Texture and Shape biased Two-Stream Networks for Clothing Classification and Attribute Recognition

Yuwei Zhang[1], Peng Zhang[1], Chun Yuan [2,3*], and Zhi Wang [2,3*]

[1]Department of Computer Science and Technology, Tsinghua University
[2]Tsinghua Shenzhen International Graduate School
[3]Peng Cheng Laboratory

Abstract

Clothes category classification and attribute recognition have achieved distinguished success with the development of deep learning. People have found that landmark detection plays a positive role in these tasks. However, little research is committed to analyzing these tasks from the perspective of clothing attributes. In our work, we explore the usefulness of landmarks and find that landmarks can assist in extracting shape features; and using landmarks for joint learning can increase classification and recognition accuracy effectively. We also find that texture features have an impelling effect on these tasks and that the pre-trained ImageNet model has good performance in extracting texture features. To this end, we propose to use two streams to enhance the extraction of shape and texture, respectively. In particular, this paper proposes a simple implementation, Texture and Shape biased Fashion Networks (TS-FashionNet). Comprehensive and rich experiments demonstrate our discoveries and the effectiveness of our model. We improve the top-3 classification accuracy by 0.83% and improve the top-3 attribute recognition recall rate by 1.39% compared to the state-of-the-art models.

1. Introduction

Nowadays, fashion image analysis has a rapid expansion in both academy and industry, with its growing application in e-commerce and online shopping. Many studies are committed to clothes recognition ([25], [1], [32]), retrieval ([9], [8], [18]), recommendation ([15], [19]) and fashion trend prediction ([1], [28]). The development of efficient deep learning methods and availability of large-scale rich-annotated fashion dataset are also stupendous impetus to the progress of these works. All these make various fashion works that seemed impossible to achieve come true. Among them, we target at *clothing classification* and *attribute recognition*, which are fundemental components for other tasks and can generally bring improvements to them (*e.g.*, clothing recommendation).

Previous works are aware of the importance of landmark detection to the clothes classification and attribute recognition and have achieved particular success. For example, Liu *et al*. [25] annotated the fashion dataset with landmark information and proposed a model that could predict the landmarks and attributes simultaneously; Wang *et al*. [32] introduced Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) for the use of the fashion grammars (*i.e.*, kinematics grammar and symmetry grammar) to detect landmarks and used landmark-aware attention and category-driven attention to enhance clothing recognition. However, these studies have neglected the importance of joint texture and shape features in fashion classification and recognition.

Though the idea of using both texture and shape features is straightforward, it is challenging when we incorporate them into real-world tasks, including fashion classification and recognition. In our experiments, we have tried to combine the two features by adding a branch to an existing model, which learns jointly with the landmark on an ImageNet pre-trained single-stream network and fine-tune it on the DeepFashion-C [25] dataset. However, the experimental results show that the accuracy of attribute recognition has not been improved. It is still challenging to design proper network architecture to integrate texture and shape features.

To make full use of these two features, we use measurement studies to find the factors that affect the accuracies when integrating the features into different models. Based on our measurement insights, we design our Texture and Shape biased Two-Stream Networks that use joint texture

* Corresponding authors. ({yuanc,wangzhi}@sz.tsinghua.edu.cn).

and shape features for fashion classification and attribute recognition. The contributions of this paper are summarized as follows.

▷ We carry out measurement studies and analysis to highlight the usefulness of texture features and shape features in fashion tasks, including clothing classification and attribute recognition. On the one hand, motivated by the recent study [12] that ImageNet [7] pre-trained CNNs are biased towards recognizing textures than shapes, we further discover that ImageNet pre-trained models can specifically promote the recognition of texture-related clothing attributes. On the other hand, we find that the detection of clothing landmarks helps the model to learn the shape characteristics of the garment. Jointly using the two features is of importance for fashion tasks.

▷ Based on our insights, we propose a two-stream architecture to better combine the advantages of the texture and shape features for the clothes classification and attributes recognition tasks. To tackle the ineffectiveness of the naive branch scheme above, we explore to extract and make use of the texture and shape features separately. In particular, we propose a two-stream structure with a texture-biased stream that is fine-tuned from an ImageNet pre-trained model, and a shape-biased stream derived from landmark features.

▷ Based on our design, we provide a simple implementation, Texture and Shape biased Fashion networks. Comprehensive experiments and evaluations demonstrate that our model outperforms the state-of-the-art, and our insights provided are valid. Mainly, TS-FashionNet improves the top-3 classification accuracy by 0.83% and improves the top-3 attribute recognition recall rate by 1.39% compared to the state-of-the-art models.

2. Related Work

Fahion Image Understanding

Deep learning based models have achieved great success in fashion field, such as clothes classification ([4], [25], [17]) and attribute recognition ([14], [1], [3], [32], [11]), fashion items recommendation ([15], [19], [30]) and clothes retireval ([33], [13], [24], [34], [9], [8], [18], [25], [21]). Earlier works used traditional image analysis methods (e.g. SIFT [27], HOG [6]) to extract fashion image features for the follow-up work, which are hard to grasp the most useful features of fashion images.

With the development of deep learning methods and the growth of the large-scale rich-annotated fashion datasets ([25], [11], [37], [39], [40]), fashion models have achieved prodigious success. They use convolutional neural networks to extract image features and process the images and obtain significant improvement in performance. In 2016, Liu *et al*. [25] introduced a large-scale fashion dataset with comprehensive annotations DeepFashion and proposed a deep model FashionNet to learn the clothing features by jointly predicting clothes category and landmark localization. In 2018, Wang *et al*. [32] introduced a deep grammar model, Bidirectional Convolutional Recurrent Neural Networks (BCRNNs), with two attention mechanisms for fashion landmark detection and clothing category classification. However, the previous work is rarely carried out from the perspective of the clothing attribute itself, and the accuracy of the recognition is rarely improved via analyzing the characteristics of various clothing attributes. In this paper, we analyze the characteristics of various attributes from the perspective of features and give corresponding optimization methods.

Landmark localization

The accurate locations and rich amounts of landmarks in the fashion images can be a good assistance to fashion tasks such as clothes attribute recognition and clothing retrieval. In the early works, they used bounding box ([13], [5]) to help the fashion tasks. A finer-level annotation landmark is desired than the existing bounding-box annotation. Recently many researches ([26], [23], [35], [36], [38], [32]) are dedicated to landmark detection. Liu *et al*. ([25]) used a branch network to predict the landmark location and visibility. Wang *et al*. [32] leveraged the high-level human knowledge of landmarks and proposed two important fashion grammars, dependency grammar capturing kinematics-like relation and symmetry grammar accounting for the bilateral symmetry of clothes. Yu *et al*. ([38]) proposed a general Layout-Graph Reasoning (LGR) layer and enforced structural layout relationships among landmarks for fashion landmark detection. In this paper, we analyze the method of landmark detection that can promote the recognition of clothing attributes, and analyze its practical effect on attribute recognition from the perspective of attributes.

ImageNet-trained CNNs are Biased Towards Texture

Previous works ([20], [22], [29], [16]) tended to think that it is the representation of shapes that counted for the impressive performance on complex understanding tasks (*e.g*. object recognition). On the other hand, some research ([10], [2], [12]) found the important role of object textures for CNNs recognition tasks. In 2015, Gatys *et al*. [10] found that texture representations based on CNNs could increasingly capture the statistical characteristics of images and optimize for object recognition. In 2016, Ballester *et al*. ([2]) showed that classic CNNs were unable to recognize sketches where textures are missing and shapes are left. More recently, Geirhos *et al*. ([12]) discovered and validated that ImageNet-trained CNNs were biased towards recognizing textures than shapes and increasing shape bias could be a benefit to increase the classification accuracy and robustness. In our work, we use the clothing attributes to verify the validity of this conclusion, and use the ImageNet pre-trained model to enhance texture feature learning and extraction.

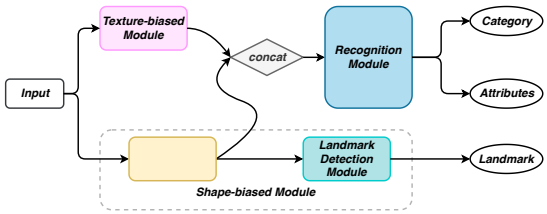


Figure 1. Texture and Shape biased Two-Stream Networks Illustration

3. Our Method

3.1. Overview

Attributes of DeepFashion-C [25] dataset are divided into five groups. Among them, the 'texture' and 'fabric' attributes are primarily determined by the texture features, while the 'shape' and 'part' attributes are primarily determined by the shape features. For the 'style' attributes, the pattern on the clothing is the basis for judging the style of the clothing. So, 'style' attributes are determined by the combination of color and texture features. If these corresponding features can be extracted and learned in a better way, they will bring a massive promotion to clothes classification and attribute recognition.

In our work, we propose an effective method to leverage texture and shape features. We propose Texture and Shape biased Two-Stream Networks (Fig. 1): one stream is **texture-biased** stream, and the other is **shape-biased** stream. For the shape-biased stream, we use a landmark branch to help extract shape features; for the texture-biased stream, we use ImageNet pre-trained model to emphasize on the extraction of texture features. Then we concatenate the features extracted by the two streams together to predict the clothing attributes and classify the clothes categories.

Moreover, we give a simple implementation of this method, TS-FashionNet (Fig. 3).

3.2. Shape-Biased Stream

We use joint learning with landmarks to enhance the model's understanding of shape features.

The Role of Landmark Information. Previous work [25, 32] has concluded that the rational use of landmark information can effectively improve the accuracy of clothing attribute recognition. We have found that the detection of fashion landmarks can improve the accuracy of the 'shape' attribute and 'part' attribute recognition (Sec. 4.3) through experiments. Then we speculate that localizations of landmarks can enhance the extraction and learning of shape features. Moreover, we also find that joint learning with fashion landmarks plays the main role in achieving such an effect. (Sec. 4.4). Therefore, we choose to jointly learn attributes and landmarks as the method to enhance shape feature extraction.

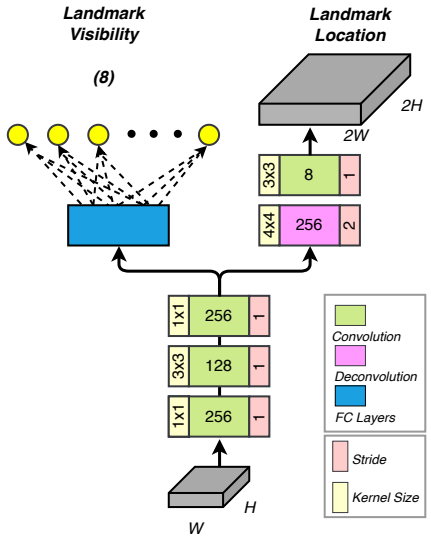


Figure 2. Landmark Processing Module

Fashion Landmark Detection. Fashion landmark detection aims to predict the positions of K landmarks corresponding to the clothing. Following [32], we transform this problem to achieve the prediction of K heatmaps, which denotes the confidence in the corresponding landmark position. The ground truth of the heatmaps is obtained by adding Gaussian filters at the ground-truth locations.

There are two approaches to handle invisible landmarks. The first one is to set the ground truth of the invisible landmarks heatmap as an all-zero heatmap. The other is to predict the visibility and location of landmarks separately, and only backpropagate the loss of the visible landmarks when predicting the locations. We have adopted the second way.

Joint Learning with Landmark. To jointly learn attributes and landmarks, we design a landmark branch. The structure of the landmark branch is shown in Fig. 2. This module will learn the visibility information of the landmarks prior to the location information of the landmarks. Visibility information is represented by the output of a fully connected layer with K neurons. The ground truth of visibility is a vector of length K. The value of the element in the vector is 0 or 1. 0 means that the landmark is invisible and 1 means that the landmark is visible. We adopt the sigmoid cross-entropy loss as the loss of landmark visibility, denoted as $L_{visibility}$, to discriminate if the landmark is visible. When learning location information, we use the ground truth of visibility to weigh the squared error to en-

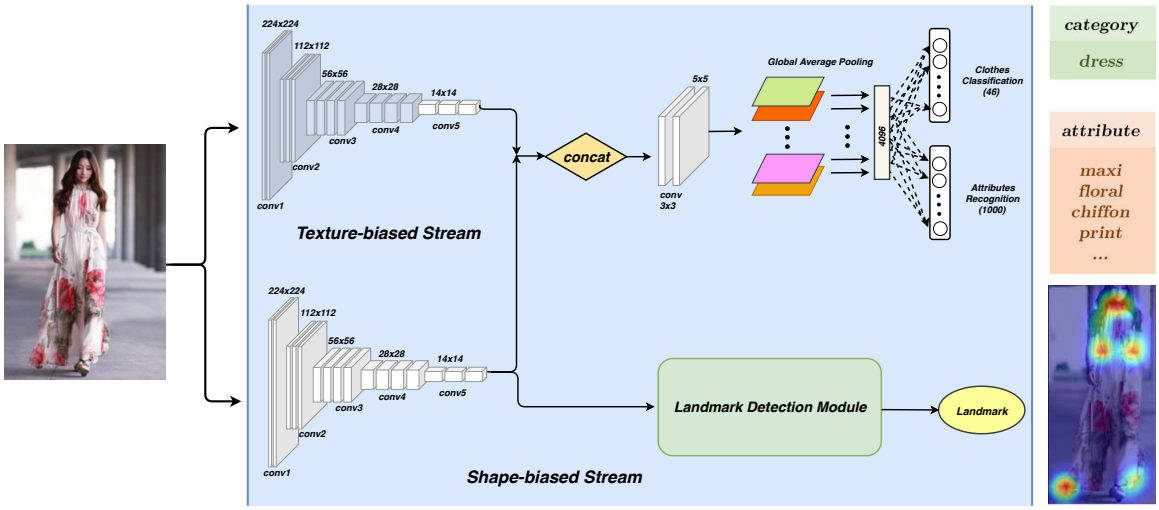


Figure 3. Texture and Shape biased Fashion Networks (TS-FashionNet) Architecture

sure that only the loss of the visible point is backpropagated:

$$L_{landmark} = \sum_{k=1}^{K} v_k^{GT} \sum_{x,y} ||S_k(x,y) - S_k^{GT}(x,y)||_2 \quad (1)$$

where S_k^{GT} is the ground truth of the landmark heatmap, S_k is the predicted heatmap, and v_k^{GT} is the ground truth of the k-th point visibility.

3.3. Texture-Biased Stream

Since clothing is the product of fashion design, the texture of the garment contains quite complex and high-level semantic information. The recognition accuracy of some attributes (stripe, print, graphic) of the clothing is also highly dependent on the understanding of the texture features. Robert Geirhos *et al*. [12] found that the ImageNet classification task was biased towards texture. We also demonstrate that the ImageNet pre-trained model focuses more on the texture features in the clothing attribute recognition task through experiments. (Sec. 4.3). In our work, we use the ImageNet pre-trained model as our texture-biased branch in virtue of the rich semantic information in ImageNet.

3.4. Two Streams Integration

We have tried two methods to integrate the shape-biased stream and texture-biased stream. The first method is to share the weights of the two streams. It is to use the pre-trained weights of ImageNet in the front part of the network and freeze them. Then we add a landmark branch to the deeper part of the network to jointly learn the attributes and landmarks. The second method is that the weights of the two streams are not shared. It is to concatenate the feature maps of the two streams together before entering the discriminative module.

We have tried these two methods separately, and find that the method of not sharing weights is better (Sec. 4.5). This is also in line with our assumption: the ability to specifically enhance the neural network's extraction and learning of texture features and shape features, respectively, can help to improve the accuracy of attribute recognition.

3.5. Network Structure

To demonstrate our method, we design TS-FashionNet. The architecture is shown in Fig. 3. Following the baseline of [25, 32], we use VGG16 [31] as our backbone. In the shape-biased stream, we use the output of the layer $conv5$-3 as the input of the landmark branch. The size of predicted heatmaps is $28 \times 28 \times K$. For texture-biased models, we freeze the model weights till $conv4$-3. Then we concatenate the output of the $pool5$ layer of the two streams together. The size of the concatenated features is $7 \times 7 \times 1024$.

In order to make the model converge efficiently, we replace two fully connected layers of VGG16 $fc6$ and $fc7$ with two convolution layers: $fc6$-$conv$ and $fc7$-$conv$. The convolution kernel size of both convolutional layers is 3×3. The padding method of $fc6$-$conv$ is 'valid', the number of channels is 2048, and the $fc7$-$conv$ is 'same' and 4096. A dropout layer with a probability of 0.5 is then added to prevent overfitting. We also replace the last fully connected layer $fc8$ with two fully connected layers to jointly predict the categories and attributes. We adopt 1-of-K softmax loss to classify categories and sigmoid cross-entropy loss to recognize attributes, denoted as $L_{category}$ and $L_{attribute}$.

	Category		Texture		Fabric		Shape		Part		Style		All	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
WTBI [4]	43.73	66.26	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
DARN [18]	59.48	79.58	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.17	66.11	71.36	42.35	51.95
FashionNet [25]	82.58	90.17	37.46	49.52	39.30	**49.84**	39.47	48.59	**44.13**	54.02	66.43	73.16	45.52	54.61
Wang *et al*. [32]	**90.99**	**95.78**	**50.31**	**65.48**	**40.31**	48.23	**53.32**	**61.05**	40.65	**56.32**	**68.70**	**74.25**	**51.53**	**60.95**
re-impl of [32]	91.16	95.86	56.48	65.85	44.10	54.40	61.30	70.30	49.24	59.36	33.58	42.44	49.19	58.80
ours	**91.99**	**96.44**	**58.52**	**68.19**	**46.44**	**57.02**	**61.86**	**70.81**	**49.82**	**60.36**	**34.40**	**43.44**	**50.58**	**60.43**

Table 1. Quantitative results for category classification and attribute prediction on the DeepFashion-C [25] dataset. Since the dataset is different, we have marked the best scores for the two cases.

4. Experiments

4.1. DeepFashion-C Dataset

DeepFashion-C [25] dataset is a universal dataset for clothing classification and attribute recognition, collected by Liu *et al*. [25] in 2016. DeepFashion-C dataset divides the clothes into 50 fine-grained categories, and 46 of them have corresponding images. Each image in this dataset is extensively labeled with 1,000 attributes, 8 landmarks, and a bounding box of target clothing. The attributes are split into five groups, characterizing 'texture', 'fabric', 'shape', 'part', and 'style', respectively. There are 289,222 images in this dataset, of which 209,222 are used for training, 40,000 are used for verification, and the remaining 40,000 are test samples.

In our experiments, we follow the split of training data and test data in DeepFashion-C. Following [32], we crop the image using the bounding box given in the dataset and scale it to 224×224. For images that are not square after cropping, we use black to complete it as a square. Horizontal flipping is the only form we use for data augmentation. For category classification, we employ the standard top-k classification accuracy as evaluation metric. For attribute prediction, our measuring criteria is the top-k recall rate following [25].

4.2. Comparison with the state-of-the-arts

Implementation Details. Our model is shown in Fig. 3. The texture stream loads the weights of ImageNet pre-trained VGG16 till conv4-3 and freezes it. The optimizer is Adam, and the batch size is 16. We first pre-train the shape-biased stream with clothing landmarks for 3 epochs. The learning rate is 1e-4 and the weights of $L_{visibility}$ and $L_{landmark}$ are both 1. Then we train our entire model on all tasks for 12 epochs. The learning rate of the first 6 epochs is 1e-4, and the learning rate of the last 6 epochs is 1e-5. The weights of $L_{category}$, $L_{attribute}$, $L_{visibility}$ and $L_{landmark}$ are 1:500:1:1.

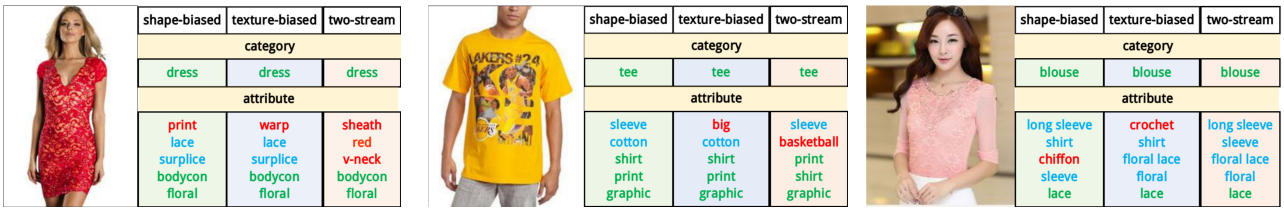
Performance Evaluation. We compare our model to Wang *et al*. [32]. As shown in the Table 1, [32] is the state-of-the-art network structure for clothing classification and attribute recognition. Nevertheless, in our experiments for its re-implementation, the experimental results of attribute recognition are quite different from theirs. We suspect that it may be because they used extra data related to fashion (mentioned in [32]). Another possible reason is that some annotations of attributes in DeepFashion-C have been changed later. So for reliability and reality, we compare our results with the re-implementation results on the same dataset, only DeepFashion-C dataset. The experimental results in Table 1 show that our method has improvement in all indicators. The classification accuracies of top-3 and top-5 are increased by 0.83% and 0.56% respectively. The top-3 recall rate and top-5 recall rate of attribute recognition are improved respectively by 1.39% and 1.63%. For each group of attributes, our method performs much better on 'texture' and 'fabric' (Δtop-3 $> 2.0\%$) while improves a little bit on 'shape' and 'part' (Δtop-3 $< 0.6\%$). The big promotion of 'texture' and 'fabric' is caused by that our proposed texture-biased stream can enable the network to learn the texture feature better. And the small promotion of 'shape' and 'part' is because that landmark detection has also been used in [32].

4.3. Pre-training and Joint Learning

We analyze the results of finetuning the ImageNet pre-trained model and joint training with landmark from scratch. Since the finetuning on the pre-trained model will result in faster convergence of the training, the number of iterations of training in these two methods is different. We select the best models that performed on the validation set and test them.

The backbone adopts VGG16 [31] architecture, and the designs of 'fc6', 'fc7' and 'fc8' follows Sec. 3.5. For joint learning, we use the output of the layer conv5-3 as the input of the landmark branch. For the pre-trained model, we freeze the model weights till conv4-3.

The results in Table 2 show that the ImageNet pre-trained model performs better on 'texture', 'fabric', and 'style' attribute recognition tasks, while the landmark joint learning model performs better on 'shape' and 'part' attribute recognition tasks. This means that the ImageNet pre-trained model is more concerned with texture features, which is consistent with the conclusions of Robert Geirhos *et al*.



(a) The comparison of the results of different bias methods: only shape-biased stream, only texture-biased stream and two streams. The 'attribute' shows the top-5 list of all the predicted attributes. The attributes predicted by the three methods are marked with green, attributes predicted by only one methods marked with red, and the left marked with blue.

(b) The result of the two stream network. The 'attribute' shows the top-5 list of all attributes predicted by the network. If ground truth has the corresponding attribute, it is shown in green, and the contrary is in red.

Figure 4. The results of clothing classification and attribute recognition.

	Category		Texture		Fabric		Shape		Part		Style	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
baseline	89.94	95.05	55.73	65.27	43.32	53.34	60.03	69.24	47.07	57.32	32.68	41.43
joint learning	**91.54**	**96.13**	56.05	65.51	44.16	54.67	**61.70**	**70.66**	**50.07**	**60.51**	32.88	42.26
pre-trained	90.82	95.70	**58.34**	**67.81**	**46.12**	**56.53**	60.41	69.58	48.24	58.60	**34.83**	**43.31**

Table 2. Results of the joint learning model and the pre-trained model.

[12]. And joint learning with landmark can make the model focus more on the shape features.

4.4. Analysis for Landmark-Aware Attention

Wang el al. [32] used an attention mechanism that combined landmark information. For the predicted heatmaps $\{S_i\}_{i=1}^{K}$, they averaged these heatmaps to get a weighed map A^L:

$$A^L = \frac{1}{K}\sum_{k=1}^{K} S_k \tag{2}$$

Then they added the A^L-weighed feature map of conv4-3 as a residual to the original network:

$$G = (1 + A^L) \circ F \tag{3}$$

where F denotes feature map, G denotes refined feature map and $\circ$ denotes the Hadamard product.

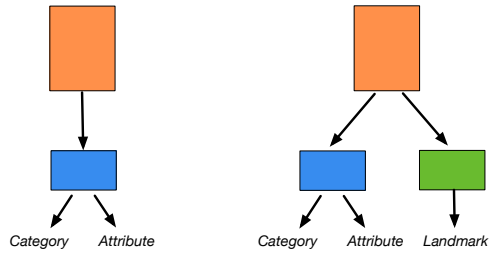
We also try a similar approach. Since we have respectively predicted visibility and position, our heatmap are weighed by the visibility vector:

$$A^L = \frac{1}{K}\sum_{k=1}^{K} v_k S_k \tag{4}$$

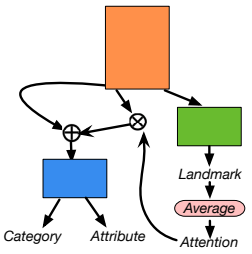
where v_k is the network's prediction of the visibility of the k-th point, which can be understood as the probability that the point is visible. We add the landmark branch after the layer conv4-3. To ensure that the input and output size of the landmark branch are the same, we remove the deconvolution layer from the landmark branch.

However, during the experiment, we find that such an attention mechanism does not effectively improve the performance of the network. It can be seen from the results in Table 3 that the performance of the network has been significantly improved after the addition of the landmark branch, but it seems to have no effect after adding the attention on the basis of landmark branch.

We initially think that this is due to the inaccurate prediction of the landmark location, which causes the model not to focus on the features near the landmark correctly. So we evaluate the precision of landmarks predicted by the model. Following [32], we adopt normalized error (NE) metric for evaluation. We find that the results after the evaluation are only a little worse than BCRNNs (NE: 0.0547 vs 0.0484). This is not a discrepancy that can make a big difference. We also visualize A^L, as shown in Fig. 6. It can be perceived that the landmark branch has a strong judgment on the vis-

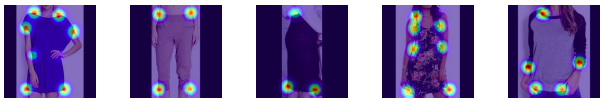


(a) Without joint learning and attention (b) Joint learning only.

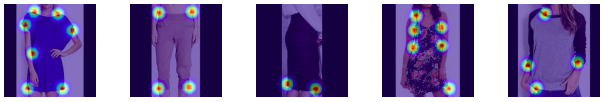


(c) Joint learning and attention

Figure 5. Three network structures for exploring the landmark attention mechanism.



(a) Landmark-Aware Attention



(b) Ground truth (only visible points)

Figure 6. Visualization of attention mechanisms and ground truth.

ibility and location of the landmark. The position of the attention mechanism is also correct. So it can be speculated that not the attention mechanism fails to detect the correct landmarks, but it is because attention does not work in this experiment.

Therefore, we surmise that the process of joint learning of landmarks has played a role in attention. Based on this inference, we choose to not use the attention mechanism in our two-stream network, and add the landmark branch after the conv5-3 layer, so that joint learning can be used in a better way.

4.5. Comparison of Single-Stream and Two-Stream

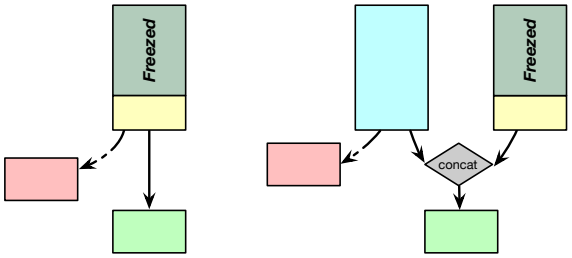
Firstly, we use a single-stream network to combine the enhancement of the two features. That is to load the weights of the ImageNet pre-trained model till conv4-3 and freezes them, and add the landmark branch after conv5-3. Experiments show that this method damages recognition performance. Afterward, we adopt another architecture: the two-stream networks we mentioned above. We do experiments for more in-depth analysis with these two methods and compare the results of extensive experiments using each enhancement (Sec. 4.3).

method	Category		Attributes	
	top-3	top-5	top-3	top-5
baseline	89.94	95.05	48.08	57.73
+joint(conv4-3)	**91.59**	96.11	**49.13**	58.81
+ attention	91.56	**96.11**	49.10	**58.90**

Table 3. Results of the three landmark processing methods.

method	Category		Attributes	
	top-3	top-5	top-3	top-5
joint learning	91.54	96.13	49.23	59.05
pre-trained	90.82	95.70	49.97	59.65
single-stream	91.24	95.86	48.77	58.70
two-stream	**91.99**	**96.44**	**50.58**	**60.43**

Table 4. Results of the single-stream model and the two-stream model.



(a) Single-stream network (b) Two-stream network

Figure 7. Illustration of single-stream network and two-stream network.

The results are shown in Table 4. Experimental results show that the method of adding two enhancements to a single-stream network performs less effectively than using only one enhancement. The top-3 attribute recall rate of the single-stream network is fewer than only shape-biased stream by 0.46%, and it is also less than texture-biased stream by 1.20%. Conversely, our two-stream network ourperforms the above three methods. Therefore, we conclude that the enhancement of the extraction of shape features and texture features is more conducive to the full use of the two features. Based on the results of such experiments, we finally adopt the two-stream network.

4.6. Further Analysis (Category and Attribute)

We also explore the relationship between clothing classification and clothing attribute recognition. The results are

	Category		Texture		Fabric		Shape		Part		Style	
	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5	top-3	top-5
category only	87.64	93.74	-	-	-	-	-	-	-	-	-	-
attribute only	-	-	55.66	65.25	42.02	52.25	58.50	67.59	45.07	55.46	31.56	40.47
category+attribute	**89.94**	**95.05**	**55.73**	**65.27**	**43.32**	**53.34**	**60.03**	**69.24**	**47.07**	**57.32**	**32.68**	**41.43**

Table 5. Results of learning category only, learning attributes only and jointly learning category and attributes.

method	Inshop		Consumer-to-Shop	
	top-30	top-50	top-20	top-30
BCRNN[32]	74.91	83.86	67.07	76.51
baseline	75.29	84.13	65.29	74.99
joint learning	75.33	84.15	68.42	77.85
pre-trained	**79.04**	**87.13**	69.33	78.66
Ours	78.45	86.69	**70.40**	**79.71**

Table 6. Results on clothes retrieval datasets.

shown in Table 5. Joint learning of category and attribute can improve the accuracy of both tasks. However, the classification task promotes the recognition of various attributes in different degrees. The classification task has little effect on the attribute recognition of the 'texture' group (only 0.07% higher for top-3 recall rate), but has more considerable influence on the 'shape' group (1.53% higher for top-3 recall rate) and the 'part' group (2.00% higher for top-3 recall rate). Moreover, according to the results in Table 2, the shape-biased stream performs significantly better on the task of clothing classification than the texture-biased stream. Therefore, we conclude that clothing classification is more dependent on the shape features. Enhancing the extraction and understanding of shape features is more helpful in improving the accuracy of clothing classification. Correspondingly, clothing classification can also promote the understanding of shape features.

4.7. Results on Clothes Retrieval Datasets

Experimental Setup. In-shop Clothes Retrieval and Consumer-to-Shop Clothes Retrieval are two of the benchmarks out of DeepFashion. In-shop Clothes Retrieval contains 52,712 images of 7,982 clothing items. And each item has 19.28 attributes on average. Consumer-to-Shop Clothes Retrieval contains 239,557 consumer/shop clothes images of 33,881 items and each item has 11.28 attributes on average. For these two datasets, we choose half of images of every item as training data and one image of every item as validation data at random. The remained are the testing data. We use top-30 and top-50 recall rate as measuring criteria of attribute recognition for In-shop Clothes Retrieval, and top-20 and top-30 recall rate for Consumer-to-Shop Clothes Retrieval.

Performance Evaluation. The results are shown in Table 6. We find that joint learning with landmark and pre-training on ImageNet both bring improvement to the recognition of Consumer-to-Shop dataset. While for In-shop dataset, landmark plays little role and sometimes the performance may exist negative transfer.

This may be caused by that: The images of In-shop dataset have no complicated background and the clothes have nearly none complex shape transforming so that the shape feature can be extracted easily. When the training data is not much, the landmark branch can help extract the shape feature, but when training data increases a lot, the baseline is enough to learn this simple shape feature. In contrast, for the Consumer-to-Shop dataset of which the shape features are hard to learn, landmark promotes the learning of shape features effectively and brings more improvement.

5. Conclusion

In this paper, we point out that compared to enhancing the ability to extract and learn shape features and texture features together, targeted enhancements are more helpful in improving the accuracy of attribute recognition. At the same time, we find that joint learning with landmarks helps extract shape features, while ImageNet pre-train process helps extract texture features. We also find that joint learning with landmarks is a major factor in promoting shape attribute recognition compared to attention mechanisms, and that clothing classification is more dependent on the shape features of clothing. Based on such knowledge and findings, we propose a network that can individually enhance shape and texture features: Texture and Shape biased Fashion Networks (TS-FashionNet). We demonstrate our model in two basic tasks (clothing classification and attributes recognition) on Deepfashion-C dataset, and the performances both surpass the previous SOTA structure.

6. Acknowledgements

This work was supported in part by NSFC under Grant No. 61872215, SZSTI under Grant No. JCYJ20180306174057899. This work was also supported by NSFC project Grant No. U1833101, Shenzhen Science and Technologies project under Grant No. JCYJ20190809172201639 and the Joint Research Center of Tencent and Tsinghua. Besides, we also would thank InfleXion Lab and LETOTE China for sponsoring this research.

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, pages 388–397, 2017.

[2] P. Ballester and R. M. Araujo. On the performance of googlenet and alexnet applied to sketches. In *AAAI*, 2016.

[3] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *ACCV*, pages 321–335. Springer, 2012.

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623. Springer, 2012.

[5] 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 *CVPR*, pages 5315–5324, 2015.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[8] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *CVPR workshops*, pages 8–13, 2013.

[9] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *ACCV*, pages 420–431. Springer, 2012.

[10] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, pages 262–270, 2015.

[11] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, pages 5337–5345, 2019.

[12] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[13] 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 *ICCV*, pages 3343–3351, 2015.

[14] 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 *ICCV*, pages 1463–1471, 2017.

[15] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *ACM MM*, pages 1078–1086. ACM, 2017.

[16] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran. Assessing shape bias property of convolutional neural networks. In *CVPR Workshops*, pages 1923–1931, 2018.

[17] W.-L. Hsiao and K. Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images (supplementary material).

[18] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, pages 1062–1070, 2015.

[19] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, pages 472–488. Springer, 2014.

[20] N. Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[21] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *ICCV*, pages 3066–3075, 2019.

[22] J. Kubilius, S. Bracci, and H. P. O. de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[23] X. Liang, K. Gong, X. Shen, and L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

[24] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337. IEEE, 2012.

[25] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

[26] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *ECCV*, pages 229–245. Springer, 2016.

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

[28] U. Mall, K. Matzen, B. Hariharan, N. Snavely, and K. Bala. Geostyle: Discovering fashion trends and events. In *ICCV*, pages 411–420, 2019.

[29] S. Ritter, D. G. Barrett, A. Santoro, and M. M. Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *ICML*, pages 2940–2949. JMLR. org, 2017.

[30] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of beauty. In *CVPR*, 2015.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv: Computer Vision and Pattern Recognition*, 2014.

[32] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, pages 4271–4280, 2018.

[33] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM MM*, pages 1353–1356. ACM, 2011.

[34] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, pages 3519–3526, 2013.

[35] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *ACM MM*, pages 172–180. ACM, 2017.

[36] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *ICCV*, pages 1281–1290, 2017.

[37] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, pages 3182–3189, 2014.

[38] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao, and L. Lin. Layout-graph reasoning for fashion landmark detection. In *CVPR*, pages 2937–2945, 2019.

[39] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *ACM MM*, pages 1670–1678. ACM, 2018.

[40] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao. Fashionai: A hierarchical dataset for fashion understanding. In *CVPR Workshops*, pages 0–0, 2019.