, and the predicted attribute value is denoted as $\hat{y}_n$. The attribute-specific subspaces are trained using independent multi-label attribute-classification tasks, which are defined as a cross-entropy loss, as follows:

$$L_{\text{cls}} = -\sum_{n=1}^{N}\sum_{a=1}^{A}\log(p(y_{n,a}|\hat{y}n, a)) \tag{1}$$

where $y_n,_a$ is the ground-truth label of the image I_n for attribute a, $\hat{y}n,_a$ is the output of the softmax layer, N is the number of samples in the training set, and A is the number of attributes.

The disentangled representation of a given image I_n is obtained by concatenating the attribute-specific embeddings r_n, resulting in a representation with the size of $r_n \in R^{A \cdot d}$, where d is the dimension of each attribute-specific embedding. This disentangled representation is used in the fashion recommendation system to provide more personalized attribute-based recommendations using more accurate 3D image-based virtual try-on.

3.6.2 Attribute manipulation retrieval

The goal of attribute manipulation retrieval is to retrieve an image that is similar to a given query image, but with different attributes. The query image, labeled as I_q, has associated attribute values $v_q = (v_q^1, v_q^2, ..., v_q^J)$, where $J = \sum_{a=1}^{A} J_a$. One-hot encoding is used to represent the attribute values, meaning that a value of 1 is assigned to the attribute present in the image and 0 to the others. A memory block, labeled as M, is used to manipulate the attributes, it saves prototype ADDE embeddings for each attribute value. The memory block is initialized by averaging the ADDE embeddings of the images that have the same attribute values.

The initial prototype embeddings are comprised of these representations and are stored in the memory block's columns: $\mathcal{M} = \begin{pmatrix} \vec{1}^1 & \ldots & \vec{1}^{J1} & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \vec{2}^1 & \ldots & \vec{2}^{J2} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \vec{A}^1 & \ldots & \vec{A}^{JA} \end{pmatrix}$ where e_a^j indicates the prototype embedding for the j-th attribute value of the attribute a.

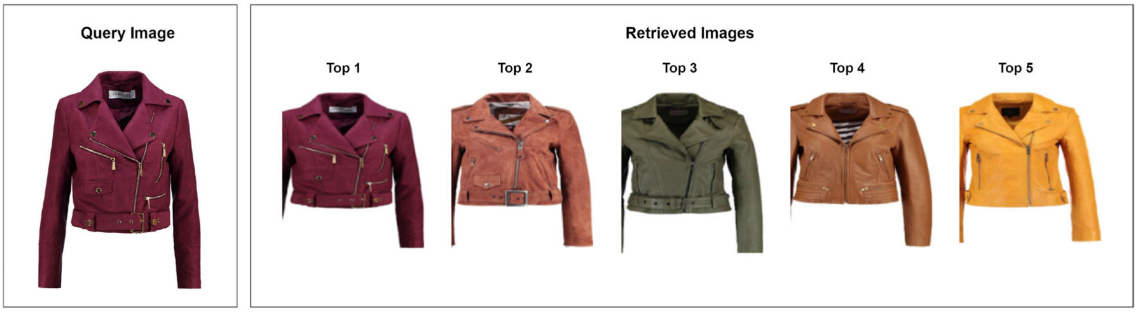

Fig. 7 A sample of the fashion retrieval result

3.6.3 Similarity search and ranking

Image similarity search aims to find images in a database that are in the same vector space as the query image; these images have less distance from the query image, which means a higher similarity. Many machine learning methods were implied for similarity search in image processing, but one of the best performing methods with higher speed and less computational cost is Facebook AI similarity search (FAISS) [56], which is used in this work. FAISS includes various methods for searching similarities. It is assumed that the instances are stored as vectors with integer identifiers and that the vectors can be compared using L2 (Euclidean) distances or dot products. Vectors that have the less L2 distance or the highest dot product with the query vector are treated similar. FAISS creates a data structure in RAM from a set of vectors x_i of dimension d. When given a new vector x in dimension d after the structure has been built, it efficiently operates:

$$j = \text{argmin}_i \|x - x_i\| \tag{2}$$

where $\|\cdot\|$ is the Euclidean distance (L^2).

The data structure is an index with an *add* method to add the x_i vectors. Then, the *search* operation is to compute the argmin on the index. In this step, after the fitting process, similar images to the chosen garment image are recommended to the user to choose one of them to try on. Figure 7 shows a sample of retrieved images that are similar to the user choice.

3.7 Performance evaluation metrics

The following standard evaluation metrics are used to measure the performance of the fashion retrieval system:

1. *Top-k Retrieval Accuracy*: defined as the number of "hit" queries divided by the total number of queries. A query is called a "hit" if there is at least one image with the desired attributes among the top-k nearest neighbors retrieved.
2. *Normalized Discounted Cumulative Gain (NDCG@k)*: defined as:

$$\frac{1}{Z}\sum j = 1^k \frac{2^{\text{rel}(j)-1}}{\log(j+1)}, \tag{3}$$

where rel(j) is the attribute relevance score for the j-th ranked image defined as the number of matching attributes between the desired label and the ground-truth label of j-th ranked image divided by the total number of attribute types; Z is a normalization constant. To better measure the ability to preserve attributes that should not be modified, two variants of the standard NDCG metric are computed: NDCG$_t$ and NDCG$_o$. Their formula is similar to Eq. 3, they only differ in the way of computing the relevance scores rel(j). NDCG$_t$ focuses particularly on the target attribute that needs to be manipulated, consequently, rel(j) will only be 0 or 1. On the other hand, NDCG$_o$ only considers the complementary attributes that should be kept fixed.

3. *Mean Average Precision @ K (mAP@k)*: defined as:

$$\text{mAP@}k = \frac{1}{N}\sum_{i=1}^{N} \text{AP@}k_i, \tag{4}$$

where N indicates the total number of queries, and AP is the average precision, which can be calculated as:

$$\begin{aligned}\text{AP@N} &= \frac{1}{m}\sum_{k=1}^{N}(P(k) \text{ if } k\text{th item was relevant})\\ &= \frac{1}{m}\sum_{k=1}^{N} P(k)\cdot \text{rel}(k),\end{aligned}$$

where N is the number of recommended items and m is the number of relevant items. rel(k) indicates whether that kth item is relevant (rel$(k) = 1$) or not (rel$(k) = 0$), and P is the Precision calculated for each pair of images. The mAP@k metric is used as a classification-based metric.

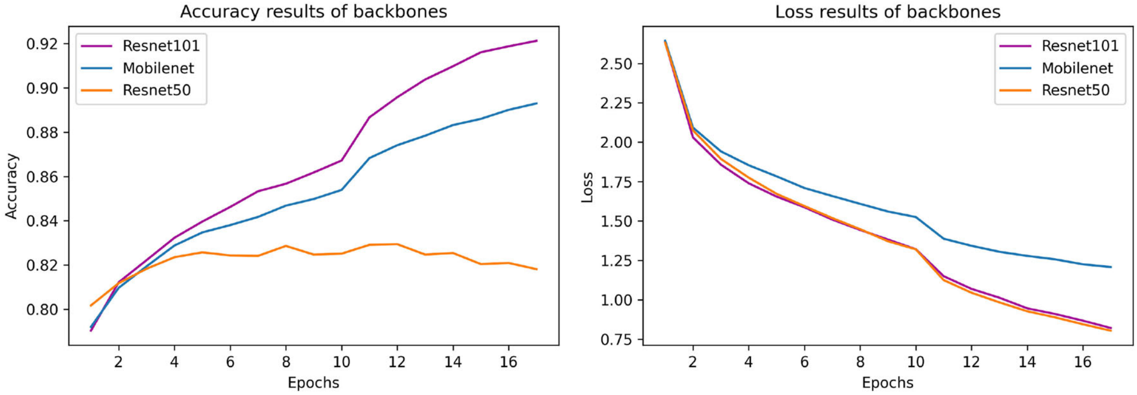

Fig. 8 The accuracy results of the testing set for the backbone deep models (left), and their corresponding loss rates (right)

Table 2 The Top@k accuracy results of the proposed recommendation system

Backbone	Baseline-AlexNet	Resnet18	Resnet50	MobileNet	Resnet101
Top@5	29.93	30.02	33.23	35.29	**50.84**
Top@10	41.17	41.42	44.71	47.11	**63.42**
Top@15	47.82	48.32	51.24	54.43	**69.68**
Top@20	52.93	53.63	56.04	59.22	**73.95**
Top@30	59.81	60.87	62.80	66.03	**80.02**
Top@40	64.10	65.62	67.62	70.66	**83.34**
Top@50	67.29	69.42	71.01	74.07	**85.74**

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

It is computed for each attribute in our fashion retrieval system.

4 Experimental results

4.1 Experiments setup

The batch size is 64 for all experiments, and the adaptive moment estimation (Adam) optimizer with a momentum of 0.9 is used. Each model was trained for 25 epochs with a learning rate of 0.0001, an lr decay rate of 0.5, and an lr decay step of 10.

4.2 The results of deep backbone models

As aforementioned, experiments were conducted over several CNN-based architectures to be tested as a backbone for the retrieval model. We used AlexNet as a baseline architecture, and the other architectures: Resnet18, Resnet50, MobileNet, and Resnet101 were investigated to obtain the best configuration of feature learning and to maximize the model's generalization ability.

Among all the experiments and compared to other architectures, as shown in Fig. 8, Resnet101 is the best model that has shown high performance in the attribute prediction task in terms of accuracy. Also, Fig. 8 (right) demonstrates the loss results of the used architectures. Notably, Resnet50 and Resnet101 are close in performance in terms of loss, whereas MobileNet has shown poor accuracy performance.

4.3 The retrieval results of attribute manipulation

For improve the performance of the recommendation system, extensive experiments were conducted on the attribute manipulation task. In this part, the performance of the deep learning architectures is measured in terms of Top@k accuracy and NDCG@k. Table 2 demonstrates the Top@k accuracy results of all backbone models. As can be observed, the best-performing model was Resnet101, achieving a Top@5 accuracy of 50.84 and a Top@30 accuracy of 80.02 compared to the baseline, with an increased accuracy of +20.21%.

Table 3 demonstrates the NDCG@k results reported for all models, in which the best model was also Resnet101, achieving an NDCG@5 of 84.08 with a decrease of +7.05% compared to the baseline and an NDCG@30 result of 80.26 with an improvement of +6.59%.

The results of NDCG$_t$@k are shown in Table 4, where it shows an improvement in the performance when Resnet101

Table 3 The NDCG@k results of the proposed recommendation system

Backbone	Baseline-AlexNet	Resnet18	Resnet50	MobileNet	Resnet101
Top@5	77.03	79.92	79.79	81.27	**84.08**
Top@10	75.84	78.70	78.67	80.03	**82.76**
Top@15	75.10	77.97	77.96	79.27	**81.89**
Top@20	74.55	77.44	77.42	78.69	**81.24**
Top@30	73.67	76.64	76.61	77.82	**80.26**
Top@40	72.95	76.02	75.99	77.14	**79.47**
Top@50	72.30	75.50	75.46	76.57	**78.81**

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

Table 4 The results of NDCG$_t$ at top-k using the target attributes

Backbone	Baseline-AlexNet	Resnet18	Resnet50	MobileNet	Resnet101
Top@5	46.85	38.44	44.06	45.24	**52.79**
Top@10	45.83	38.14	42.88	44.01	**50.14**
Top@15	44.97	37.58	41.94	42.99	**48.25**
Top@20	44.27	37.09	41.17	42.06	**46.82**
Top@30	43.05	36.15	39.81	40.63	**44.63**
Top@40	42.08	35.32	38.78	39.49	**43.01**
Top@50	41.25	34.62	37.91	38.51	**41.71**

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

Table 5 The results of NDCG$_o$ at top-k using the complementary attributes

Backbone	Baseline-AlexNet	Resnet18	Resnet50	MobileNet	Resnet101
Top@5	81.11	85.31	84.49	86.01	**88.22**
Top@10	79.89	83.96	83.37	84.76	**87.07**
Top@15	79.16	83.21	82.68	84.03	**86.33**
Top@20	78.63	82.66	82.17	83.49	**85.78**
Top@30	77.79	81.87	81.43	82.68	**84.94**
Top@40	77.09	81.27	80.85	82.06	**84.25**
Top@50	76.47	80.77	80.36	81.53	**83.67**

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

is used with an increased accuracy of NDCG$_t$@5 by + 5.94% compared to the baseline, and by + 1.58% of NDCG$_t$@30.

In Table 5, the results of NDCG$_o$@k are shown; it also shows an improvement in the performance when Resnet101 is used as it increased the results of NDCG$_o$@5 by + 7.11% compared to the baseline, and by + 7.15% in NDCG$_o$@30.

4.4 Comparison with the state-of-the-art

Table 6 shows the performance comparison of our proposed recommendation system with the related previous works in terms of Top@k accuracy. It demonstrates that our work outperforms the state-of-the-art and shows a significant performance for all garment images ranked at Top10 to Top50.

Table 7 also demonstrates that our work outperforms the state-of-the-art and shows a performance improvement in terms of NDCG, NDCG$_t$, and NDCG$_o$.

4.5 Attribute-level retrieval accuracy

We compute the mAP for the top k images which was examined by the best performing model, i.e., Resnet101. Extensive experiments were conducted to measure the mAP@k for each attribute separately to test the ability of the recommendation model in retrieving garments include the same attributes in the query image.

As shown in Table 8, the gender attribute achieved the highest mAP values with a result of 99.56, while the color attribute achieved the lowest mAP values with 63.03. This difference can be attributed to the number of labels in each

Table 6 Top@k accuracy results compared to the state-of-the-art

Model	Top@10	Top@20	Top@30	Top@40	Top@50
AMNet [25]	25.62	36.13	42.94	47.71	51.64
FSN [26]	38.41	47.44	57.17	61.62	66.70
ADDE-M [55]	41.17	52.93	59.81	64.10	67.29
Ours	**63.42**	**73.95**	**80.02**	**83.34**	**85.74**

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

Table 7 NDCG@30 compared to the state-of-the-art

Model	NDCG@30	NDCG$_t$@30	NDCG$_o$@30
AMNet [25]	71.48	40.10	75.71
ADDE-M [55]	73.67	43.05	77.79
Ours	**80.26**	**44.63**	**84.94**

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

attribute. The gender attribute has only 2 labels, while the color attribute has 19 labels, which is the highest number of labels. This means that attributes with fewer labels tend to have higher mAP values and vice versa. Figure 9 displays the results of the proposed system in terms of the top 10 retrieved images.

4.6 Analysis of retrieval speed and memory size

As part of our experimental process, we also assessed the retrieval system's performance by evaluating the speed and disk size requirements of the image vectors it utilizes. The average time taken for a single image to retrieve its similar images from 100K images then rank them on the top of the list is 3.2 milliseconds. Same thing applies for the virtual try-on fitting speed; the average time needed to perform the process of fitting for each image is nearly 3.6 milliseconds. Therefore, the system needs about 6.8 milliseconds on average to respond to the user request and to display the 3D fitting with a recommended list of similar garments that they can also select to try them on. Additionally, the average size of actual disk needed for each image vector was only about 16.3 KB, which means that the whole collection of images (100,586) needs about 1.56 GB. The significance of these findings lies in their relevance to real-time E-commerce systems and their impact on large-scale fashion systems. This is crucial because there is a growing need to develop recommendation and retrieval systems that not only save time but also reduce disk space expenses.

5 Conclusion

This paper presents a novel approach that integrates fashion retrieval, recommendation, and virtual try-ons. The objective of our work is to introduce a personalized recommendation system that empowers consumers to select appropriate garments and virtually try them on. This is achieved by uploading a frontal image of their entire body, which then

Table 8 The results of mAP@k for the attributes learnt by Resnet101

Label	Top@5	Top@10	Top@15	Top@20	Top@30	Top@40	Top@50	mAP
Category	96.73	95.91	95.34	94.90	94.22	93.67	93.18	**94.85**
Collar	98.49	98.03	97.69	97.41	96.98	96.62	96.31	**97.36**
Color	70.82	67.45	64.83	62.93	60.17	58.26	56.78	**63.03**
Fabric	77.28	74.62	72.80	71.57	69.96	68.91	68.17	**71.90**
Fastening	97.34	96.47	95.89	95.46	94.83	94.38	93.99	**95.48**
Fit	82.74	80.13	78.46	77.25	75.65	74.48	73.61	**77.47**
Gender	99.73	99.68	99.63	99.58	99.49	99.43	99.38	**99.56**
Neckline	96.70	95.68	94.95	94.38	93.55	92.90	92.38	**94.36**
Pattern	89.13	87.22	85.96	84.99	83.63	82.64	81.85	**85.06**
Pocket	98.32	97.79	97.45	97.18	96.77	96.48	96.25	**83.43**
Sleeve length	96.55	95.59	94.89	94.35	93.50	92.84	92.28	**94.29**
Sport	96.07	95.31	94.81	94.43	93.82	93.34	92.95	**94.39**
Average	**91.66**	**90.30**	**89.39**	**88.70**	**87.71**	**86.99**	**86.43**	–

The BOLD values highlight the highest results achieved by the best-performing model/algorithm in each experiment.

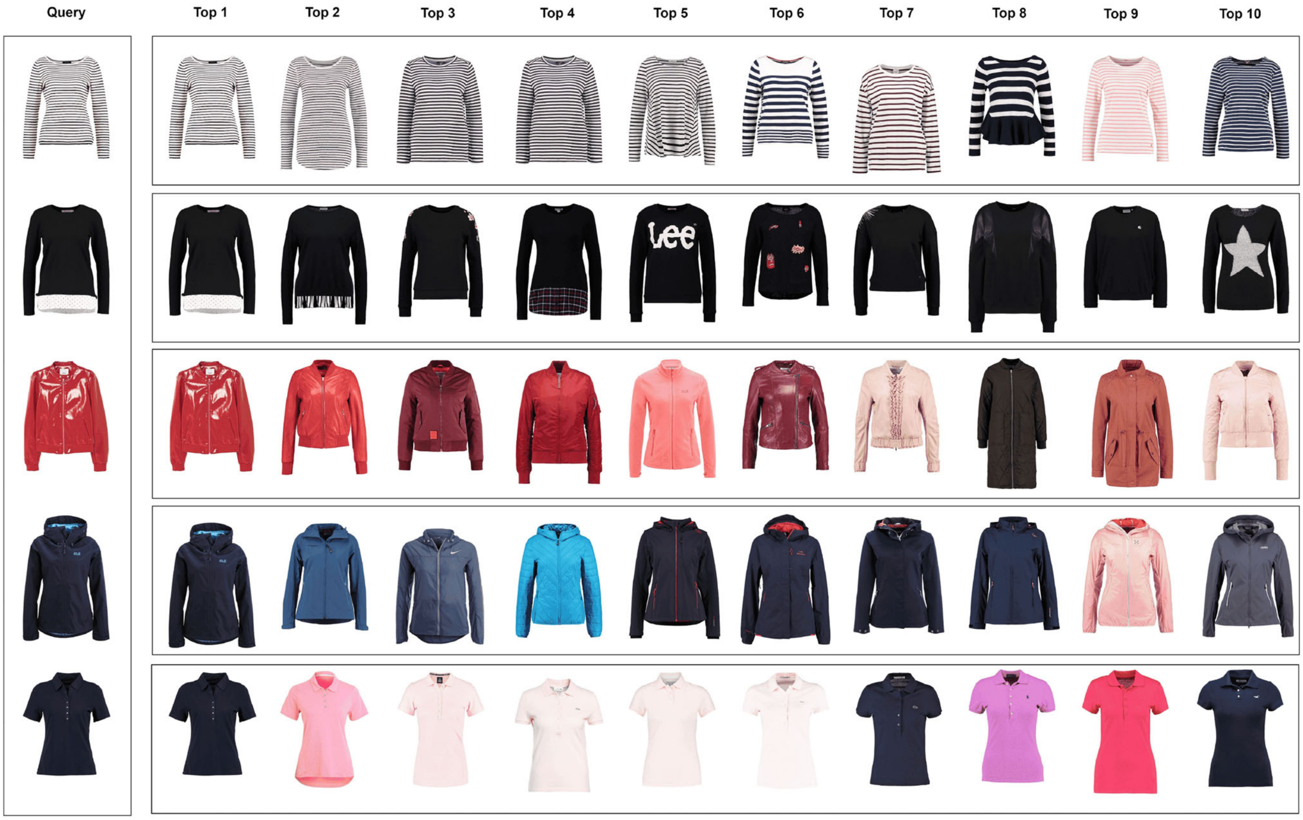

Fig. 9 Sample garments retrieved and ranked at top 10

generates a 3D fitting image. Several preprocessing techniques are applied to the user and garment images, and the user's pose is estimated by obtaining body joints with OpenPose. The M3D-VTON is used to fit the chosen garment image to the user's body in the provided image, where the ADDE is utilized for the fashion retrieval and recommendation. Extensive experiments were conducted using several CNN-based models, and the Resnet101 was used as a backbone in the fashion recommendation pipeline. One of the main findings is that the mAP@K results can be effected by the number of labels in the attributes of the garments, where this means that the less the label's amount the higher the mAP@K result, and vice versa. However, certain challenges were encountered in this research, one challenge was that only a limited number of garment categories were evaluated, such as tops, to be consistent with the used VTON system, since it performs better only with tops. As a result, the experiments focused exclusively on female fashion because currently, virtual try-on models are primarily focused on these types of garments. In future work, we aim to improve the performance of the proposed framework by considering a wider range of categories of garments, as well as utilizing more attribute manipulations in the recommendation phase to give users a variety of choices.

Author Contributions A.Z. and L.BY. conceived of the presented idea. A.Z. and L.BY. developed the theory and performed the experiments. A.Z. and A.M. verified the analytical methods. A.Z. and A.M. investigated and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript write-up.

Data availability Shopping100k dataset is available upon reasonable request from Kenan E. AK, emir.ak@u.nus.edu, and has been properly referenced in the text.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The photographs in this study were taken and processed from a publicly archived dataset with no ethics implications.

References

1. Chen W, Huang P, Xu J, Guo X, Guo C, Sun F, Li C, Pfadler A, Zhao H, Zhao B (2019) POG: personalized outfit generation for fashion recommendation at alibaba ifashion. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2662–2670

2. Park J, Ciampaglia GL, Ferrara E (2016) Style in the age of Instagram: predicting success within the fashion industry using social media. In: Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing, pp 64–73
3. Fashion—worldwide: statista market forecast. https://www.statista.com/outlook/dmo/ecommerce/fashion/worldwide#revenue
4. Jo J, Lee S, Lee C, Lee D, Lim H (2020) Development of fashion product retrieval and recommendations model based on deep learning. Electronics 9(3):508
5. Joo S, Ha J (2016) Fashion industry system and fashion leaders in the digital era. J Korean Soc Cloth Text 40(3):506–515
6. Guo Z, Wong WK, Leung S, Li M (2011) Applications of artificial intelligence in the apparel industry: a review. Text Res J 81(18):1871–1892
7. Alzu'bi A, Abuarqoub A (2020) Deep learning model with low-dimensional random projection for large-scale image search. Eng Sci Technol Int J 23(4):911–920
8. Wieczorek M, Michalowski A, Wroblewska A, Dabrowski J (2020) A strong baseline for fashion retrieval with person re-identification models. In: Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27, pp 294–301
9. Alzu'bi A, Amira A, Ramzan N, Jaber T (2015) Robust fusion of color and local descriptors for image retrieval and classification. In: 2015 international conference on systems, signals and image processing (IWSSIP), pp 253–256
10. Sachdeva H, Pandey S (2020) Interactive systems for fashion clothing recommendation. In: Mandal JK, Bhattacharya D (eds) Emerging technology in modelling and graphics. Springer, Singapore, pp 287–294
11. Wang J, Tan S, Zhen X, Xu S, Zheng F, He Z, Shao L (2021) Deep 3D human pose estimation: a review. Comput Vis Image Underst 210:103225
12. Hashmi MF, Ashish BKK, Keskar AG, Bokde ND, Geem ZW (2020) FashionFit: analysis of mapping 3D pose and neural body fit for custom virtual try-on. IEEE Access 8:91603–91615
13. Ak KE, Hwee Lim J, Kassim AA, Yew Tham J (2018) Efficient multi-attribute similarity learning towards attribute-based fashion search. In: The IEEE winter conference on applications of computer vision (WACV)
14. Ak KE, Kassim AA, Hwee Lim J, Yew Tham J (2018) Learning attribute representations with localization for flexible fashion search. In: The IEEE conference on computer vision and pattern recognition (CVPR)
15. Sarkar R, Bodla N, Vasileva M, Lin Y-L, Beniwal A, Lu A, Medioni G (2022) OutfitTransformer: outfit representations for fashion recommendation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2263–2267
16. Song X, Feng F, Liu J, Li Z, Nie L, Ma J (2017) NeuroStylist: neural compatibility modeling for clothing matching. In: Proceedings of the 25th ACM international conference on multimedia, pp 753–761
17. Song X, Feng F, Han X, Yang X, Liu W, Nie L (2018) Neural compatibility modeling with attentive knowledge distillation. In: The 41st International ACM SIGIR conference on research & development in information retrieval, pp 5–14
18. He X, He Z, Du X, Chua T-S (2018) Adversarial personalized ranking for recommendation. In: The 41st International ACM SIGIR conference on research & development in information retrieval, pp 355–364
19. Quadrana M, Karatzoglou A, Hidasi B, Cremonesi P (2017) Personalizing session-based recommendations with hierarchical recurrent neural networks. In: Proceedings of the eleventh ACM conference on recommender systems, pp 130–137
20. Smirnova E, Vasile F (2017) Contextual sequence modeling for recommendation with recurrent neural networks. In: Proceedings of the 2nd workshop on deep learning for recommender systems, pp 2–9
21. Abugabah A, Cheng X, Wang J (2020) Learning context-aware outfit recommendation. Symmetry 12(6):873
22. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1096–1104
23. Morelli D, Cornia M, Cucchiara R (2021) Fashionsearch++: improving consumer-to-shop clothes retrieval with hard negatives. In: Italian information retrieval workshop
24. Li W, Xu B (2020) Aspect-based fashion recommendation with attention mechanism. IEEE Access 8:141814–141823
25. Zhao B, Feng J, Wu X, Yan S (2017) Memory-augmented attribute manipulation networks for interactive fashion search. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1520–1528
26. Ak KE, Kassim AA, Hwee Lim J, Yew Tham J (2018) Fashionsearchnet: fashion search with attribute manipulation. In: Proceedings of the European conference on computer vision (ECCV) Workshops, pp 0–0
27. Baldrati A, Bertini M, Uricchio T, Del Bimbo A (2022) Effective conditioned and composed image retrieval combining clip-based features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 21466–21474
28. Shimizu R, Saito Y, Matsutani M, Goto M (2023) Fashion intelligence system: an outfit interpretation utilizing images and rich abstract tags. Expert Syst Appl 213:119167
29. De Divitiis L, Becattini F, Baecchi C, Del Bimbo A (2023) Disentangling features for fashion recommendation. ACM Trans Multimed Comput Commun Appl 19(1s):1–21
30. Han X, Wu Z, Wu Z, Yu R, Davis LS (2018) VITON: an image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7543–7552
31. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
32. Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M (2018) Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV), pp 589–604
33. Lee HJ, Lee R, Kang M, Cho M, Park G (2019) LA-VITON: a network for looking-attractive virtual try-on. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
34. Yu R, Wang X, Xie X (2019) VTNFP: an image-based virtual try-on network with body and clothing feature preservation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10511–10520
35. Yang H, Zhang R, Guo X, Liu W, Zuo W, Luo P (2020) Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7850–7859
36. Du C, Xiong S (2023) CF-VTON: multi-pose virtual try-on with cross-domain fusion. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1–5
37. Yang H, Yu X, Liu Z (2022) Full-range virtual try-on with recurrent tri-level transform. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3460–3469
38. Zhao F, Xie Z, Kampffmeyer M, Dong H, Han S, Zheng T, Zhang T, Liang X (2021) M3D-VTON: a monocular-to-3D virtual try-on network. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13239–13249

39. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: a skinned multi-person linear model. ACM Trans Graph (TOG) 34(6):1–16
40. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AA, Tzionas D, Black MJ (2019) Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10975–10985
41. Joo H, Simon T, Sheikh Y (2018) Total capture: a 3D deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8320–8329
42. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K (2019) Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/cvf international conference on computer vision, pp 2252–2261
43. Choutas V, Pavlakos G, Bolkart T, Tzionas D, Black MJ (2020) Monocular expressive body regression through body-driven attention. In: European conference on computer vision. Springer, pp 20–40
44. Zanfir M, Popa A-I, Zanfir A, Sminchisescu C (2018) Human appearance transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5391–5399
45. Bhatnagar BL, Tiwari G, Theobalt C, Pons-Moll G (2019) Multi-garment net: learning to dress 3D people from images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5420–5430
46. Mir A, Alldieck T, Pons-Moll G (2020) Learning to transfer texture from clothing images to 3D humans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7023–7034
47. Santesteban I, Otaduy M, Thuerey N, Casas D (2022) ULNeF: untangled layered neural fields for mix-and-match virtual try-on. Adv Neural Inf Process Syst 35:12110–12125
48. Bowen Wu FZ (2021) 2D-Human-Parsing. GitHub, San Francisco
49. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
50. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision. Springer, pp 740–755
51. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3686–3693
52. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol. 25
53. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
54. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
55. Hou Y, Vig E, Donoser M, Bazzani L (2021) Learning attribute-driven disentangled representations for interactive fashion retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 12147–12157
56. Johnson J, Douze M, Jégou H (2019) Billion-scale similarity search with GPUs. IEEE Trans Big Data 7(3):535–547

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.