Learning Attribute Representations with Localization for Flexible Fashion Search

Kenan E. Ak,[1,2] Ashraf A. Kassim[1], Joo Hwee Lim [2], and Jo Yew Tham[3]
[1]National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]ESP xMedia Pte. Ltd., Singapore
emir.ak@u.nus.edu, ashraf@nus.edu.sg, joohwee@i2r.a-star.edu.sg, thamjy@espxmedia.com

Abstract

In this paper, we investigate ways of conducting a detailed fashion search using query images and attributes. A credible fashion search platform should be able to (1) find images that share the same attributes as the query image, (2) allow users to manipulate certain attributes, e.g. replace collar attribute from round to v-neck, and (3) handle region-specific attribute manipulations, e.g. replacing the color attribute of the sleeve region without changing the color attribute of other regions. A key challenge to be addressed is that fashion products have multiple attributes and it is important for each of these attributes to have representative features. To address these challenges, we propose the FashionSearchNet which uses a weakly supervised localization method to extract regions of attributes. By doing so, unrelated features can be ignored thus improving the similarity learning. Also, FashionSearchNet incorporates a new procedure that enables region awareness to be able to handle region-specific requests. FashionSearchNet outperforms the most recent fashion search techniques and is shown to be able to carry out different search scenarios using the dynamic queries.

1. Introduction

Over the last few years, there has been remarkable progress in fashion related research, which includes recognition [14, 31, 36], attribute discovery [19, 42], recommendation [4, 20, 35], retrieval [10, 18, 36] and human/fashion parsing [29, 34, 43]. While the key issue addressed in mainstream image retrieval research is in cross-domain retrieval, a major challenge is managing situations where there are many attributes.

It is especially difficult to conduct a search by changing a certain attribute of the query image since it may be hard to find a balance between "maintaining the current attributes"

Figure 1: Given a query image, two different fashion search scenarios are conducted. Each attribute manipulation operation provides the desired attribute, that redefines the representation of the query image. The proposed FashionSearchNet then retrieves the most similar three images by incorporating the query image with the requested manipulations as shown.

and "adding a new attribute". In [46], this problem is defined as attribute manipulation and a solution is provided by combining features of the query image with a representation of the desired attribute. Another approach involves allowing users to decide which image is more preferred through "relevance feedback" but it can be computationally intensive [26, 28, 40]. In any case, these methods do not explore the localization aspect of feature representations which could help leaving out some artifacts of the unwanted attributes.

This paper introduces the FashionSearchNet which is able to conduct fashion searches using just query images and also query images with basic and region-specific attribute manipulations. Figure 1 illustrates two attribute manipulation scenarios for a given query image. For the first scenario, the color and collar attributes are changed to "beige" and "hood" respectively. The second scenario involves, a more specific attribute manipulation which changes the color of torso region to "red". While for the first attribute manipulation scenario resulted in suitable matches

by FashionSearchNet, this was not the case for the second example as there was no exact match in the dataset for the desired attributes.

FashionSearchNet's global average pooling layer gives it remarkable a localization ability [47], enabling it to generate attribute activation maps (AAMs) using weakly supervised labels i.e., with no information about the location of the attributes. Consequently, each attribute can be represented with the most relevant region. Focusing on regions is more feasible for images with multiple attributes since some attributes may only be present at certain locations. Region of interest pooling is performed to feed attribute-relevant features into a set of fully connected layers. The classification and ranking losses are applied to these layers to learn attribute representations/similarities. By doing so, the triplets of images ranking constraint becomes triplets of regions as each region is estimated from the AAMs. Therefore, the necessary conditions for picking triplets are relaxed which increases the possible number of triplets immensely as the similarity learning is now independent from the number of attributes.

After the training of attribute representations is finished, the undesired attribute representation of the query image can be directly replaced for a given attribute manipulation. Next, attribute representations of the query image are combined into a single global representation and used in the fashion search. This is made possible through the global ranking loss function, applied to decide which representation should have more importance depending on the attribute manipulation.

To our best knowledge, there are no other reported fashion search methods that are able to handle different regions of attributes (e.g. different color attributes for torso and sleeve regions). In order to conduct region-specific attribute manipulations using only image level annotations, FashionSearchNet incorporates a procedure that is able to discover relevant regions from the AAMs. These discovered regions then can be combined with the learned attribute representations (fully connected layers) so that new features can be included into the fashion search. For example, FashionSearchNet is made aware of say the color attribute of the sleeve regions through the AAMs corresponding to these regions of interest (ROI). Through ROI pooling of sleeve regions, the fully connected layer associated with the color attribute outputs a new set of features for region-specific fashion search. The same procedure can be used for any attribute-region combination to extract new region-specific features.

Here is a summary of the main contributions of this paper:

- Introduce a new network structure called FashionSearchNet which is able to conduct attribute representation learning by localizing towards attributes and acquire better attribute representations for attribute manipulations.
- Regions and region-specific attribute representations discovered using attribute activation maps which enable region-specific attribute manipulations.
- Experiments show that our proposed FashionSearchNet is up to 16% more accurate compared to baseline methods.

2. Related Work

Clothing Attribute Recognition. Clothing attributes provide a useful tool to assess the clothing products. Clothing attribute recognition has gained an increasing attention in the recent years. Preliminary works [6, 7, 25] relied on combining hand-crafted features such as SIFT [32] and HoG [12] with Support Vector Machines [11]. With deep neural networks, more accurate methods have been proposed. Mix and match [44] combined a deep network with conditional random fields to explore the compatibility of clothing items and attributes. Chen [8] focused on solving the problem of describing people based on fine-grained clothing attributes. Several works utilized weakly labeled image-text pairs to discover attributes [42, 45]. Abdulnabi et al. [2] proposed a multi-task based approach to learn an algorithm to predict multi-attributes. Attribute recognition helps both localization and retrieval abilities of our FashionSearchNet.

Attribute Localization. Being able to localize towards objects has been proven to be quite efficient in fine-grained recognition [23, 24] and image retrieval [5, 17]. In terms of fashion products, DARN [22] utilized from a module to detect clothing items which improved the performance of the network. Similar to DARN [22], Song et. al [38] proposed a unified method to localize and classify apparels. More interestingly, in FashionNet [31] joint prediction of clothing attributes and landmarks proved to be quite effective. However, most aforementioned methods require annotation of bounding boxes or key points and use localization to detect the main image. In [37], an end-to-end method is proposed to simultaneously localize and rank relative attributes in a weakly supervised manner with the help of Spatial Transformer Networks [24]. Nevertheless, the fact that a new model must be trained for each attribute for the method proposed in [37], makes it hard to implement for images with multiple attributes. Recently, class activation maps [47] has been shown to be very efficient in localizing on most representative regions using only the image level annotations. As it is not quite possible to annotate bounding boxes for every attribute, we were inspired to innovate by incorporating a weakly annotated attention mechanism to conduct multi-attribute based similarity learning.

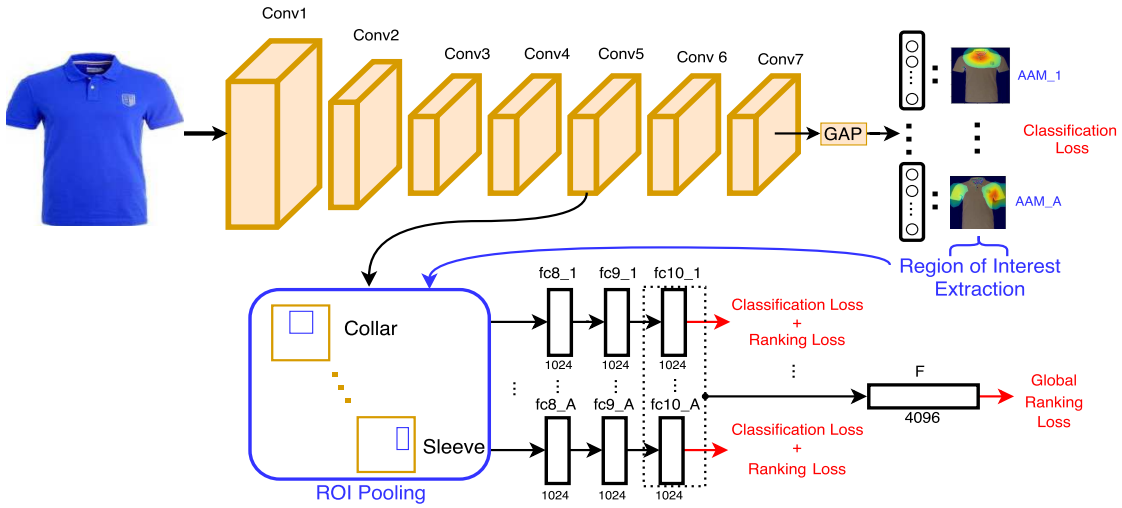


Figure 2: Overview of the FashionSearchNet. With the input image fed through the network, the GAP layer is used to generate attribute activation maps (AAMs) for each attribute which are used to estimate several regions of interests (ROIs). Next, attribute-specific features are pooled from the $conv5$ layer by using the estimated ROIs. The pooled features are linked to a set of fully connected layers where the similarity learning between attributes is conducted and attribute representations are produced ($fc_{10\_1}, fc_{10\_2}, ...$). Finally, attribute representations are combined into a global representation (F) for use in the fashion search.

Fashion Retrieval. Fashion retrieval can be grouped into several categories. The most popular involves searching for same/similar items from images [18, 31, 33, 36] or videos [9, 15], while another set of works investigate the problem of fashion recommendation [4, 30, 39, 41]. There has not been much work done on image retrieval with attribute manipulation except [19, 46]. Han et. al [19] focuses on spatially aware concept discovery from image/text pairs and conducts attribute-feedback product retrieval with the word2vec model. On the other hand, Zhao et. al [46] proposed a method to manipulate clothing products using a memory gate and their approach was called "fashion search with attribute manipulation". In contrast to [46], our proposed method follows a different approach by including the localization of attributes, focusing on attribute representation learning and enabling region-specific attribute manipulations.

3. FashionSearchNet

This section presents an overview of our FashionSearchNet. First, the network is trained with the classification loss to generate attribute activation maps (AAMs) [47] which are able to identify regions of attributes thus ignoring unrelated features. The network is then trained once more to learn attribute representations with a combination of the ranking and the classification losses. A weighted combination of the attribute representations into a global representation is adopted in our proposed network through the global ranking loss. Lastly, AAMs produce region-specific attribute representations that enable the FashionSearchNet to carry out region based analysis of fashion images.

3.1. Architecture Overview

The proposed FashionSearchNet network structure shown in Figure 2 is based on the AlexNet [27] architecture with the following modifications applied to the baseline network: All fully connected layers are removed and two convolutional layers are added after $conv5$ layer to compensate the effect of removing the fully connected layers. Regions of Interest (ROIs) are extracted using the AAMs, that represent the most activated regions of attributes. Features from $conv5$ layer are pooled using the ROIs into a set of fully connected layers, that serve as attribute representations and trained with the classification and ranking losses. These learned attribute representations are combined into a global representation to represent the input image with manipulated attributes, if any. The global ranking loss is required to estimate the importance of attribute representations depending on the attribute manipulation.

3.2. Learning Attribute Representations

Attribute Activation Maps: The classification loss is used to discover the most relevant regions of attributes. Initially, the global average pooling (GAP) layer is applied to the last convolutional layer which corresponds to $conv7$ as follows:

$$\sum_{k} \left(x_I(k) = \sum_{i,j} conv7_k(I,i,j) \right) \tag{1}$$

where $x_I(k)$ is the features extracted from the image I for channel k and $conv7_k(I,i,j)$ is the k-th feature map of $conv7'th$ layer at spatial location (i,j). The multi-attribute classification network is trained using the following classification loss:

$$L_C = -\sum_{I=1}^{N}\sum_{a=1}^{A} log(p(g_{Ia}|x_I w_a)) \tag{2}$$

where g_{Ia} represents the ground truth of the $a'th$ attribute of the $I'th$ image. $x_I w_a$[1] calculates weighted linear combination of x_I for attribute a, N is the number of training examples and A is the number of attributes. The posterior probability estimates the probability of $x_I w_a$ to be classified as g_{Ia}. We next define $M_{a_c}(I,i,j)$ as the attribute activation map (AAM) for class c of an attribute a as follows:

$$M_{a_c}(I,i,j) = \sum_k w_{a_{(k,c)}} conv7_k(I,i,j) \tag{3}$$

where $w_{a_{(k,c)}}$ is the weight variable of attribute a associated with $k'th$ feature map of class c and c is determined from the class, that maximizes the classification confidence. Using M_{a_c}, attribute localization can be added to the network. In order to do so, A number of maps are estimated with a simple hard threshold technique. As per the implementation in [47], the pixel values that are above 20% of the maximum value in the generated map are segmented. This is followed by estimating a bounding box, that covers the largest connected region in the AAM. This step is repeated for each attribute.

Ranking with Triplets of Regions: FashionSearchNet's ability to identify ROIs enables it to ignore regions with unrelated features which may confuse the attribute similarity learning capability of the network. A structure similar to ROI pooling layer [16] is used to pass features from the $conv5$ layer to a set of fully connected layers.

The example in Figure 3 for collar attribute similarity shows the intuition behind the triplets of regions ranking loss function. Anchor image $(\hat{I})$ may look similar to the negative image (I^-), due to color similarities. Actually, the collar attribute is very similar to that of I^+. If the output of the network's h'th layer without triplet ranking is used to check Euclidean distances, $||h(\hat{I}), h(I^+)||_2 > ||h(\hat{I}), h(I^-)||_2$ would be the situation, meaning $\hat{I}$ would be closer to I^- rather than I^+ in the feature space.

The first step of the proposed method involves estimating the corresponding AAMs as shown in Figure 3. Note that the heatmaps of $\hat{I}$ and I^+ cover a smaller area compared to I^- thus confirming the localization ability since the collar attribute of I^- covers a wider region. It is evident that as the AAMs localize towards the collar attribute,

[1]The dimensions of w_a is [number of feature maps by number of classes associated with a]

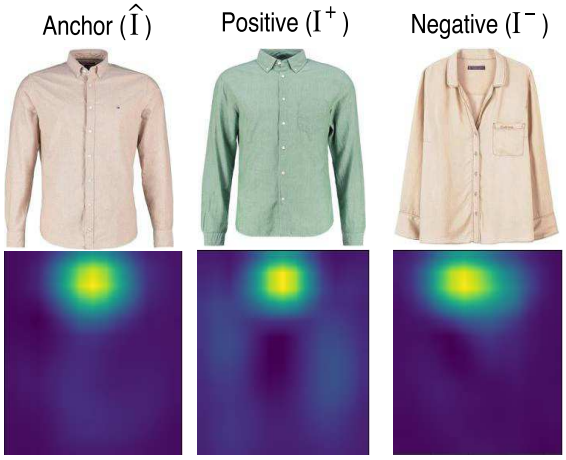


Figure 3: Examples for the triplets of regions of the collar attribute: Anchor $(\hat{I})$, Positive (I^+) and Negative (I^-). The generated collar attribute activation maps tend to be near the collar region and irrelevant regions are eliminated; thus, enabling a better attribute similarity learning.

the unrelated regions such as sleeve and torso are ignored without any intervention. Thus, FashionSearchNet is able to differentiate the collar attribute while ignoring irrelevant attributes (e.g., color, pattern etc.).

When the triplet ranking loss function defined in [22, 33] is used in FashionSearchNet, the observed loss was tremendous unless a very small learning rate is used. Inspired by [21], the soft-triplet ranking function is utilized which normalizes the distances to the range of (0,1) with the softmax function and formulated as follows:

$$d^+(h(\hat{I}),h(I^+),h(I^-)) = \frac{exp(||h(\hat{I})-h(I^+)||_2)}{exp(||h(\hat{I})-h(I^+)||_2)+exp(||h(\hat{I})-h(I^-)||_2)} \tag{4}$$

$$d^-(h(\hat{I}),h(I^+),h(I^-)) = \frac{exp(||h(\hat{I})-h(I^-)||_2)}{exp(||h(\hat{I})-h(I^+)||_2)+exp(||h(\hat{I})-h(I^-)||_2)} \tag{5}$$

Given $||d^+, d^- - 1||_2^2 = d^+$ and $h = fc_{10\_a}$ the ranking loss function can be written as:

$$L_T = \sum_{I=1}^{N}\sum_{a=1}^{A} d^+(fc_{10\_a}(\hat{I}), fc_{10\_a}(I^+), fc_{10\_a}(I^-)) \tag{6}$$

where A is the number of the fully connected layers which is also the number of attributes. The role of Eq. 6 is to learn a representation for each attribute using the final set of fully connected layers: fc_{10_a}. We minimize $||fc_{10\_a}(\hat{I}), fc_{10\_a}(I^+)||_2$ and maximize $||fc_{10\_a}(\hat{I}), fc_{10\_a}(I^-)||$. The rule for picking triplets is quite simple, $\hat{I}$ and I^+ must share the same label while I^- is chosen randomly from a different label.

Both ranking and classification losses are used in the optimization processes leading to the attribute representations. It is necessary to use the classification loss as it was observed from experiments that using only the ranking loss significantly diminishes the discriminative ability of the network. This classification loss denoted as L_{TC} is formulated as in Eq. 2, except $x_I w_a$ is replaced with the output of fc_{10_a} layers.

Attribute Manipulation: By associating each attribute with a different fully connected layer ($fc_{10\_1}, ..., fc_{10\_a}$), the fashion search with attribute manipulation becomes straightforward. After the training, features are extracted from the training images with the same attribute value and averaged. These averaged features can be used to replace the "undesired" attribute representations i.e., attribute manipulation (a*).

3.3. Learning Global Representation

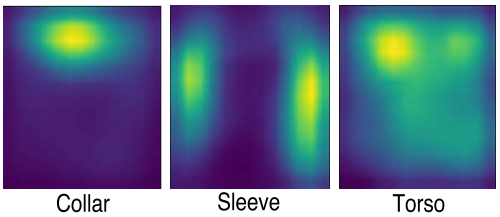
In the previous subsection, we showed how FashionSearchNet is taught to localize and learn attribute representations. Combining all these learned features would lead to better results when conducting image searches. However, such combinations may be too big to handle and thus slow down the search process. To address this issue, a weight parameter w_{a^*} is applied to reduce the concatenated feature length to 4096. Moreover, using an additional ranking function and letting a weight variable learn "which features are more important" improves the performance vastly. This is because some attribute representations such as "color" could be more important than other attribute representations say, "fabric". A+1 weight parameters denoted by λ_{a,a^*} are learned, where the '+1' weight is the one without attribute manipulation. The training is conducted with the following global ranking loss function L_G using $F(I, a^*)$ as follows for a given attribute manipulation a^*:

$$F(I, a^*) = [fc_{10\_1}(I)\lambda_{1,a^*}, ..., fc_{10\_A}(I)\lambda_{a,a^*}]w_{a^*} \quad (7)$$

$$L_G = \sum_{I=1}^{N} \sum_{a^*=1}^{A+1} d^+(F(\hat{I}, a^*), F(I^+, a^*), F(I^-, a^*)) \quad (8)$$

For the training of global representations, the unwanted attribute representation of the query image is replaced with the desired one which corresponds to $F(\hat{I}, a^*)$. With the weight variables applied as shown in Eq. 7, the training is conducted with the loss function given in Eq. 8. The same procedure is applied to all attributes. The rule for picking triplets for the global ranking loss is that $\hat{I}$ and I^+ must be identical in terms of attributes after the attribute manipulation a^* while I^- is chosen randomly.

Optimization: FashionSearchNet uses different loss functions in its optimization steps. It is first trained using L_c as the other processes depend on how reliable the AAMs are. Then, L_C, L_T, L_{TC} losses are accumulated and the network is trained once again. Finally, the parameters of the global representation are learned using L_G without changing any attribute representation.



Region	Attribute
Collar	Collar, Neckline
Torso	Pattern, Color
Sleeve	Sleeve length

Figure 4: The mean of the discovered collar, sleeve and torso regions for images in Shopping100k dataset [3], using the linkage provided in the table.

3.4. FashionSearchNet with region awareness

The weakly annotated structure of fashion datasets limits the ability to conduct detailed fashion searches. Consider an example of a dress image where the ground truth for the color attribute is "red". This may mean that the most dominant color is "red" while the color attribute for the sleeve and collar regions are "white". To address this issue, FashionSearchNet is extended to incorporate a procedure that facilitates "region awareness".

Intuitively, it is possible to define which attribute corresponds to which region. For example, in the Shopping100k dataset [3], the Collar, Sleeve and Torso regions can be defined. The region awareness procedure combines AAMs based on the linkages between regions and attributes.

For an attribute, the classes that have classification confidences higher than 0.1 are chosen to be relevant classes. Next, each AAM from the relevant classes is weighted by their associated classification confidences and accumulated. AAMs from different attributes are combined by accumulating heatmaps (see Figure 4). Once again, the pixel values that are above 20% of the maximum value in the generated map are segmented enabling the "awareness" of three different regions. These discovered regions can be associated with any attribute representation and thus enable new region-specific features to be extracted. The means of the discovered regions for Shopping100k dataset [3] images are shown in Figure 4. As can be seen, the sleeve and collar maps are visually better positioned than the torso map. To resolve this issue, the highlighted regions of the sleeve and collar are removed from the torso region.

4. Experiments

Datasets: Of the many datasets available for fashion research [3, 18, 22, 25, 31], we used Shopping100k [3] and DeepFashion [31] which provide real-world fashion images along with their detailed attributes. The images in the DeepFashion [31] dataset have people, posing with the clothing products while Shopping100k [3] only has images of clothing items.

Shopping100k dataset [3] contains 101,021 images with 12 attributes. 79,000 images are used to train the network and the remaining 22,021 images are used for the fashion search. 20,021 images are reserved for the retrieval gallery, leaving 2,000 images for the queries.

For the DeepFashion dataset [31], we use category and attribute prediction benchmark which has 289,222 images with 6 attributes. Looking at the attributes, Texture, Shape and Category attributes are used as they are the most suitable for the attribute manipulation experiments. By removing unrelated attributes, the dataset is reduced to around 113,000 images. 90,000 images are used to train the network and the remaining 23,000 images are used to conduct the fashion search. For the retrieval gallery, 21,000 images are reserved leaving 2,000 images for the queries.

Competing Methods: We investigated several state-of-the-art approaches and found that FashionNet [31] and StyleNet [36] which solve different fashion retrieval problems and are not suitable for attribute manipulation. We compared the performance of FashionSearchNet with AMNet [46] and an attribute-based method which uses AlexNet [27] and predicts attributes of query images and substitutes unwanted attributes with the desired ones. For the comparison study, we used a variation of FashionSearchNet: FashionSearchNet without the localization ability AAMs (i.e., no AAMs) which feeds $conv5$ features without ROI pooling. FashionSearchNet with region awareness is used for the user study.

Evaluation Metric: For the qualitative experiments, we define the retrieval accuracy as follows. Given a query image and an attribute manipulation, the search algorithm finds the "best K" image matches i.e., "Top-K" matches. If there is a match (i.e., having the same attributes as the query image) after attribute manipulation, it corresponds to a hit (1) otherwise it is a miss (0).

Search Strategy: In [13], the search involves the use of very precise terms resulting in short search sessions leading to an end-result. We also adopt simple fashion queries by only changing a single attribute from the query image.

4.1. Search by Query

"*Search by Query*" experiments essentially involve finding similar images to the query image. These experiments require extracting features of the query image and comparing with the retrieval gallery without any attribute manipulations. Attribute manipulation by AMNet [46] is "turned off" by setting the attribute manipulation indicator with zeros and $fc7$ features of the attribute-based model is directly used to conduct the search.

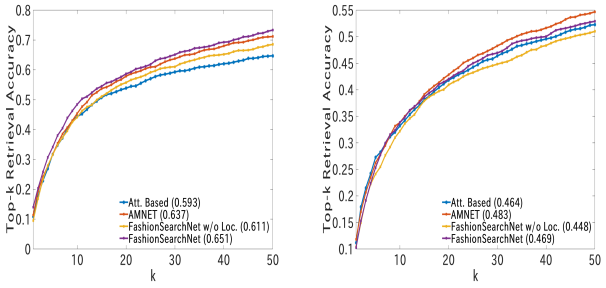
From the Top-K retrieval accuracy results presented in Figure 5 for Shopping100k and DeepFashion datasets, it is evident that all methods perform quite similar to each other. FashionSearchNet achieves the best performance on Shopping100k while AMNet [46] performs slightly better on DeepFashion dataset. While "*search by query*" is the easiest fashion search scenario, we have also carried out more challenging experiments by including attribute manipulations.

4.2. Search by Query and Attribute Manipulation

"*Search by query and attribute manipulations*" experiments involve replacing certain feature representations of the query image with the desired and comparing Euclidean distances of global representations with the retrieval gallery. For these experiments, every possible attribute manipulation that is available in the retrieval gallery is applied to the query images.

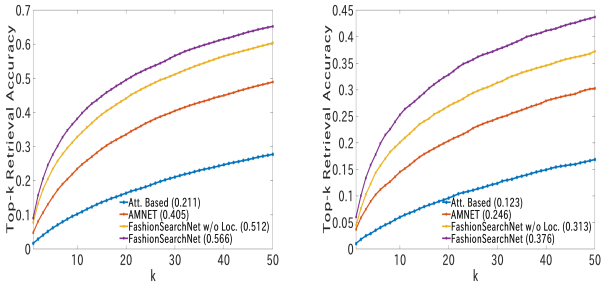
Top-K retrieval accuracy results are presented in Figure 6 for Shopping100k and DeepFashion datasets. FashionSearchNet achieves the best performance, giving 56.6% and 37.6% Top-30 accuracy on Shopping100k and DeepFashion datasets respectively which is 5.4% and 6.3% higher than the nearest competitor which is the FashionSearchNet without the localization ability. The performance of FashionSearchNet without localization shows that using the whole feature map to learn attribute representations may have introduced some noisy features that harm the retrieval accuracy. FashionSearchNet performs 16% and 13% better than AMNet [46]. AMNet [46] is a good baseline as it outperforms the attribute-based method by 19% and 12%. The attribute-based method does not perform well as attribute manipulation depends on having decent predictions for all attributes. The main drawback of AMNet [46] is that it combines the query image with a representation of the desired attribute, while FashionSearchNet replaces the undesired attribute with the desired one and estimates a global representation using a weighted combination. Also, AMNet [46] is unable to carry any localization which affects its performance.

Compared to the "*Search by Query*" experiment, there is a significant reduction in performance accuracy for all methods which highlights the difficulty of this task. With regards to the datasets, DeepFashion has more complex images, unlike Shopping100k. However, Shopping100k has a larger number of attributes thus making it a challenging benchmark for the attribute manipulation problem.



(a) Shopping100k **(b)** DeepFashion

Figure 5: Top-K Retrieval Accuracies for *search by query* experiments using (a) Shopping100k and (b) DeepFashion datasets. The number in the parentheses is the Top-30 retrieval accuracy.



(a) Shopping100k **(b)** DeepFashion

Figure 6: Top-K Retrieval Accuracies for *search by query and attribute manipulation* experiments using (a) Shopping100k and (b) DeepFashion datasets. The number in the parentheses is the Top-30 retrieval accuracy.

4.3. Visual Examples of Fashion Search

Figure 7 shows several examples of the fashion search experiments. Images with the green bounding box mean that all attributes are matched with the query image and attribute manipulation. In the first row, a query image is selected and fashion search is conducted without changing any attributes which is called "*search by query*". As a result, images similar to the query image are retrieved by FashionSearchNet.

Next, an attribute manipulation involving the change of the collar attribute to "High" is applied to the query image. FashionSearchNet uses the provided attribute manipulation and query image to retrieve the Top-3 images that meet the criteria. The second row of Figure 7 shows another example with the same procedure and more attribute manipulation experiments are provided in the supplementary material.

4.4. Search by Region-specific Attribute Manipulations

As none of the datasets contain region-based attributes and it would be challenging to label each region for the whole dataset, a user study on the Shopping100K dataset was conducted to overcome this limitation. Note that, for these set of experiments lower weights were assigned to other attribute representations such as collar, pocket etc. to give more importance to the desired attributes.

User Study: This user study consists of 10 volunteers and 30 query images along the region-specific color attribute manipulations (e.g. change sleeve region from red to green color) and leaves other attributes for the future studies. Each user was given a query image along with the attribute manipulation similar to Figure 8 and shown the Top-5 retrieved images by AMNet [46], FashionSearchNet and FashionSearchNet with region-awareness. The users were asked to choose which of the retrieved images satisfy the search criteria and the success rate was evaluated as follows: (number of picked images) / (total number of displayed images). The results of this user study are presented are present the results in Table 1. For FashionSearchNet and AMNet [46], as it is not possible to conduct region-specific attribute manipulations, the attribute manipulation performed is similar to "*Search by Query and Attribute Manipulation*". Our FashionSearchNet with region awareness performs much better than the other two methods by utilizing additional region-specific representations. FashionSearchNet works better than AMNet [46] because it can attend to color regions to find several relevant images.

Table 1: Success rates for region-specific attribute manipulation experiments.

	AMNet [46]	FashionSearchNet	FashionSearchNet with region awareness
Success Rate	1.2%	8.6%	22.2%

Of the three examples in Figure 8, the color of the torso attribute is changed to the yellow color in the first example. After conducting this region-specific attribute manipulation, the first retrieved image is the one where the color of the torso is yellow. Two other examples given in Figure 8 focus on manipulation of attributes to the sleeve and collar regions.

4.5. Implementation Details

We use pre-trained ImageNet weights for AlexNet [27] up until $conv4'th$ layer and reinitialize other layers. For $conv5$, $conv6$, $conv7$ layers, sizes of feature maps are 384, 512, 512 respectively. As we use regions of triplets ranking constraint, the selection of $(\hat{I}, I^+, I^-)$ becomes easy. For each mini-batch, images with the same attribute are chosen to be $\hat{I}$ and I^+. I^- is picked such that it has a different attribute. Gradients are calculated for each loss function and they are back propagated through the network. The network is trained with the stochastic gradient descent algorithm us-

Figure 7: The Top-3 retrieval results of fashion search with or without attribute manipulation for a given a query image. The green bounding boxes denote those images retrieved by FashionSearchNet that match all desired attributes when the search is conducted after the attribute manipulations are applied.

Figure 8: Examples of region-specific attribute manipulations. For a given query image and region-specific attribute manipulation (change), the image retrieved by FashionSearchNet that best correlates with the attributes of the query image and attribute manipulation is shown on the right of the query image.

ing the learning rate of 0.01. No pre-processing steps other than removing the means from each channel of images were conducted. For DeepFashion dataset, as there are only 3 attributes, the sizes of the fully connected layers are increased to 2048. For the region of interest pooling, Tensorflow's [1] "tf.image.crop_and_resize" function is used to feed the $conv5$ features with the estimated bounding boxes into a set of fully connected layers. Dropout was used on all fully connected layers with $p = 0.5$.

4.6. Run Time Performance

Our FashionSearchNet is trained on Intel i7-5820K CPU and 64 GB RAM memory with GeForce GTX TITAN X GPU. FashionSearchNet can extract features from 10,000 images in around 60 seconds which is very close to the attribute-based method. Compared to the AlexNet implementation [27], the proposed FashionSearchNet has several additional layers to learn attribute representations. However, by using smaller fully connected layers, the run-time performance of FashionSearchNet is still efficient compared to the attribute-based method. Moreover, using ROI pooling layer, all images just need to be fed into the network once which saves a lot of computation time. Extraction of ROIs for all attributes is efficient as it takes about only 0.002 seconds for each image.

5. Conclusion

This paper presents a new approach for conducting fashion searches using just query images and also query images with basic and region-specific attribute manipulations. The proposed FashionSearchNet is able to generate efficient feature representations for fashion search and its good localization ability enables it to identify the most relevant regions for attributes of interest. In addition to being able to combine attribute representations into a single global representation for attribute manipulations, FashionSearchNet incorporates a procedure that facilitates "region awareness" to accommodate region-specific requests. The proposed FashionSearchNet is shown to outperform the baseline fashion search methods including AMNet [46]. An interesting problem for future work would be to extend FashionSearchNet to other types of image retrieval problems.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. 8

[2] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959, 2015. 2

[3] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 1671–1679. IEEE, 2018. 5, 6

[4] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. *arXiv preprint arXiv:1705.06394*, 2017. 1, 3

[5] S. Bell and K. Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 34(4):98, 2015. 2

[6] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Asian conference on computer vision (ACCV)*, pages 321–335, 2012. 2

[7] H. Chen, A. Gallagher, and B. Girod. Describing Clothing by Semantic Attributes. In *European Conference on Computer Vision (ECCV)*, 2012. 2

[8] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, volume 7, pages 5315–5324, 2015. 2

[9] Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4048–4056, 2017. 3

[10] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. *arXiv preprint arXiv:1709.09426*, 2017. 1

[11] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 2

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 886–893, 2005. 2

[13] R. Datta, D. Joshi, J. I. A. Li, and J. Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys (Csur)* , 40(2):1–60, 2008. 6

[14] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 520–529, 2017. 1

[15] N. Garcia and G. Vogiatzis. Dress like a star: Retrieving fashion products from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2293–2299, 2017. 3

[16] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[17] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer, 2016. 2

[18] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3343–3351, 2015. 1, 3, 6

[19] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1463–1471, 2017. 1, 3

[20] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. *arXiv preprint arXiv:1707.05691*, 2017. 1

[21] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 4

[22] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1062–1070, 2015. 2, 4, 6

[23] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182, 2016. 2

[24] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2

[25] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, pages 472–488, 2014. 2, 6

[26] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch : Image Search with Relative Attribute Feedback. 2012. 1

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 6, 7, 8

[28] H. Li, M. Toyoura, K. Shimizu, W. Yang, and X. Mao. Retrieval of clothing images based on relevance feedback with focus on collar designs. *Visual Computer*, 32(10):1351–1363, 2016. 1

[29] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1386–1394, 2015. 1

[30] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM international conference on Multimedia*, pages 619–628, 2012. 3

[31] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 1, 2, 3, 6

[32] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[33] D. Shankar, S. Narumanchi, H. Ananya, P. Kompalli, and K. Chaudhury. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344*, 2017. 3, 4

[34] E. Simó Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A high performance crf model for clothes parsing. In *Computer Vision-ACCV 2014, Vol 9005 of Lecture Notes in Computer Science*, pages 64–81, 2014. 1

[35] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2015. 1

[36] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 298–307, 2016. 1, 3, 6

[37] K. K. Singh and Y. J. Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016. 2

[38] Y. Song, Y. Li, B. Wu, C.-Y. Chen, X. Zhang, and H. Adam. Learning unified embedding for apparel recognition. *arXiv preprint arXiv:1707.05929*, 2017. 2

[39] P. Tangseng, K. Yamaguchi, and T. Okatani. Recommending outfits from personal closet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2279, 2017. 3

[40] S. Tong and E. Chang. Support Vector Machine Active Learning for Image Retrieval. (C), 2000. 1

[41] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015. 3

[42] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi. Automatic Attribute Discovery with Neural Activations. *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[43] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3519–3526, 2013. 1

[44] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi. Mix and match: Joint model for clothing and attribute recognition. In *BMVC*, pages 51–1, 2015. 2

[45] T. Yashima, N. Okazaki, K. Inui, K. Yamaguchi, and T. Okatani. Learning to describe e-commerce images from noisy online data. In *Asian Conference on Computer Vision*, pages 85–100. Springer, 2016. 2

[46] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *IEEE Conference Computer Vision Pattern Recognition (CVPR)*, 2017. 1, 3, 6, 7, 8

[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 3, 4