Which is Plagiarism: Fashion Image Retrieval based on Regional Representation for Design Protection

Yining Lang[1], Yuan He[1], Fan Yang[1], Jianfeng Dong[2,3], Hui Xue[1] *
Alibaba Group[1], Zhejiang Gongshang Univresity[2],
Alibaba-Zhejiang University Joint Institute of Frontier Technologies[3]

Abstract

With the rapid growth of e-commerce and the popularity of online shopping, fashion retrieval has received considerable attention in the computer vision community. Different from the existing works that mainly focus on identical or similar fashion item retrieval, in this paper, we aim to study the plagiarized clothes retrieval which is somewhat ignored in the academic community while itself has great application value. One of the key challenges is that plagiarized clothes are usually modified in a certain region on the original design to escape the supervision by traditional retrieval methods. To relieve it, we propose a novel network named Plagiarized-Search-Net (PS-Net) based on regional representation, where we utilize the landmarks to guide the learning of regional representations and compare fashion items region by region. Besides, we propose a new dataset named Plagiarized Fashion for plagiarized clothes retrieval, which provides a meaningful complement to the existing fashion retrieval field. Experiments on Plagiarized Fashion dataset verify that our approach is superior to other instance-level counterparts for plagiarized clothes retrieval, showing a promising result for original design protection. Moreover, our PS-Net can also be adapted to traditional fashion retrieval and landmark estimation tasks and achieves the state-of-the-art performance on the DeepFashion and DeepFashion2 datasets.

1. Introduction

Fashion-related works have attracted increasing attention, due to the boom of online shopping in these years. The rapid growth of deep learning-based approaches further enhances the ability of fashion image classification [30, 34], fashion landmark detection [39, 27], and fashion retrieval [45, 4, 49, 31]. The traditional clothes retrieval methods [26, 16] typically perform similarity learning in the entire instance of clothes without any focus, which is easily interfered by irrelevant features. The recent clothes retrieval methods [2, 4, 20, 49] learn the attribute representations to guide the retrieval, thus improve the performance.

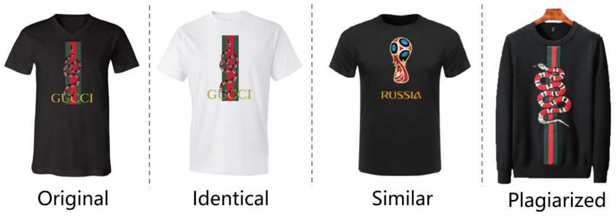

Figure 1. Examples for identical, similar, and plagiarized clothes with respect to the original item.

Different from the exiting methods [26, 45, 2, 4, 49] typically aim to retrieve visually similar or identical clothes, we focus on a novel problem of *plagiarized clothes* retrieval. The plagiarized clothes retrieval is somewhat ignored in the academic community, while it has great application value in the industry. The similar clothes retrieval task is somewhat similar to the plagiarized clothes retrieval task, as some retrieved similar items may be plagiarized one. However, plagiarized items are not always very similar to the original fashion items. As shown in Figure 1, the plagiarized item is relatively more dissimilar than the similar item with the original one. Hence, the retrieved target of both tasks is different. Moreover, in the plagiarized clothes retrieval task, the ground-truth images may be in a different category with the original item (a long-sleeved T-shirt and a short-sleeved T-shirt in the example). But in the similar or identical clothes retrieval task, they are usually in the same category. It also shows that the plagiarized clothes retrieval task is more challenging.

Actually, plagiarized clothes are very complicated and appear in a wide variety of forms. For example, an item which only plagiarizes the design of a certain part can be considered as plagiarism or an item which completely copies another item without any authorization, etc. Moreover, the form of plagiarized clothes is dynamic, as illegal businesses continue to update their plagiarized ways. Therefore, it is difficult to use a uniform definition to include all plagiarized types. As the first work for plagiarized clothes retrieval task, we initially define the plagia-

*Corresponding Author: Hui Xue (hui.xueh@alibaba-inc.com).

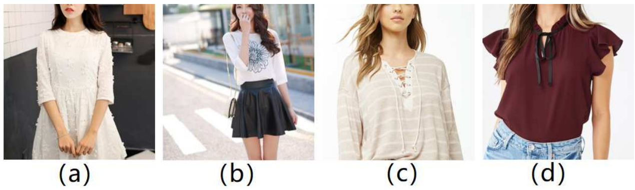

Figure 2. Hard cases for attribute-driven retrieval method: Indistinguishable sleeves (a & b); Unrecognizable collars (c & d).

rized clothes as samples that are modified in less than or equal to two regions on the original design (e.g., change the shape of the collar, modify the pattern within the chest region). This kind of plagiarized clothes occupies the high proportion in e-commerce platforms. Besides, the defined plagiarized clothes are relatively easy for evaluation, thus helpful for the study of the plagiarized clothes retrieval task.

In the fashion-related works [1, 2], clothes attributes are common used. However, clothes attribute is somewhat subjective, which is not very suitable for the plagiarized clothes retrieval task. For instance, it is difficult to judge the length of the sleeves or the style of the collar in some hard cases, as Figure 2 shows. Besides, for some clothes with deformations and occlusions, the retrieval performance also decreases obviously. On the contrary, the geometric properties of the clothes are highly deterministic and can maintain stability for deformed and occluded samples. Hence, we propose a novel PS-Net based on regional representation, where clothes landmarks are employed to guide the learning of regional representations and clothes are compared region by region. Besides, we find that different categories of plagiarized clothes are easy to be modified in different regions. Therefore, we would like to learn different groups of region weights for each category of clothes in order to manipulate the region weights automatically during similarity learning. By doing so, a plagiarized clothes image with a modified region could be recalled more easily. Additionally, there is no available dataset for the plagiarized clothes retrieval task. Hence, we collect a new dataset named "Plagiarized Fashion", where clothes images are annotated by experts who majored in intellectual property protection.

In summary, the major contributions of our paper are:

- We introduce a novel problem of plagiarized clothes retrieval and a new dataset named "Plagiarized Fashion" for plagiarized clothes retrieval, which provides a meaningful complement to the fashion retrieval field.
- A multi-task network named PS-Net based on the regional representation is proposed, which is superior to other instance-level counterparts for plagiarized clothes retrieval.
- Besides the plagiarized clothes retrieval, our proposed PS-Net can also be used for traditional fashion retrieval and landmark estimation tasks, achieving the state-of-the-art performance on both DeepFashion [27] and DeepFashion2 [14] datasets.

2. Related Work

Visual Fashion Analysis. Visual fashion works have attracted lots of attention due to the boom of e-commerce and online shopping in these years. With the development of large-scale fashion datasets [27, 14], deep learning-based techniques further boosted the interest in fashion-related tasks, like clothes recognition [6, 17, 19], retrieval [16, 26, 45, 2, 49], recommendation [23, 18], clothes synthesis [5, 24] and fashion landmark detection[28, 39]. Recently, some multi-task neural network, such as FashionNet [27] and Match-RCNN [14] can even perform the above tasks simultaneously. Earlier works [40, 12] on clothes recognition mostly relied on hand-crafted features, such as SIFT [29], HOG [11]. The performance of these methods was limited by their ability of feature representation. Recently, plenty of deep learning-based models have been introduced to learn more discriminative representation [49, 22], which can even handle cross-domain scenarios [16] and near-duplicate detection task [33]. Moreover, some related works have performed clothes retrieval using parsing [45, 44], or achieved the search by attribute-driven methods [12, 1, 2, 49]. However, we found in practice that, for retrieving the images of plagiarized clothes, the existing methods are not effective enough due to the characteristic of plagiarized clothes: modified less than or equal to two regions on the original design.

Different from the above works, in this paper, we focus on the new task of plagiarized clothes retrieval. To the best of our knowledge, this paper is the first work for plagiarized clothes retrieval. Besides, the task is aimed to retrieve the plagiarized clothes with regional manipulation, which to some extent has a similar idea with Deepfake detection tasks [7, 15].

Landmark Guided Attention. Landmark detection technique is widely used in many tasks nowadays, like face alignment [42] and human pose estimation [36]. To obtain much stronger feature representations of clothes, fashion landmark estimation task is proposed in recent years [28, 46, 39]. On the other hand, attention technique is also an effective way to obtain stronger feature representations. Previous works [43, 47, 38] have proved that attention mechanism is helpful due to it enables the network to focus on the critical features and filter out the irrelevant ones.

Given an image, the typical attention model learns to obtain one whole image feature vector by weighted summing with attention weights. However, in this work, we go a step further by dividing fashion items into several regions under the guidance of predicted landmarks and learning to obtain several weighted region feature vectors. With the proposed regional attention, we compare images region by region and find it is better than the typical attention for the plagiarized clothes retrieval task.

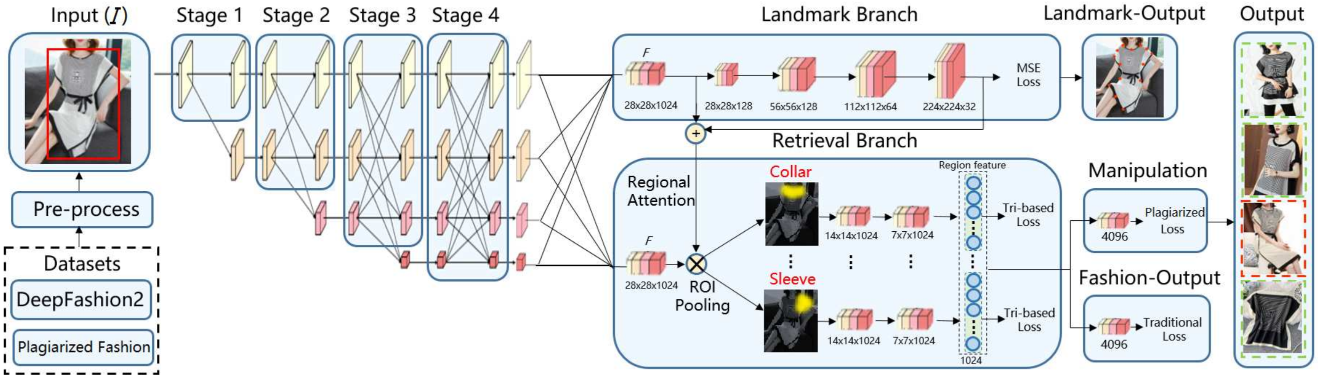

Figure 3. Structure of the proposed PS-Net which consists of a landmark branch and a retrieval branch, based on the HR-Net backbone (some convolution layers are hidden). Two output feature maps $F \in \mathcal{R}^{28\times28\times1024}$ of the backbone are identical for demonstration. The landmark guided regional attention is introduced to the retrieval branch during the ROI pooling. The retrieval branch is also split into two parts for output, one for traditional fashion retrieval and the other for plagiarized clothes retrieval. The images bounded by green and red boxes indicate plagiarized clothes and identical clothes, respectively.

3. Our Approach

Our work aims to retrieve the images of plagiarized clothes which are modified in less than or equal to two regions on the original design. Hence the key is to compute the similarity between two images of clothes. To this end, we propose a Plagiarized-Search-Net (PS-Net), which obtain the regional representation of images and compute the similarity region by region. Specially, given an image of clothes I, we propose to represent the image by multiple regional features $f_1(I), f_2(I), ...f_R(I)$, where R is the number of image regions. Besides, we find from practice that different categories of clothes are easy to be plagiarized in different regions. Therefore, we would like to learn the region weights $(\lambda_1, \lambda_2, ...\lambda_R)$ for different categories of clothes in order to manipulate the region weights automatically for plagiarized clothes retrieval. Finally, the similarity between images of clothes I and I' is:

$$\sum_{r=1}^{R} \lambda_r \cos(f_r(I), f_r(I')), \tag{1}$$

where $\cos$ indicates cosine similarity between two feature vectors. Figure 3 illustrates the structure of our proposed PS-Net, it is composed of a backbone, a landmark branch and a retrieval branch. As our PS-Net has a landmark branch, so it can also be used for fashion landmarks detection task.

In what follows, we firstly describe the detailed structure of our proposed PS-Net, followed by the description of its optimization.

3.1. Network Architecture

Network Backbone. In the proposed PS-Net, we choose the HR-Net [36] as our backbone. With its multi-stage parallel structure, the HR-Net can maintain high resolution in deep networks, which is especially important for landmark estimation task. Note that the choice of the backbone is not mandatory, which can be replaced by any backbone with a similar effect (e.g., ResNet [19], VGG-Net [35]). Besides, as shown in Figure 3, the landmark branch and the retrieval branch in PS-Net share the same type of backbone (but not identical one). Before feeding an image of clothes to the backbone, we first detect the clothes in the image. Hence we trained a Faster R-CNN [32] (Res50-FPN) model on the DeepFashion2 [14] dataset as a detector to obtain the clothes and their category labels. The cropped images are resized to 224×224 pixels as the input I.

Landmark Branch. We design a landmark branch to predict landmarks on each image of clothes. More specifically, we transform the fashion landmark estimation task to predicting k heatmaps, where each the i-th heatmap indicates the location confidence of the i-th landmark. Given the output feature map F of the backbone, we use one 1×1 convolution to convert it to $28 \times 28 \times 128$. Then, several groups of transposed convolution are utilized to produce a high-resolution landmark heatmap with the same scale as the input. Finally, we use a regressor to estimate the heatmaps where the landmark positions are chosen.

Regional Attention-based Retrieval Branch. On the other hand, the output feature map F of the backbone is fed to the retrieval branch. In our experiment, we first train the model on the Deepfashion2 [14] dataset to obtain the ability for identical clothes image retrieval. After that, we get a pretrained model for further step training on plagiarized clothes retrieval task. Utilizing the regional representation achieved by the landmark branch, we fine-tune the retrieval model by manipulating the region weights. Finally, we can

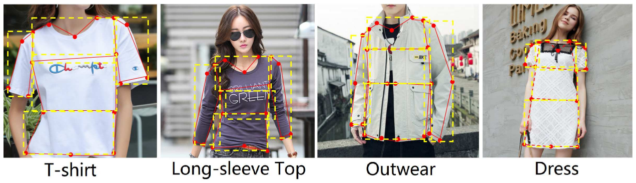

Figure 4. The visualization of the landmark guided region division. Five bounding boxes are estimated which covers the largest segmented region, respectively.

obtain one retrieval model with two types of output form, which are "Fashion Output" and our target plagiarized "Output", as indicates in Figure 3.

The attention generated by the landmark branch is introduced to the retrieval branch by the following process: Firstly, we take the concatenation of the representations output $F \in \mathcal{R}^{28\times28\times1024}$ by the backbone and the bilinear downsampled landmark information $M_{ij} \in \mathcal{R}^{28\times28\times32}$ as the input. Second, we reshape the input attention map A to $28\times28\times1024$, which has the targeted scale of the retrieval branch. Then, inspired by previous fashion analysis work [25], the attention is introduced to the retrieval branch by making $F^{'} = F \circ (1/2 + A)$, where $\circ$ stands for Hadamard product. By adding $1/2$ to the attention feature map, the range of the element becomes $(1/2, 3/2)$. The critical features are strengthened by elements greater than 1, while irrelevant features are filtered out via elements less than 1. For instance, the landmarks around critical areas like cuff and collar can guide the extraction of features, which makes these key features have more possibility to retain.

To learn regional representations for the plagiarized retrieval task, we go a step further by dividing fashion items into several regions, as shown in Figure 4, under the guidance of predicted landmarks. Five bounding boxes are estimated as regions of proposal, which covers the largest segmented area, respectively. Then, we achieve an ROI pooling based the proposed regions on the feature map of the Hadamard product. By this way, the landmark guided regional attention is introduced to the retrieval branch and the input image I is represented by multiple regional features.

Different from the previous work [50] in the field of person re-ID, which generates regions by an RPN [32] network for feature decomposition, we directly divide the regions by the distribution of landmark outputs. In this way, the regions generated by our approach are explicit rather than implicit, which is more controllable with the high accuracy of landmark estimation.

3.2. Optimization

The optimization process of our approach can be divided into two phases: a pre-trained phase and a fine-tune phase.

Pre-trained Phase. For the landmark branch, we choose the mean squared error (MSE) as our loss function. The ground-truth heatmaps are generated by applying 2D Gaussian with a standard deviation of 1 pixel centred on the location of each landmark.

Category	Sleeves	Collar	Chest	Waist	Sum
T-shirts	12%	6%	72%	10%	15,300
Tops	13%	34%	38%	15%	15,500
Outwear	25%	21%	33%	21%	14,200
Dress	6%	17%	21%	56%	15,000

Table 1. The distribution of modified regions among different categories of plagiarized clothes in our proposed Plagiarized Fashion Dataset.

For the regional attention based retrieval branch, we utilize triplet (tri) ranking loss which are commonly used in the retrieval tasks [13, 48]. Formally, the loss is defined as

$$\mathcal{L}_{tri}(I, I^+, I^-) = \sum_{r=1}^{R} \max(D_r^{I,I^+} - D_r^{I,I^-} + m, 0) \quad (2)$$

$$\mathcal{L}_{tra} = \sum_{n=1}^{N} \mathcal{L}_{tri}(I, I^+, I^-) \quad (3)$$

where I corresponds to the input image, N is the number of training examples, R is the number of the regions and m represents the margin. The loss aims to minimize $D_r^{I,I^+} = ||f_r(I) - f_r(I^+)||_2$ and maximizing $D_r^{I,I^-} = ||f_r(I) - f_r(I^-)||_2$. $f_r(I^+)$ and $f_r(I^-)$ represent the feature maps of image I^+ and I^- corresponded to region r, respectively. Note that the triplets are chosen from identical mini-batch. For each triplet: I and I^+ must share the same label while I^- is chosen randomly from others. By doing so, images of identical clothes are made to be close to each other in the feature space. After that, we also combine region representations $(f_1, f_2, ...f_5)$ into a global one. The concatenated feature is used to obtain the "Fashion Output" mentioned in Figure 3.

In general, our approach is able to learn critical features representations by leveraging landmark guided regional attention, which can increase the focus on specific regions during the training process. Also, the geometric properties of clothes are highly stable with few false predictions compared to the attribute-driven ones, which can enhance the retrieval performance for some hard samples (e.g., samples with deformations and occlusions).

Fine-tune Phase. The most challenging problem for plagiarized clothes retrieval is that plagiarists typically modify the clothes in a certain region on the original design to escape the supervision by traditional retrieval methods. We find from practice that different categories of clothes are easy to be plagiarized in different regions, as shown in Table 1. Therefore, we would like to learn the region weights for different categories of clothes in order to manipulate the region weights automatically for plagiarized clothes retrieval.

During the training, each image of clothes is divided into 5 regions (2 sleeves included) automatically, guided by the geometric distribution of landmarks. On the output features of the last convolution layer, plagiarized retrieval loss $\mathcal{L}_{pla}$ as shown below is imposed to enable region weights learning, which shares the same network framework with traditional fashion retrieval:

$$\mathcal{L}_{tri}^{'}(I,I^{+},I^{-}) = \sum_{r=1}^{R} \max(D_{r}^{I,I^{+}} - D_{r}^{I,I^{-}} + m, 0) \cdot \lambda_{r}, \tag{4}$$

$$\alpha_{tri} = \frac{\mathrm{avg}\{||f_{r}(I) - f_{r}(I^{+})||_{2}; r = 1,2,...R\}}{\max\{||f_{r}(I) - f_{r}(I^{+})||_{2}; r = 1,2,...R\}}, \tag{5}$$

$$\mathcal{L}_{pla} = \sum_{n=1}^{N} [\mathcal{L}_{tri}^{'}(I,I^{+},I^{-}) \cdot \alpha_{tri}]. \tag{6}$$

The loss $\mathcal{L}_{pla}$ is only used to update the weight λ_r of each region, which is decoupled with the parameter update of traditional retrieval task. $\mathcal{L}_{tri}^{'}$ is a triplet-based loss which contains region weights λ_r. α_{tri} is the weight for loss functions $\mathcal{L}_{tri}^{'}$, which is updated during the training. Through the adjustment of α_{tri}, the loss $\mathcal{L}_{tri}^{'}$ of samples with large feature difference in a single region and small difference in other regions will be lower.

We utilize Coordinate Ascent as our optimization method. The λ_r of each region is set to 1 with a step size $\Delta\lambda$ of 0.1 at the beginning. The step size drops to 0.05 after 40 epochs, and 0.01 after 60 epochs. The weights of each region are sampled with a step size before each iteration ($\lambda_r \pm \Delta\lambda \rightarrow \lambda_r^{'}$). After each iteration, if the loss decreased, the current weight $\lambda_r^{'}$ is accepted; otherwise, the weight turn back to λ_r. Note that the weights of the five regions (2 sleeves included) are always normalized in proportion to ensure a sum of 1. The weights of each region are updated iteratively to reduce the loss until the last epoch.

Finally, we also combine region representations into a global one to complete the plagiarized clothes search. The retrieval branch can recall more partially modified samples by manipulating the region weights of the features. Note that the region weights for four categories of clothes are trained separately.

4. Plagiarized Fashion Dataset

Fashion datasets (e.g., DeepFashion [27], Shopping 100K [3]) provide a variety of data for the training of clothes retrieval model. But there is not yet a benchmark dataset for the retrieval of plagiarized clothes. Hence, in this paper, we propose a new dataset named Plagiarized Fashion for plagiarized clothes retrieval. The dataset contains 60,000 images in total, where 40,000 images for training, and 20,000 images for testing. Among them, 1500 are query images, and the others are gallery images. The dataset consists of four categories of clothes: short-sleeved T-shirts, long-sleeved tops, outwears and dresses. The numbers of samples for them are approximately balanced. Table 1 shows the distribution of modified regions among different categories of plagiarized clothes. Since the design of shorts, trousers and skirts are not recognizable enough, we do not include these three types of clothes. We consider expanding the category of clothes in future work to enable more powerful design protection ability.

We collect the dataset by crawling from Taobao, the biggest e-commerce website in Asia. Given an original clothes image, we can obtain a set of images (top-100) of similar clothes on the website by the traditional retrieval method. Then, we invite three experts who majored in intellectual property protection to achieve the annotation. The experts need to annotate the clothes images in each set by identical, plagiarized, or irrelevant. If it is plagiarized clothes, they also need to label the modified region. The challenge in constructing the dataset is to mark out the clothes with minor variations in style from a large number of identical outfits.

5. Experiment

In order to verify the effectiveness of our proposed PS-Net for the plagiarized fashion task, we evaluate it on the Plagiarized Fashion dataset. Additionally, as mentioned that PS-Net can also be adapted to traditional fashion retrieval and landmark estimation tasks, so we also conduct experiments on both DeepFashion and DeepFashion2 datasets.

Implementations. Our proposed multi-task network requires training on two datasets: 1) learn landmark estimation and retrieval abilities on 13 categories of clothes in DeepFashion2 [14] dataset; 2) obtain “reasonable” region weights for plagiarized retrieval on four categories of clothes in Plagiarized Fashion dataset. The training is carried out in sequence and finally, combined to achieve the goal of plagiarized clothes retrieval. For the landmark branch, the initial learning rate is set as 0.001. It decreases at the 9th and 12th epochs with a factor of 0.1. The training is completed after 12 epochs. For the retrieval branch, the initial learning rate is set as 0.001 and decreases at the 61st and 71st epochs with a factor of 0.1. The training is completed after 80 epochs. Specifically, given a query, it takes approximately 0.75 seconds to retrieve images from the Plagiarized Fashion dataset. The performance is tested on a computer with 64G RAM and a GTX 1080TI GPU.

5.1. Plagiarized Clothes Retrieval

Experimental Setup. We conduct the plagiarized clothes retrieval on the Plagiarized Fashion dataset. We compare our approach with the traditional method without landmark guided regional attention, manual manipulation

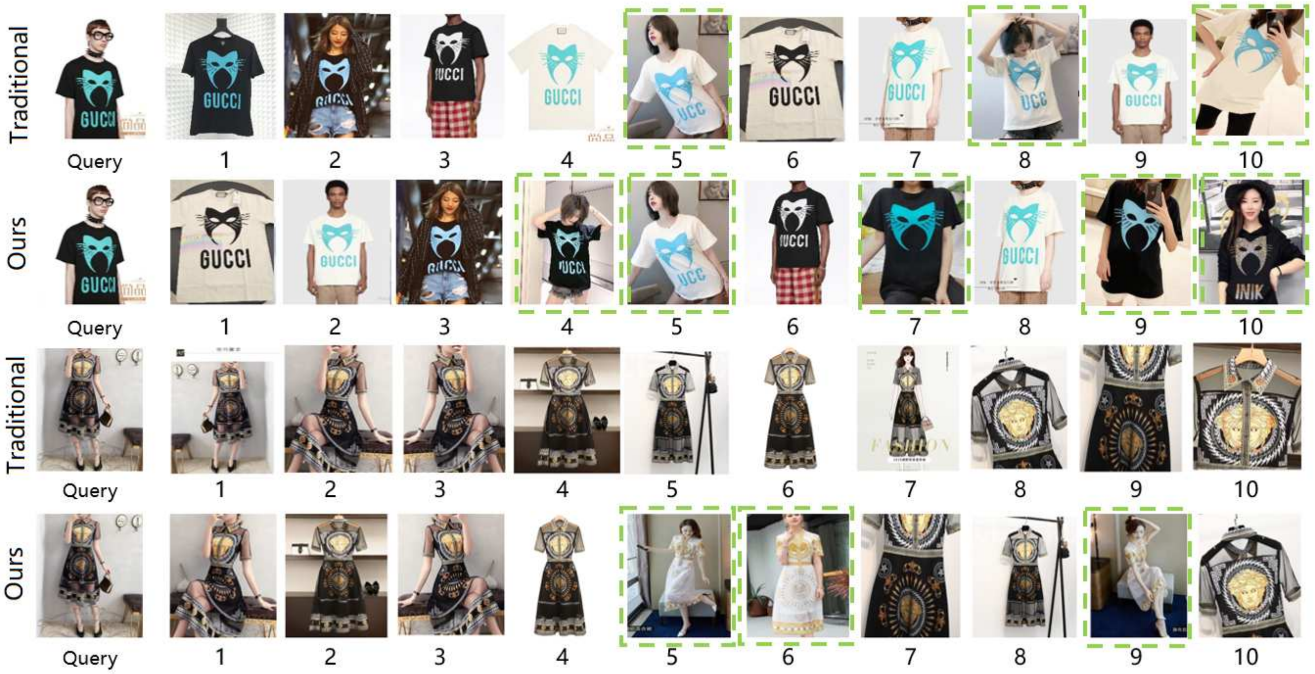

Figure 5. Example results of plagiarized clothes retrieval. The query is the clothes image with the original design, and the target recalls with green boxes are the plagiarized clothes in the gallery. For each query, the results of the traditional method are shown above, and our results are shown below.

	T-shirts			Outwears			Tops			Dress			Overall		
	Top-10	Top-20	mAP	Top-10	Top-20	mAP	Top-10	Top-20	mAP	Top-10	Top-20	mAP	Top-10	Top-20	mAP
PCB [37]	0.388	0.610	0.306	0.390	0.632	0.321	0.383	0.667	0.349	0.401	0.645	0.334	0.391	0.640	0.328
Traditional	0.395	0.622	0.313	0.398	0.640	0.325	0.390	0.672	0.353	0.406	0.650	0.338	0.397	0.645	0.332
Manual	0.564	0.803	0.465	0.532	0.772	0.401	0.556	0.793	0.451	0.532	0.768	0.429	0.542	0.783	0.443
Ours	**0.627**	**0.862**	**0.513**	**0.587**	**0.834**	**0.482**	**0.613**	**0.854**	**0.505**	**0.577**	**0.827**	**0.474**	**0.597**	**0.842**	**0.493**

Table 2. Quantitative results for plagiarized retrieval evaluated by Top-K recall and mAP. We compare our approach with the traditional method without landmark guided regional attention, manual manipulation method without learned region weights, and the PCB [37] method, which is widely used in near-duplicate retrieval task. The other settings of the model are identical.

method without learned region weights, and the PCB [37] method, which is widely used in near-duplicate retrieval task. For the traditional method, we set the five region weights as 1 by default. For the manual method, we collect the manual manipulation results from 25 participants, and use the average weight values to complete the plagiarized retrieval. Specifically, on the interactive interface we provide, the user can lower or raise the weight of each region by dragging the slider. The other settings of the model are identical (e.g., the backbone). The results of the three methods are evaluated by the metrics of Top-K recall and mAP.

Evaluation Resutls. Quantitative results for plagiarized clothes retrieval are shown in Table 2. The traditional retrieval method obtains the top-20 recall of 0.645 and an overall mAP of 0.332 on four categories of clothes, which is similar to the PCB [37] method. The manual method obtains over 10 percent improvement on recall rate and an overall mAP of 0.443, which is better than the traditional one. Then, we utilize the learned weights from the training to complete the retrieval. Our approach obtains 0.852 top-20 recall and an overall mAP of 0.493 on four categories of clothes, which improves the performance by a large margin, compared to other counterparts. Especially, for the categories of T-shirts and long-sleeves tops, our approach gets an mAP of 0.513 and 0.505, which are obviously higher than the manual method (0.465 & 0.451) and traditional method (0.313 & 0.353).

Results Visualization. Figure 5 shows two groups of plagiarized retrieval results of our approach and the traditional method. The images bounded by green boxes are correct recalls, which indicate the plagiarized clothes.

For the T-shirt and long-sleeved top categories, the tricks which commonly used by plagiarisers are replacing logo

Method	Top-10	Top-20	Top-30	mAP
No Attention	0.567	0.811	0.854	0.466
No Manipulation	0.413	0.682	0.743	0.361
None	0.397	0.645	0.698	0.332
Ours (HR-Net)	0.597	0.842	0.887	0.493
Ours (Ensemble)	**0.602**	**0.852**	**0.893**	**0.501**

Table 3. Quantitative results for the ablation study, evaluated by the Top-K recall rate and mAP.

text, adding mosaics or patterns and flipping clothes prints. Taking the first query as an example, after using the region manipulation, we successfully recall five plagiarized samples within the top-10 results. By contrast, the counterpart method only completes three plagiarized recalls within the top-10.

For the plagiarized sample of dresses, it usually not only has a small local modification but an imitation of the overall style. Therefore, the modification of this magnitude makes it difficult for traditional retrieval methods to complete the recall. For the second group of the query, the traditional method fails to recall any plagiarized clothes in the top-10 results. By manipulating the region weights, we can recall three plagiarized samples within the top-10 results.

The results show that our approach has significantly improved the ability of retrieval plagiarized clothes and alleviates the difficulty of recalling samples with partial modification. In conclusion, the region weights we learned through training are reasonable, and the region manipulation mechanism is effective for plagiarized clothes retrieval.

5.2. Ablation Study

Experimental Setup. We conduct an ablation study on the Plagiarized Fashion dataset. The factors we consider are: attention mechanism, region manipulation, and model ensemble. The results are evaluated by the Top-K recall rate and mAP.

Evaluation Results. The quantitative results of the ablation study are shown in table 3. The complete model of our approach achieved a 0.842 top-20 recall rate and a 0.493 mAP on the Plagiarized Fashion dataset. When we ensemble five models together (with different initial learning rates from 0.0005 to 0.01), the top-20 recall is increased to 0.852, and the mAP becomes 0.501. When the attention mechanism is removed from the complete model, it achieved a top-20 recall rate of 0.811 and the mAP drops to 0.466, which proves that the attention mechanism is vital for the retrieval task. To verify the effect of region manipulation, we adjust the learned region weights to the default ones. The top-20 recall rate drops significantly for more than 15 percent. Finally, we test the model without any component mentioned above, the top-20 recall rate drops about 20 percent on the Plagiarized Fashion dataset.

Method	Collar	Sleeve	Waist	Hem	Overall
FashionNet[27]	.0878	.0954	.0854	.0818	.0872
DFA[28]	.0633	.0640	.0714	.0661	.0660
DLAN[46]	.0591	.0660	.0699	.0626	.0643
BCRNN[39]	.0410	.0660	.0513	.0544	.0484
DAFE [9]	.0296	.0362	.0312	.0398	.0342
Ours	**.0293**	**.0358**	**.0310**	**.0396**	**.0339**

Table 4. Quantitative results for clothes landmark detection on the DeepFashion [27] dataset, evaluated by normalized error (NE). The best scores are marked in bold.

From the above comparison results, we can find that two essential designs of our approach: landmark guided regional attention and region manipulation are vital for plagiarized clothes retrieval. Moreover, the model ensemble is also beneficial.

5.3. Landmark Estimation

Experimental Setup. The landmark estimation experiments are conducted on both DeepFashion [27] and DeepFashion2 [14] datasets. DeepFashion is the most widely used fashion landmark dataset with 123,016 clothes images. Followed by previous work [39], we use the standard dataset split and the evaluation is performed on 40,000 images. DeepFashion2 is the most challenging fashion landmark dataset at present, which has a different number of landmarks for each type of clothes. It contains 491,895 clothes images, and the experiment is evaluated on 33,669 images. According to the previous methods[39, 14], the results on DeepFashion dataset are evaluated by normalized error (NE). The results of DeepFashion2 dataset are evaluated by Average Precision (AP).

Evaluation Results. Our approach obtains an average result of 0.0339 on DeepFashion dataset, as shown in Table 4, which is much better than previous methods like D-LAN [46] (0.0643), DFA [28] (0.0660), Fashion-Net [27] (0.0872), BCRNN [39] (0.0484) and the recent method with dual attention [9] (0.0342).

The results of DeepFashion2 dataset is shown in Table 5. Note that we conduct the experiment on both visible and occluded landmarks. Since the DeepFashion2 dataset is a new dataset with none comparison method, we compare our approach with the released networks of Simple-Baseline [41] and CPN [10] (two of the best methods for human landmark estimation). Our approach obtains an overall AP of 0.633, which greatly higher than the Simple-Baseline [41] (0.591), CPN [10] (0.579) and the Match-RCNN [14] (0.563).

In general, our approach obtains SOTA results in the landmark estimation task on two fashion datasets. It indicates that the design of the landmark branch is effective.

Method	Scale			Occlusion			Zoom-in			Viewpoint			Overall
	small	moderate	large	slight	medium	heavy	no	medium	large	no wear	frontal	side	
Match-RCNN[14]	0.497	0.607	0.555	0.643	0.530	0.248	0.616	0.489	0.319	0.510	0.596	0.456	0.563
CPN*[10]	0.512	0.619	0.560	0.663	0.542	0.261	0.625	0.501	0.330	0.523	0.621	0.468	0.579
Simple-Baseline*[41]	0.523	0.632	0.574	0.671	0.562	0.277	0.638	0.512	0.349	0.543	0.632	0.485	0.591
Ours	**0.581**	**0.682**	**0.633**	**0.713**	**0.606**	**0.332**	**0.691**	**0.567**	**0.408**	**0.592**	**0.679**	**0.533**	**0.633**

Table 5. Landmark estimation results on different subsets of DeepFashion2 [14], evaluated by Average Precision (AP). The best performance are marked in bold. * represents the results we achieved by the released network [10, 41].

5.4. Traditional Retrieval

Experimental Setup. The main goal of our work is to retrieve plagiarized clothes. On the other hand, we would like to demonstrate that the landmark guided regional attention can also enhance the performance of traditional retrieval task. Thus, we evaluate our approach with the same ResNet [19] backbone of previous methods on Deepfashion [27] and DeepFashion2 [14] datasets.

According to the previous methods [33, 14], Top-K recall on (a) DeepFashion [27] and Top-K accuracies on (b) DeepFashion2 [14] datasets are plotted in Figure 6. DeepFashion provides 52,712 clothes images for the retrieval within shops, while DeepFashion2 provides a Customer-to-Shops application scenario with 491,895 images. We conduct the evaluation on 26,830 and 33,669 images, respectively.

Evaluation Results. On the DeepFashion dataset, our approach achieves 0.889 recall rate at top-1, which is obviously better than the previous retrieval methods like DARN [21] (0.381), WTBI (0.347) [8], FashionNet [27] (0.533), and the various form of FashionNet. When retrieving the top-50 results, the recall rate of our approach has reached 0.991. Shin et al. [33] proposed a semi-supervised feature-level manipulation for fashion image retrieval. It achieves the same top-50 performance (0.991) compared to us, but its top-1 performance (0.887) is lower than ours.

Figure 6 (b) shows the retrieval results on DeepFashion2. Our approach obtains a top-10 result of 0.745, which is much higher than the result of PCB method [37] (0.703), achieving based on the released model. Compared to the best performance by Match-RCNN [14] (0.573), our performance obtains a huge improvement of 18 percent. The geometric distribution of the landmarks allows the model to know the structure of the clothes in any case, which is especially important for partial occluded or deformed clothes. By coincidence, these samples are often encountered in Customer-to-Shops scenario. Match R-CNN [14] method also trains the landmark branch and the retrieval branch simultaneously, but it only accumulates the loss functions without introducing any attention.

As can be seen from the result, introducing landmark guided regional attention during the similarity learning is very effective. In conclusion, our method has achieved state-of-the-art effects on both In-Shops and Customer-to-Shops search, which verifies the retrieval ability of our approach on different application backgrounds and image quality.

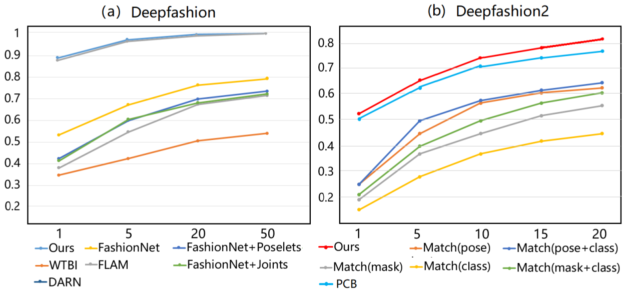

Figure 6. The results of the traditional clothes retrieval experiment. Top-K recall on (a) DeepFashion [27] and Top-K accuracies on (b) DeepFashion2 [14] datasets are plotted.

6. Conclusion

In this paper, we introduce a novel problem of plagiarized clothes retrieval for original design protection and provide a novel network named PS-Net with a dedicated Plagiarized Fashion dataset, which fills the gap in the field of fashion retrieval. We propose an attentive network based on regional representation and let the geometric information of landmarks guide the similarity learning, which outperforms the other SOTA counterparts. We design a region manipulation mechanism to solve the problem of plagiarized clothes retrieval. We learn the region weights for different categories of clothes on the proposed dataset, in order to manipulate the region weights automatically during similarity learning. By doing so, a plagiarized clothes image with a modified region can also be recalled, which has significant improvement compared to other methods. In general, our work can effectively alleviate the problem of plagiarized clothes retrieval and has great potential for original design protection.

Acknowledgement. This work was partly supported by the National Natural Science Foundation of China (No.61902347), the Zhejiang Provincial Natural Science Foundation (No.LQ19F020002) and the Alibaba-ZJU Joint Research Institute of Frontier Technologies.

References

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Fashionsearchnet: Fashion search with attribute manipulation. In *ECCV*, 2018. 2

[2] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, 2018. 1, 2

[3] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *WACV*, 2018. 5

[4] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Which shirt for my first date? towards a flexible attribute-based fashion query system. *Pattern Recognition Letters*, 2018. 1

[5] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *ICCV*, 2019. 2

[6] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 2

[7] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *ICCV Workshops*, 2019. 2

[8] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 8

[9] Ming Chen, Yingjie Qin, Lizhe Qi, and Yunquan Sun. Improving fashion landmark detection by dual attention feature enhancement. In *ICCV Workshops*, 2019. 7

[10] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 7, 8

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[12] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *CVPR Workshops*, 2013. 2

[13] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *CVPR*, 2019. 4

[14] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, 2019. 2, 3, 5, 7, 8

[15] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *AVSS*, 2018. 2

[16] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 1, 2

[17] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 2

[18] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *ACM MM*, 2017. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 8

[20] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W Zheng, and Qi Liu. Explainable fashion recommendation: A semantic attribute region guided approach. *arXiv preprint arXiv:1905.12862*, 2019. 1

[21] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 8

[22] Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa. Multi-label fashion image classification with minimal human supervision. In *ICCV*, 2017. 2

[23] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2

[24] Yining Lang, Yuan He, Jianfeng Dong, Fan Yang, and Hui Xue. Design-gan: Cross-category fashion translation driven by landmark attention. In *ICASSP*, 2020. 2

[25] Jingyuan Liu and Hong Lu. Deep fashion analysis with feature map upsampling and landmark-driven attention. In *ECCV*, 2018. 4

[26] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012. 1, 2

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1, 2, 5, 7, 8

[28] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2, 7

[29] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[30] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, 2017. 1

[31] Zhe Ma, Jianfeng Dong, Yao Zhang, Zhongzi Long, Yuan He, Hui Xue, and Shouling Ji. Fine-grained fashion similarity learning by attribute-specific embedding network. *arXiv preprint arXiv:2002.02814*, 2020. 1

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3, 4

[33] Minchul Shin, Sanghyuk Park, and Taeksoo Kim. Semi-supervised feature-level attribute manipulation for fashion image retrieval. *arXiv preprint arXiv:1907.05007*, 2019. 2, 8

[34] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, 2015. 1

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2, 3

[37] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 6, 8

[38] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *TIP*, 2017. 2

[39] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 1, 2, 7

[40] Xianwang Wang and Tong Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM MM*, 2011. 2

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 7, 8

[42] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 2

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[44] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 2

[45] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1, 2

[46] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *ACM MM*, 2017. 2, 7

[47] Yijun Yan, Jinchang Ren, Genyun Sun, Huimin Zhao, Junwei Han, Xuelong Li, Stephen Marshall, and Jin Zhan. Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognition*, 2018. 2

[48] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *TIP*, 2017. 4

[49] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, 2017. 1, 2

[50] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 4