Survey on clothing image retrieval with cross-domain

Chen Ning¹ · Yang Di¹ · Li Menglu¹

Received: 5 June 2021 / Accepted: 8 April 2022 / Published online: 13 May 2022
© The Author(s) 2022

Abstract
The paper summarizes the research progress on critical region recognition and deep metric learning to achieve accurate clothing image retrieval in cross-domain situations. Critical region recognition is of great value for the clothing feature extraction, effectively improving retrieval accuracy. The accuracy will decrease when solving difficult samples with similar features but different categories. Nowadays, deep metric learning is an effective way to solve this problem, which utilizes the optimization of different loss functions and ensemble network to strengthen the discrimination of clothing features. Therefore, through comparison of the experimental results of different algorithms and analysis of the accuracy of cross-domain clothing retrieval, it is demonstrated that the improvement of the retrieval accuracy in the future mainly depends on clothing important feature extraction and clothing feature discrimination.

Keywords Cross-domain clothing retrieval · Critical region recognition · Deep metric learning · Deep learning

Arabic keywords عميق تعلم · العمق المترى التعلم · الحبرة بالمنطقة الاعتراف · المجال عبر الملاص استرجاع

Introduction
The clothing image retrieval technology is that the computer recognizes the given clothing image and recommends clothing images with similar styles. Clothing image retrieval has been widely used in e-commerce platforms and search fields, such as Taobao, Jingdong, Baidu, Google, etc. benefit from clothing image retrieval. People prefer to take photos in daily life and find their favorite clothing from the Internet. Cross-domain clothing retrieval technology can help people quickly and accurately find similar styles of clothing on the Internet. This not only meets the needs of people’s daily life and improves the quality of life, but also promotes the consumption of the clothing industry. According to the fashion industry survey, the scale of domestic and foreign fashion markets is growing steadily. By 2024, the size of the domestic fashion market is expected to increase by 8.8% compared with the size of 2020, thus the market size will reach 26.288 million US dollars [1,2].

Although clothing image retrieval technology has made great progress in the past 10 years, clothing image retrieval in cross-domain situations still faces great challenges. Cross-domain clothing retrieval means that the image to be queried and the image retrieval database come from two different domains. In the scene domain, clothing images in online shopping malls are retrieved from daily street clothing images based on their similarity. Referring to recent surveys, it is found that it has two main difficulties: (1) clothes are flexible items, and when viewed from different shooting angles or the appearance can be very different when wearing different body types. The query image provided by the user may be taken under complex conditions, with complex background, various shooting angles, various lighting conditions, and even occlusion. Instead, most store images feature clean backgrounds, good lighting, and good frontal angles. (2) The intra-class variance is large and the inter-class variance is small, which is an inherent characteristic of clothing images. For example, two dresses from different categories are very similar in color and design, but have subtle differences in the shape of the neckline, one is V-shaped and the other is U-shaped. Given an image of a user’s clothing with a V-neck,
returning a dress with a U-neck is not considered a correct search result in the retrieval system.

Motivation

Clothing image retrieval in cross-domain situations has been widely used in daily life. In online shopping, when a user provides a photo of daily life, the retrieval system can return clothing images with the same or similar characteristics, which reduces the retrieval system’s dependence on text, so that the desired clothing images can be retrieved more directly and accurately. Cross-domain clothing image retrieval is also widely used in physical stores. To accurately avoid purchasing and avoiding inventory, store owners need to understand the clothing preferences of people in nearby blocks. The traditional algorithm is to manually count and classify the clothing styles of consumers around the store. However, if the program can automatically take pictures of nearby pedestrians and analyze the attributes of the clothes, the number of observers can be greatly increased, and the time and labor costs can be reduced at the same time, providing a stronger reference and basis for purchasing decisions. Therefore, the research and summary of cross-domain clothing image retrieval has far-reaching significance for both individuals and society.

In actual demand, in the major domestic e-commerce shopping platforms, clothing images are mainly retrieved through keywords or texts, and the essence is to search for pictures by text. This technique requires that clothing images be finely classified and labeled accordingly. But with the explosive growth of clothing images, the shortcomings of this method become more and more obvious. First, keywords can only describe the easy-to-extract and abstract semantic features, and cannot fully reflect the visual features of clothing images, especially some fine and difficult-to-describe features; secondly, due to the huge number of clothing images, it takes a lot of Human and material resources are used for manual labeling, and manual labeling is prone to subjective bias; finally, if the description of the search keywords entered by the user is not accurate enough, it is difficult to retrieve the desired product. Therefore, this paper studies the content-based cross-domain clothing image retrieval, summarizes and evaluates it from different technical perspectives, and hopes to bring some inspiration to researchers and find new research hotspots for future research.

Previous work

With the development of deep learning [3–6], the framework of cross-domain clothing image retrieval is shown in Fig. 1. Cross-domain clothing image retrieval mainly includes two key steps: feature extraction and similarity measurement. For clothing image features, the methods of critical region recognition is generally used to identify important areas of clothing. The new methods of similarity measurement have also appeared. At present, the best effect is use the deep metric learning methods

As shown in Table 1, different methods from past studies are summarized based on the latest research survey. In cross-domain clothing retrieval, deformation, occlusion, complex background and other phenomena will occur, which bring challenges to the accuracy of cross-domain clothing retrieval. At present, most of the critical region recognition of clothing is detect foreground objects before extracting features. The main purpose is to suppress background differences, enhance the identification of relevant local details, and provide more discriminative features in image feature extraction, which is convenient for distinguishing different types of objects. Another major challenge of cross-domain clothing image retrieval is distinguish similar images of different categories, and to cluster the images with larger differences in the same category. Deep metric learning maps images to feature vectors in space through deep neural networks. In this space, Euclidean distance or cosine distance can

Fig. 1 Framework of cross-domain clothing image retrieval
Table 1 The comparison of cross-domain clothing image retrieval methods

| Methods                  | Detailed methods          | Label information | Accuracy | Computation complexity |
|--------------------------|---------------------------|-------------------|----------|------------------------|
| Critical region recognition | Bounding box method       | Less              | Lower    | Low                    |
|                          | Human body landmarks recognition | Less  | Lower    | Low                    |
|                          | Clothing landmarks recognition | More | Higher   | Higher                 |
|                          | Attention map recognition | Less              | Higher   | Higher                 |
| Deep metric learning     | Siamese network           | No                | Low      | High                   |
|                          | Triplet network and Variants | No               | High     | Higher                 |
|                          | Ensemble network          | No                | Higher   | Highest                |

be directly used as a distance metric between two points. The contribution of many deep metric learning algorithms design a loss function that can learn more discriminative features. Therefore, there is a large amount of work researching deep metric learning and its related loss functions, including contrastive loss, triplet loss, triplet loss of more complex mutations, and the combination of multiple networks or methods. The ensemble method of output grouping together.

**Critical region recognition learning**

**Bounding box method**

The bounding box method uses some detection methods to identify the clothing regions in the image, and uses a rectangular box to mark the frame, as shown in Fig. 2. The purpose reduces the clothing image from the complex background and other external environmental factors during the retrieval process. Enhancing the effect of neural network on the feature extraction of clothing images.

Kiapour et al. [7] used a selective search method [8] to filter out any image with a width less than one-fifth of the image width, and directly used manual labeling of the bounding box of the clothing to limit the influence of the background regions and obtain a more accurate search. And other steps help reduce some of the variability observed in different online stores and item descriptions. Chen et al. [9] based on the R-CNN [10] target detection method made some improvements to the clothing detection problem in the image, using selective search to generate expected region suggestions, and using the Network-in-Network (NIN) model feature extraction was performed on the local regions. Then, Huang et al. [11] embedded additional semantic information in the tree structure layer of the attribute perception network, after they obtained the depth features of attribute perception, use Support Vector Regression (SVR) to predict the overlap ratio of each candidate frame, limit the size range and aspect ratio of the bounding box, and discard some inappropriate candidate solutions, thereby enhancing the positioning effect of the clothing regions bounding box in the image. In general, with the development of target detection [12], bounding box method is relatively easy to implement, the speed and accuracy of the recognition are also improved. However, when solving more complicated clothing images with background, human posture, and occlusion, the image features extracted in the clothing regions set by the bounding box will have many interference features, so the accuracy of retrieval accuracy is decreased.

**Human body landmark recognition method**

The landmark recognition method of the human body focuses on the limbs of the person wearing clothes. As shown in Fig. 3, according to the important body parts, it identifies the important regions of clothing, and then uses convolutional neural network to perform feature analysis on this region.

A method of research work is based on the human body pose estimation method proposed by Marcin Eichner’s team [13] to realize the detection of human body important nodes, they associated the clothing regions with the important parts of the human body and divided them into 9 parts: the torso, the left and right upper arms, the left and right lower arms, the upper left and right legs, and the lower left and right legs [14, 15]. In the recognition process, the upper body detection [16] and face detection technology [17] are combined to estimate the regions of the upper body. On this basis, the human body is segmented using the GrabCut algorithm [18], and finally combined with the human body proposed in the literature [13]. The appearance estimation model estimates the posture of the human body, thereby further dividing the important regions. The human body important node recognition method uses the landmarks of the human body to realize the detection of the important parts of the clothing. When the posture of the human body in the image is complex, it can still detect the clothing features of the relevant parts. However, when some regions are blocked, the expressive power of this method is limited.
Clothing landmark recognition method

Clothing landmark recognition directly detects the landmark marked by the clothing itself, and it is a newer method to realize the location of important regions of clothing, as shown in Fig. 4.

The DeepFashion database proposed by Liu et al. [19] defines a set of clothing landmarks, corresponding to a set of landmarks on the clothing structure. For example, a set of landmarks on the upper body is positioned as left/right collar ends and left/right cuffs. The landmarks of the lower end, lower left/right lower end, lower body clothing and full body clothing are also defined, because some landmarks often included in the image are occluded, so the visibility of each set of landmarks is also marked. On this dataset, Deep Fashion Alignment (DFA) network [20] is proposed to detect landmarks. DFA consists of three stages. In each stage, the output of the previous stage is used as input, and the network uses VGG-16 as the skeleton. In the first stage, DFA uses the original image as input to predict rough landmark locations...
and pseudo-labels. The pseudo-labels represent landmark such as clothing category and posture; In the second stage, the output of the entire network needs to predict the offset of the landmark, and the pseudo-label represents the offset of the local landmark; the third stage uses two CNNs as branches, and each branch has the same input and output. The choice of branch is determined by the pseudo-label of Phase 2. DeepFashion also has its own shortcomings. For example, each image has only one piece of clothing and each clothing category has only 4 to 8 landmark. To solve this problem, the DeepFashion2 [21] database was proposed, which has more landmark labeling and annotation information, the Match R-CNN model is proposed on this database, which is mainly composed of three parts: Feature Network (FN), Perception Network (PN) and Matching Network (MN) composition. After the query image passes through the FN network, it is input to the PN network, and then the landmark positions are obtained through the convolutional layer and the deconvolutional layer, and finally combined with the MN network for clothing retrieval. These two large clothing databases provide great support for future clothing retrieval research. The clothing landmark recognition method is also more capable of processing clothing deformation, occlusion and details, and the retrieval accuracy is greatly improved, but a lot of labeling and annotation information, and the professional knowledge of the clothing industry is very demanding.

**Attention map recognition method**

The attention map recognition method uses the idea of the attention mechanism to extract the image features of the saliency or visual attention regions from the original image, and usually needs to combine the attribute information of the clothing to complete the clothing retrieval task, as shown in Fig. 5.

Clothing attributes usually refer to the semantic attributes of objects or scenes shared across categories. The attributes of clothing include color, texture, fabric and style, so attributes can be used as potential and interpretable connections between image content and abstract labels. By constructing a latent space between fine-grained labels and low-level features, it helps models find inter-class and intra-class correlations between clothing categories. The attention map recognition method is an algorithm model that does not require a large number of human-labeled bounding boxes and feature extraction of landmark. Extract effective image representation from the spatial location of the salient regions reduce the cost of annotations while ensuring the effect of clothing image retrieval.

Most of the popular attention map recognition algorithms use image attributes as external information to locate the attention of the image in the database, and use the database image as the context to infer the attention of the query image. The attention model ignores the noise background and extracts discriminative features for retrieval. Wang et al. [22] proposed a deep convolutional neural network system, TagCtxYNet, which includes a convolutional layer for image feature extraction and an attention layer for spatial attention modeling. It extracts an effective representation of the image by learning attention weights. Gu et al. [23] proposed an autonomous learning Visual Attention Model (VAM) to extract attention maps from clothing images. It includes two branch networks, one is a global branch network based on CNN, which is used to extract attention maps from clothing images. Extract the bottom layer features of the image to get the image feature map. The other is the introduction of the attention branch of the Full Convolutional Network [24] (FCN), which is used to predict the saliency regions of the image to obtain the image attention map. The Imp-drop module is used to connect two branches to obtain the attention feature map, and the module introduces randomness between the attention map and the feature map. The addition of this randomness can reduce the risk of overfitting and allow the neural network to learn to more robust features, so the robustness of the model is improved. Zheng et al. [25] proposed an Attention-Based Region Transfer (ART) module to highlight the importance of foreground, which works in a rough way that is not classified. The attention mechanism in the advanced features is used to extract the foreground objects of interest and mark them when the feature distribution is aligned. Through multi-layer adversarial learning, the use of complex detection models can achieve effective cross-domain retrieval.

Attribute learning models usually treat attribute prediction as a multi-label classification problem, and treat each attribute as a category. In fact, the clothing images trained by each model are associated with a series of attributes, such as "silk pocket shirts". However, traditional attribute learning
models ignore sequence information. Although [22,23] use the combination of the attention feature map and the image feature map to find a more effective feature representation, which improved the clothing retrieval effect. However, there is a lack of more local information and research on the contextual connection of different parts of clothing. Luo et al. [26] proposed an attention-based learning strategy in the clothing image retrieval task. By integrating global information and local information, the features of clothing images can be intuitively extracted, because these two kinds of information provide complementary mechanisms. It can describe clothing images accurately, and uses the Long Short-Term Memory (LSTM) mechanism [27] to simulate the top-down spatial relationship of different parts of clothing to obtain more discriminative feature representations. Luo et al. [28] proposed a Deep Multi-task Cross-domain Hashing (DMCH) method to jointly establish the sequence correlation between clothing attributes, and learn the attention and perception vision of clothing images features to further enhance the effect of cross-domain clothing image retrieval.

Deep metric learning

Siamese network

Chopra et al. [29] first applied the contrastive loss function to the Siamese network based on deep neural network. Kiapour et al. [7] used Siamese network to predict whether two features represent the same category. Sean Bell et al. [30] used the traditional contrastive loss function to design an end-to-end Siamese network for modeling learning. Huang et al. [11] proposed a Dual-attribute perceptual Ranking Network (DARN) for feature learning based on the Siamese network. All in all, the contrastive loss of the Siamese network is the most widely used pair of losses in calibration learning.

As shown in Fig. 6, the Siamese network architecture has two parallel feature networks, followed by a normalization operation (L2) and a contrastive loss. Jia et al. [31] defined the contrastive loss function in the literature as shown in Eq. (1).

\[
\text{Loss}(f(a_i), f(b_i)) = (1 - y) \max [0, m - D(f(a_i), f(b_i))]^2 + yD(f(a_i), f(b_i))^2.
\]

Among \( f() \) is an embedding function that maps an image to a feature vector, \( y \) is a label, \( D() \) represents the distance between two feature vectors. The margin parameter \( m \) forces the distance between images of different categories to increase, which has a certain effect on learning ordering. On this basis, Xiong et al. [32] proposed a contrastive loss function with a bilateral distance parameter, as shown in Eq. (2).

\[
\text{Loss}(f(a_i), f(b_i)) = (1 - y) \max [0, m_1 - D(f(a_i), f(b_i))]^2 + y \max [0, D(f(a_i), f(b_i)) - m_2]^2.
\]

Here, if the positive margin (PM) parameter is equal to the negative margin (NM), it is called symmetric double margin, otherwise we call it asymmetric double margin. The existence of positive margins makes clothing images diversified to a certain extent, which is more reasonable than forcing them to be exactly the same. Wang et al. [33] optimized the contrastive loss by adding penalty constraints, and proposed a robust contrastive loss function to improve the generalization ability of the learning network.

Triplet network and variants

The triplet loss [34] function is widely used in the triplet network model, and has achieved better results in cross-domain clothing image retrieval. The structure of the triplet network is shown in Fig. 7. The main structure is three parallel feature networks that map images into feature vectors, then normalize the feature vector, and input it into the three sets of loss functions. The triplet loss makes the distance between different clothing images larger and the distance between the same clothing images smaller.

Different from the contrastive loss function that considers the absolute distance of the pair, the triplet loss calculates the relative distance between the positive pair and the negative pair of the same reference sample, and the specific definition is shown in Eq. (3).

\[
\text{Loss}(f(a_i), f(p_i), f(n_i)) = \max \left\{0, m + D(f(a_i), f(p_i))^2 - D(f(a_i), f(n_i))^2\right\}.
\]

Among \( a_i,p_i \) and \( n_i \), respectively, represent the reference sample, the positive sample, and the negative sample. The \( a_i \) and \( n_i \) label of is the same, the \( a_i \) and \( p_i \) label is different. \( m \) is the margin between the positive and negative pairs.

Due to the triplet contains reference samples, positive samples and negative samples, so \( N \) images can generate \( O(N^3) \) sample, even if it is a medium number of images, it is impossible to consider all samples, because not all samples can provide the same information for training a model. Randomly selects triples is a very inefficient training deep embedding network, which has inspired a lot of recent work to mine difficult samples for training. Wang et al. [35] randomly selected samples as triples in the first 10 rounds of training, and dig out difficult triples in each small batch after
10 rounds of training. Cui et al. [36] used manual methods to mark difficult negative images from images with high confidence scores assigned to them in each round. Simo-Serra et al. [37] analyzed the impact of difficult positive sample mining and difficult negative sample mining, and found that the combination of positive sample mining and negative sample mining improved the discrimination ability. Song et al. [38] designed a small batch of triple loss that considers all possible three-group associations in the small batch. Liu et al. [39] proposed a cluster-level triple loss, which considered the correlation between the cluster center, the positive sample and the nearest negative sample. Ge et al. [40] introduced Hierarchical Triplet Loss (HTL) to solve the random sampling problem in the triplet training process. These studies solve the problem of how to mine difficult samples in training. They use more complex triples for training, which can not only speed up the convergence speed of the learning algorithm, but also can use the positive and negative samples of the given reference learn clearer margins and better improve the global structure embedded in the cyberspace.

However, the methods based on difficult sample mining aims to find those triples that are difficult to find in the existing network from the existing training samples. It is essentially a greedy algorithm, which makes the trained feature embedding network vulnerable to bad local optima [41]. Therefore, Zhao et al. [42] seek a method that can intentionally generate difficult triples to optimize the overall network of the network, instead of using a greedy strategy to explore existing samples only for the current network. As shown in Fig. 8, to generate the goal of difficult triples, a Hard Triplet Generation (HTG) network algorithm is proposed to optimize the network’s ability to distinguish similar samples of different categories and group different samples of the same category.

Chopra et al. [43] proposed a novel Grid Search Network (GSN) to learn feature embedding for clothing retrieval. Similar to the triplet network variant, this method assumes that the training process is a search problem, and it finds a match of reference sample images in a grid containing positive and negative images. The framework also uses reinforcement learning-based strategies to learn special feature vector conversion functions, instead of simply connecting feature vectors. When applied to feature embedding networks, it further improves the clothing image retrieval accuracy. Kuang et al. [44] proposed a Graph Reasoning Network (GRNet), the similarity pyramid network, which uses global similarity to learn and query the similarity between clothing images and clothing databases.

**Ensemble network**

Ensemble is a widely used method of training multiple learners to obtain a combined model, and its performance is better than a single model [45,46]. For deep metric learning, the ensemble network connects the feature embedding learned by multiple learners. Under the constraint of the distance between a given image pair, a better embedding space can usually be obtained. A better ensemble network depends on the high performance of individual learners and the diversity...
between learners. However, in deep metric learning, there is not much research on the optimal architecture to generate feature embedding diversity.

According to the above-mentioned use of Siamese network or triplet network for deep metric learning, they have achieved better results. Although the image of clothing of the same category is closer, and clothing of different categories is further away from the image. However, it is difficult to directly optimize the target because of the size of the sample. Therefore, difficult sample mining is widely used to solve this problem, and it costs expensive calculations on a subset of samples considered to be difficult. However, the difficulty sample is defined relative to a specific model. Such a complicated model will treat most of the samples as easy samples, and a simple model will treat most of the samples as difficult ones, both of which are not conducive to training. Since different samples have different difficulty levels, it is difficult to define a moderately complex model, and it is also difficult to fully select difficult samples. To solve the above problems, we summarize and analyze the different methods put forward by researchers.

The above-mentioned triplet loss and variant methods are only based on a single model to mine difficult sample images, and cannot make full use of samples of different difficulty levels. Therefore, Yuan et al. [47] proposed the Hard-Aware Deeply Cascaded (HDC) Embedding model, which uses increasingly complex model sets in a cascaded manner to mine negative samples of different difficulty levels during the training process. They take advantage of the deep supervision network [48,49], and use a contrastive loss function to train the lower layers of the network to handle easier samples, and the higher layers of the training network to handle more difficult samples. Compared with this multi-layer method, the Boosting Independent Embedding Robustly (BIER) model [50] uses an ensemble method of high-dimensional embedding method, focusing on reducing the correlation on a single layer, and dividing the high-dimensional embedding into several learners, trained with Online Gradient Boosting (OGB). Continuous learners are trained on reweighted samples, which greatly reduces the correlation between learners, thereby reducing the correlation within the embedding and improving the robustness of the embedding. In addition, compared with the HDC model, the method allows continuous weighting of samples according to the loss function. Inspired by the BIER method, Xuan et al. [51] proposed a different method to learn robust, high-dimensional embedding spaces. Instead of reweighting the input samples to create independent output embedding.

As an important aspect of ensemble network, learners should have diversity in feature embedding. So, Kim et al. [52] proposed an Attention-based Ensemble (ABE) model, as shown in Fig. 9, (a) based on ordinary ensemble learning; (b) based on attention ensemble learning. The model ensembles multiple attention models so that each learner can pay attention to different parts of the object. For different positions of different learners, different feature embedding functions are trained. A divergence loss is proposed to regularize features to distinguish feature embedding from different learners. All in all, although the ensemble method increases the complexity of the model, it can further improve the accuracy of cross-domain clothing retrieval.

**Comparison of experimental methods**

**Clothing databases**

As shown in Table 2, the detailed introduction to the clothing databases. In recent years, the size of popular clothing databases and types of annotations is different. For example, Street2Shop and DARN contain 425K and 540K clothing images respectively. They contain two types of images (1) street images, which are images of people actually wearing clothes under daily uncontrollable environmental conditions; (2) shop images, which are clothing images of online clothing stores, which are made by professionals shot of a more controlled environment. Since the clothing category tags are extracted from the metadata of the images collected by online shopping sites, this makes their tags appear a lot of errors and confusion. DeepFashion and ModaNet obtains labels by manually annotating clothing categories. In addition, different types of annotations are provided with these databases. DeepFashion is a large clothing database with comprehensive annotations, with four benchmarks, among which Consumerto-shop Benchmark is a database corresponding to street images and store images, and each clothing item’s folder contains a street image and several store images, with a total of 33,881 clothing items and 239,557 clothing images. Each image has 4–8 clothings functional regions (such as “collar”) and other related fashion labels. The recognitions of these fashion landmarks are shared with all clothing categories, which makes it difficult for them to capture the
Table 2  The detailed introduction of the different clothing databases

| Databases      | Institutions                                      | Details                                                                 | Download                                      |
|----------------|---------------------------------------------------|------------------------------------------------------------------------|-----------------------------------------------|
| Street2Shop    | University of Illinois at Urbana-Champaign        | 404,683 store images and 20,357 street images. 39,479 street and store pairs. 11 categories | http://www.tamaraberg.com/street2shop.       |
| DARN           | National University of Singapore                  | 450,000 store and 90,000 street images. Each image contains 5–9 semantic attributes. 20 categories | arXiv:1505.07922                             |
| DeepFashion    | The Chinese University of Hong Kong               | 800,000 images, including four benchmarks, with a large amount of clothing labeling information, 50 categories | http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html |
| ModaNet [53]   | eBay Inc                                          | 55,000 street images, clothing is more fashionable. 13 categories. Used in clothing object detection and segmentation | https://github.com/hrsm2i/modanet             |
| FashionAI [54] | The Hong Kong Polytechnic University, Hong Kong SAR | 375,000 images, 6 categories of women’s clothing, a total of 41 subcategories, with landmark annotations | http://tianchi.aliyun.com/markets/tianchi/FashionAI |
| DeepFashion2   | The Chinese University of Hong Kong               | 491,000 images. 13 categories with more annotation information. A comprehensive database that can be used for clothing detection, pose estimation, segmentation and retrieval | http://github.com/switchablenorms/DeepFashion2 |

rich variety of clothing images. In contrast, ModaNet’s street images have a mask of a single person, but there are no landmarks. Unlike the datasets above, DeepFashion2 contains 491K images and 801K annotations of landmarks, masks, and bounding boxes and 873,000 pairs of images, which is the most comprehensive benchmark in a clothing dataset.

**Experiment preparation**

The hardware and software environment used in the experiment is: Intel(R) Core(TM) i5-3570 CPU @ 3.40 GHz processor, NVIDIA GeForce GTX 1070 8GB graphics card, 8 GB memory. The operating system is Ubuntu 16.04, and the programming language is Python, the deep learning framework is Pytorch.

As shown in Table 3, the datasets used in this paper are two subsets of the DeepFashion dataset, namely In-shop Clothes Retrieval Benchmark subset and Consumer-to-shop Clothes Retrieval Benchmark subset.

**The evaluation of clothing retrieval**

There are many clothing databases, but the commonly used evaluation methods in clothing retrieval are as follows: precision, MAP, and accuracy.

1. The precision is shown in Eq. (4):

\[ P = \frac{A}{B}, \quad (4) \]

where \( A \) is the number of similar clothing in the search results, and \( B \) represents the number of search results returned. From the precision rate formula, it can be seen that the precision rate can effectively investigate the proportion of the correct return results of the retrieval model in the retrieval results in all the returned results.

2. Although the precision rate is a statistical evaluation of the proportion of correct search results, it lacks the evaluation of the location information of the search results. Therefore, the MAP value is used as the evaluation of the location information of the search results. The MAP is shown in Eq. (5):

\[ \text{MAP} = \frac{1}{Q} \sum_{q=1}^{NAP} (a), \quad (5) \]

where \( Q \) is the number of clothing images in the retrieval database, which represents the average correct rate and the change in recall rate, that is, the regions under the \( P-R \) curve. MAP can reflect the overall performance of the retrieval method, but lacks insight into the details of the retrieval results.

3. The accuracy of cross-domain clothing retrieval is the most commonly used evaluation criterion. The Top-k method is generally used, as shown in Eq. (6):

\[ P@K = \frac{1}{|Q|} \sum_{q \in Q} \text{hit}(q, k), \quad (6) \]

where \( Q \) is the number of clothing images in the search database, and \( q \) represents the specific clothing image to
Table 3: A detailed introduction to the experimental dataset in this paper

| Datasets           | Introduction                     | Number of training sets | Number of test sets |
|--------------------|----------------------------------|-------------------------|---------------------|
| Consumer-to-shop   | Cross-domain clothing dataset    | 8000                    | 2000                |
| In-shop            | Same domain clothing dataset     | 8000                    | 2000                |

Table 4: Different models based on critical region recognition

| Models   | Year | Attribute | Landmark |
|----------|------|-----------|----------|
| WTBI     | 2015 | Y         | N        |
| DARN     | 2015 | Y         | N        |
| FashionNet | 2016 | Y         | Y        |
| VAM      | 2017 | Y         | N        |
| TagCtxYNet | 2017 | Y         | N        |
| MGN&SCN  | 2019 | Y         | N        |
| Match R-CNN | 2019 | Y         | Y        |
| DMCH     | 2019 | Y         | N        |

Table 5: Different models based on deep metric learning

| Models   | Year | Attribute | Landmark |
|----------|------|-----------|----------|
| Partial-sharing | 2016 | N         | N        |
| HDC      | 2017 | N         | N        |
| BIER     | 2017 | N         | N        |
| HTL      | 2018 | N         | N        |
| HTG      | 2018 | N         | N        |
| ABE      | 2018 | N         | N        |
| GSN      | 2019 | N         | N        |
| GRNet    | 2019 | N         | N        |

be queried. If at least one clothing image in the Top-k list matches the image \( q \), then \( \text{hit}(q, k) \) will be set to 1, otherwise it will be set to 0.

Analysis of the results of the experiment

By summarizing the clothing image retrieval models based on deep learning in recent years, they mainly solve some difficulties in cross-domain situations. Tables 4 and 5, respectively, show two different models based on clothing critical region recognition and deep metric learning two ideas. Among them, “Y” and “N” indicates whether the algorithm uses the corresponding attribute and landmark annotations. It can be seen from the table that the network model based on critical region recognition has higher requirements for clothing to attribute labeling, and supervised learning is generally used. Even if the attention mechanism is used to greatly reduce the annotations of landmark in clothing, some weakly supervised networks are proposed, but some annotations of clothing attributes are still needed. However, most of the network models based on deep metric learning do not need landmark annotations or clothing attribute annotations, because deep metric learning is based on the characteristics of the image itself, in-depth mining of different difficult samples, and strengthening the discrimination of different extracted clothing features, then further extracts the important clothing features that are discriminative.

At present, most popular clothing retrieval networks are implemented based on the DeepFashion database. The DeepFashion database has two subsets, Consumer-to-shop Benchmark and In-shop Benchmark. The application scenarios are cross-domain clothing image retrieval and same-domain clothing image retrieval, as shown in the figure. As shown in Fig. 10, the performance of using the critical region recognition idea in solving the cross-domain clothing retrieval problem. The figure shows that different clothing critical region recognition algorithms have a greater impact on clothing retrieval performance. Among them, clothing landmark recognition and attention map recognition have a higher accuracy in clothing retrieval. However, FashionNet and Match R-CNN networks use clothing landmark recognition to rely too much on clothing attributes and landmark annotation information in the retrieval process, while the attention maps recognition method can solve this problem better in clothing retrieval, it can still achieve better retrieval accuracy without the clothing landmark annotation, which provides new ideas of cross-domain clothing image retrieval.

As shown in Fig. 11, the performance of several deep metric learning-based clothing image retrieval models in different datasets. From the comparison of the performance of (a) and (b), We find that the deep metric learning is better in the same domain clothing image retrieval than the cross-domain clothing image retrieval, because the same domain clothing image itself is less affected by the external environment. Mainly considering the influence of the inherent attributes of clothing images, deep metric learning can use different loss functions combined with network models to achieve better similar clothing matching. However, for cross-domain clothing image retrieval, deep metric learning using contrastive loss, ternary loss and variant, and ensemble learning needs to consider the influence of background and other factors, and needs to mine difficult samples. At present, ensemble learning is widely used, and it can be very use-
It is good to mine samples of different difficulty levels to improve the accuracy of cross-domain clothing retrieval. Clothing image retrieval requires feature extraction and similarity match. Different research methods have different focuses. Fig. 12 and Table 6 show the performance of deep network models in cross-domain clothing retrieval in recent years. It can be seen that the overall effect of deep metric learning is not as good as clothing critical region recognition of solving cross-domain retrieval problems, indicating that the main problem to be solved in cross-domain clothing retrieval are the recognition of clothing important regions in the image. This is important step in using convolutional neural networks to extract features, then combine clothing attributes can achieve better retrieval results. Therefore, in the future, we can combine critical region recognition and deep metric learning to propose a new algorithm without additional annotation information, it can achieve better cross-domain clothing image retrieval accuracy.

**Conclusion**

This paper reviews the cross-domain clothing retrieval methods. First, it analyzes the common methods of the critical
region recognition and deep metric learning in cross-domain clothing retrieval. Then, the research results show that attention map recognition method not only saves time and cost, but also further improves the effect of clothing retrieval. Deep metric learning research is widely used, and has achieved better results in the same-domain clothing retrieval and cross-domain clothing retrieval. At last, we find that the critical region recognition can extract more important clothing detail features, and deep metric learning makes the extracted features more discriminative, both affect the effect of cross-domain clothing retrieval.

In summary, although cross-domain clothing retrieval has achieved better retrieval results using clothing critical region recognition and deep metric learning methods, there are still many issues need to be solved, it mainly includes:

1. Attribute labeling problem: Most of the deep network models need the assistance of clothing attribute labels, which is a supervised learning or weakly supervised method. It requires high clothing labeling and is a time-consuming and labor-intensive work. How to reduce clothing attribute labels, save costs, and improve accuracy still needs further research.

2. Model complexity problem: In recent years, the research of cross-domain clothing image retrieval tasks has mainly focused on ensemble methods. Although better results have been achieved, the long training time and memory loss brought by model ensemble methods are difficult to solve. Therefore, how to solve the high model complexity brought by ensemble learning under the premise of ensuring the retrieval effect is a big challenge.

3. Clothing databases: At present, the clothing databases contain different types of clothing, which are distinguished from clothing categories, such as dresses, jeans, shirts, etc. However, due to the development of the clothing fashion industry, different clothing combinations can produce ever-changing clothing styles, such as sports, Japanese, punk, etc. The retrieval of clothing styles will be another important research direction in the future.

Acknowledgements This work is supported in part by China National Textile and Apparel Council No. 2018097, National Natural science Foundation of China (61902301), Shaanxi Provincial Education Department (19JK036418JK0334) and Science and Technology Plan Project of Shaanxi Province (2022JM-146,2022JZ-35). Thanks all reviewers.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding This work is supported in part by China National Textile and Apparel Council No. 2018097, National Natural science Foundation of China 61902301, Shaanxi Provincial Education Department 19JK036418JK0334 and Science and Technology Plan Project of Shaanxi Province 2022JM-146, Thanks all reviewers.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Korea Federation of Textile Industries (2019) Korea fashion market trend 2019 Report; Korea Federation of Textile Industries: Seoul, Korea
2. Korea Fashion Association (2019) Global fashion industry survey. Seoul, Korea, Korea Fashion Association
3. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR. arXiv:1409.1556
4. Lin M, Chen Q, Yan S (2013) Network in network. arXiv:1312.4400
5. Chen L, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. CoRR. arXiv:1706.05587
6. 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
7. Kiapour MH, Han X, Lazebnik S (2015) Where to buy it: matching street clothing photos in online shops. IEEE Int Conf Comput Vis 2015:3343–3351
8. Sande KEA van de, Uijlings JRR, Gevers T, Smelnders AWM (2011) Segmentation as selective search for object recognition. In: ICCV
9. Chen Q, Huang J, Feris R, Brown L, Dong J, Yan S (2015) Deep domain adaptation for describing people based on fine-grained clothing attributes. In: CVPR
10. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR

11. Huang J, Feris RS, Chen Q (2015) Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of 2015 international conference on computer vision, pp 1062–1070

12. Ning C, Menglu L, Hao Y, Xueping S, Yunhong L (2020) Survey of pedestrian detection with occlusion. Compil Intell Syst 2020:5

13. Eichner M, Ferrar V (2012) Appearance sharing for collective human pose estimation. Comput Vis ACCV 2012:38–151

14. Chen H, Andrew G, Bernd G (2012) Describing clothing by semantic attributes. In: Proceedings of the 12th European conference on computer vision, pp 609–623

15. Chen K, Luo T, Jai B (2017) When fashion meets big data: discriminative mining of best selling clothing features. In: Proceedings of the 26th international conference on world wide web companion, pp 15–22

16. Pedro FF, Ross BG, David AM (2010) Object detection with discriminatively trained part-based models. IEEE Trans 2010:1627–1645

17. Viola P, Jones M (2001) Robust real-time object detection. Int J Comput Vis 57:2

18. Carten R, Vladimir K, Aadrew B (2014) “GrabCut”: interactive foreground extraction using iterated graph cuts. ACM Trans 23(3):309–314

19. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: powering robust clothes recognition and retrieval with rich annotations. Comput Vis Pattern Recogn 2016:1096–1104

20. Liu Z, Yan S, Lou P (2016) Fashion landmark detection in the wild. In: proceedings of the 14th European conference of computer vision (ECCV), pp 229–245

21. Ge Y, Zhang R, Wu L (2019) DeepFashion2: a versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: proceedings of the 2019 conference on computer vision and pattern recognition, pp 5337–5345

22. Ji X, Wang W, Liu MH, Yang Y (2017) Cross-domain image retrieval with attention modeling. In: Proceedings ACM on multimedia conference, ACM, pp 1654–1662

23. Wang Z, Gu Y, Zhang Y, Zhou J, Gu X (2017) Clothing retrieval with visual attention model. IEEE Vis Commun Image Process 2017:5

24. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2015:640

25. Zheng Y, Huang D, Liu S, Wang Y (2020) Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13766–13775

26. Luo Z, Yuan J, Yang J, Wen W (2019) Spatial constraint multiple granularity attention network for clothes retrieval. In: 2019 IEEE international conference on image processing (ICIP), IEEE, pp 859–863

27. Xu K, Ba J, Kirov R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille, France, pp 2048–2057

28. Luo Y, Wang Z, Huang Z, Yang Y, Lu H (2019) Snap and find: deep discrete cross-domain garment image retrieval. IEEE Trans Image Process 2019:5

29. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: Computer vision and pattern recognition (CVPR), pp 539–546

30. Bell S, Bala K (2015) Learning visual similarity for product design with convolutional neural networks. ACM Trans Graph 34(4):98
51. Xuan H, Souvenir R, Pless R (2018) Deep randomized ensembles for metric learning. In: ECCV, pp 723–734
52. Kim W, Goyal B, Chawla K, Lee J, Kwon K (2018) Attention-based ensemble for deep metric learning. In: ECCV, pp 760–777
53. Zheng S, Yang F, Kiapour MH, Piramuthu R (2018) Modanet: a large-scale street fashion dataset with polygon annotations. In: ACM multimedia
54. Zou X, Kong X, Wong W, Wang C, Liu Y, Cao Y (2019) Fashionai: a hierarchical dataset for fashion understanding. In: CVPR workshop

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