The Visual Computer
https://doi.org/10.1007/s00371-020-01885-7

ORIGINAL ARTICLE

# An improved landmark-driven and spatial–channel attentive convolutional neural network for fashion clothes classification

**Majuran Shajini**[1] · **Amirthalingam Ramanan**[1]



**Abstract**
Fashion clothes classification encompasses spotting and identifying items of clothing in an image. This area of research has involved using deep neural networks to make an impact in the field of social media, e-commerce and fashion world. In this paper, we propose an attention-driven technique for tackling visual fashion clothes analysis in images, aiming to achieve clothing category classification and attribute prediction by producing regularised landmark layouts. For enhancing clothing classification, our fashion model incorporates two attention pipelines: landmark-driven attention and spatial–channel attention. These attention pipelines allow our model to represent multiscale contextual information of landmarks, thus improving the efficiency of classification by identifying the important features and locating where they exist in an input image. We evaluated the proposed network on two large-scale benchmark datasets: DeepFashion-C and fashion landmark detection (FLD). Experimental results show that the proposed architecture involving deep neural network outperforms other recently reported state-of-the-art techniques in the classification of fashion clothes.

**Keywords** Dilated convolutions · Fashion clothes classification · Landmark-driven · Spatial–channel attention

## 1 Introduction

Visual fashion clothes analysis has generated a lot of interest due to its wide spectrum of human-centric applications. Fashion clothes classification relies on the detection of clothing area using bounding box prediction, analysing important landmarks to distinguish various clothing categories, and describing information about clothing based on its attributes. Earlier fashion models mostly relied on handcrafted features and search for powerful clothing depictions such as graph models, contextual information, general object proposals, human parts, bounding boxes, and semantic masks. At present, an evolving technique for clothing-related tasks is the deep learning method, which uses deep convolutional neural networks to simultaneously learn feature illustration and classify clothing-related images. Deep learning methods demonstrate the potential of automatically learning effective image representations to complete clothing image classification tasks (Fig. 1). More recently, with the introduction of large-scale fashion datasets, deep learning-based models achieved astonishing success in this area [1–5].

A real-time clothing classification system facilitates automated fashion stylists, outfit recommendation, discovering similar fashion pieces, surveillance context, automatic annotation of images with tags or descriptions, context-aided people identification, occupation recognition, and improvement in information retrieval from various areas such as social medias and consumer sites. Digital analysis in clothing images inherently endures difficulties based on varieties in texture, style and cutting, deformation and occlusion of subjects, and different scenarios where the images are taken from such as selfies and online shopping images. Discerning by overcoming these obstacles is a highly challenging task in clothing classification. Thus, a robust solution should make use of scale invariance along with both spatial and contextual feature extractors. Moreover, the fashion clothes classification integrates landmark localisation, attribute prediction, and category classification for multiple clothing items into formulated compact multitask learning. This is intended to allow knowledge sharing while solving multiple clothing-related tasks simultaneously. It has been demonstrated that

✉ Majuran Shajini
shayu.kiri@gmail.com

Amirthalingam Ramanan
a.ramanan@univ.jfn.ac.lk

[1] Department of Computer Science, Faculty of Science, University of Jaffna, Jaffna, Sri Lanka

Published online: 29 June 2020

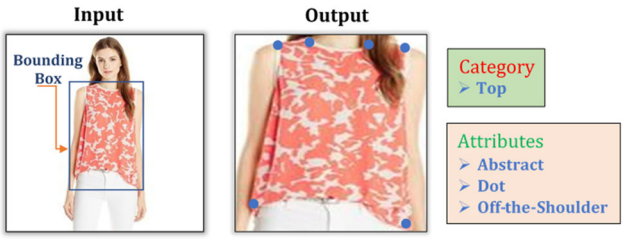


**Fig. 1** Overall problem specification of the fashion clothes classification and its attribute prediction

this sharing can boost the performance of such tasks through finding some common feature structures shared by all tasks or mining the related features.

We note that contextual information increases the performance of localisation by combining multiscale extracted features. Different scales have different influences on the input images. Nowadays, many approaches use multiscale information by employing several pooling and subsampling layers within convolutional neural networks which reduce the resolution of input. In contrast, the work in this paper is motivated by the fact that the module enhanced by dilated convolutions provide contextual reasoning without loss of resolution and also can fit into any architecture. As clothing landmarks fertilise the classification of clothing items, we employ stacked dilated convolutional (SDC) blocks accompanying upsampling mechanism in order to produce localised heatmaps.

In classification tasks, many visual attention mechanisms are used to improve the performance. Humans have an ability to focus and understand the novel visual features in difficult scenarios which is called "visual attention". Conventionally, it is important to perform some fine-grained visual processing on images and even multiple steps of inference to produce high-quality output. In clothing-related tasks, landmarks are well-formed attentive layouts for distinguishing different clothing items. Furthermore, the process of convolutional neural networks incorporates characteristics along spatial regions which are object-dependent features and along channels which replicates the various representation of global image features (e.g. edges, colours, corners, etc.). Various clothing categories and wide-ranging attributes can effectively be identified through integrating and probing those concepts as the attentive representation. Based on that, we adopt the spatial- and channel-wise attentions through convolutional neural network in order to empower feature extraction. Our main contribution in the proposed method is twofold:

1. To develop a scale-driven structure that utilises parallel convolutions with different scale variants through deep neural network followed with upsampling convolutions for fashion clothes landmark localisation.
2. To present attribute prediction along with clothes classification as multitask network accompanying two attentions: landmark-driven and spatial–channel-wise attentions which effectively learn features of items of clothing.

The rest of this paper is structured as follows: Sect. 2 summarises recently carried out works in the topics of fashion landmark detection, fashion clothes classification, and attribute prediction. Section 3 details the stacked dilated convolutional block and spatial–channel attention block-based approaches proposed in our methodology. Section 4 provides details of the experimental set-up and implementation along with the results. Finally, Sect. 5 concludes this work.

## 2 Related work

### 2.1 Fashion landmark detection

Localising effectual regions known as landmarks such as the collar, sleeves, and the waistline in clothes can be an important attention in clothing classification. Landmark localisation involves global integration of information and also the ability to retain local pixel-level details. It adjoins more challenges due to the various appearances, deformation, and occlusion of clothes. To conquer these complications, this study utilises multiscale contextual information without losing resolution.

To manipulate landmarks of fashion clothes, researchers focus on techniques such as regression and heatmaps production. FashionNet [6] is based on VGG-16 [7] and employs regression-based landmark localisation alongside predicting visibility of landmarks. Application of pooling at the predicted landmark layer is then facilitated to classify clothing items. Liu et al. [8] confront the large variations in fashion images, so that their work embellishes the auto-routing mechanism using the pseudo-label scheme to improve the obstacle in variability of the fashion landmarks. Deep landmark network (DLAN) [4] fused different scale variations of images using maximum selection in dilated convolutions and exploited hierarchical spatial transformer network by iteratively updating the landmark locations to handle background clutters.

Recently, Wang et al. [3] proposed a grammar network for landmark detection by using bidirectional convolutional recurrent neural network (BCRNN) to grasp the landmark heatmaps. Spatial-aware non-local (SANL) attention [9] mechanism also demonstrated good performance in landmark detection. The authors established a framework on feature pyramid network, equipped with four SANL blocks which are constructed from the non-local block in a residual manner [10] to learn the spatial-related representation by taking a spatial attention map from Grad-CAM [11].

Another method inspired by part affinity fields (PAF) architecture [12] was reported by Chang Qin et al.[1]. This method employs probabilistic maps of landmark points by taking element-wise summation of probabilities of landmark points' positions and interrelation between landmarks through two branches from convolutional neural network.

### 2.2 Attribute prediction and clothes classification

Most of the work in attribute prediction and clothes classification involved the use of clothing landmarks to enhance the performance of fashion clothes classification. Works from [1–3,9] showed the importance of the landmark-driven knowledge in classification. Li et al. [13] also reported the model which uses knowledge sharing strategy by utilising boundaries and structure of the target predicted from landmarks. In contrast, Lee et al. [14] came up with the idea of landmark-free clothes classification via exploiting feature selective network. In their methodology, they proposed a multitask learning network, divided into two tasks: attribute prediction and category classification, where attribute prediction was enhanced with the help of class activation map derived by [11] and higher activated feature selection. By contrast, Hyunsoo et al. citefiftyspsseven proposed a neural network-based image encoder followed by a hierarchical classifier over a set of annotated categories. The hierarchical classification block combines the merits of both the local and global nature of feature extraction which is transformed as a type of multitask learning problem. Chen et al. [15] proposed a deep convolutional neural network architecture that captures spatial details and selects task-related features to detect precise fashion landmark detection through a dual attention feature enhancement (DAFE) module. Moreover, detailed human body configuration analysis has also drawn much attention in application of fashion synthesis where Wang et al. [16] formulated human parsing task as a neural information fusion process by defining the human body as a hierarchy of multilevel semantic parts.

### 2.3 Attention mechanism

Fashion landmark detection enhances the performance of fashion clothes classification by aggregating features from functional regions of clothes. Such input may not extract further information to determine the categories and attributes. In this case, global attention derived from semantic information helps to constraint the classification network. Attention mechanism was absorbed into many machine learning approaches such as image recognition [17–19], visual question answering [20,21], image captioning [22], image editing [23], and image generation [24]. These methods demonstrate effective top-down attention mechanisms to better understand how the regions of interest (ROIs) are selected in images.

A landmark-aware attention along with category-driven attention is reported by Wang et al. [3] based on message passing over fashion grammar and top-down approach in category and attribute prediction. Besides, Jingyuan et al. [2] proposed a landmark-driven network which utilises upsampling technique through basic encoder-decoder architecture and feeding landmark attention map into classification of clothes to facilitate added advantage.

In this work, two attentions are proposed, landmark- and spatial–channel-driven attentions, in terms of elevating the performance of the fashion clothes classification. We employ spatial- and channel-wise awareness cooperative with dilated convolutions [25] for localising landmarks which aggregates multiscale contextual information.

## 3 Methodology

We propose a network based on VGG-16 [7] that seamlessly embraces fashion clothes classification by providing two attention pipelines: (1) landmark-driven attention through a multiscale contextual extraction of landmark localisation and (2) spatial–channel attention constructed through both spatial region and channels of feature map blocks.

### 3.1 Landmark-driven attention

Fashion landmarks are known as the keypoints of functional area of clothes [8]. These keypoints can represent features of items of clothing in an effective way. This subnetwork utilises stacked dilated convolutional (SDC) blocks. We illustrate the implementation of the proposed landmark localisation subnetwork in Fig. 2.

*Stacked Dilated Convolutional Block*

Dilated convolution is simply a convolution applied to the input feature map with designated distance between the kernel points. The distance is determined by the dilation rate $\mathcal{D}$. Dilated convolutions increase the receptive field size and let the convolution operation wrap more relevant information from input feature maps using different values for $\mathcal{D}$ [25]. An certain way of explanation for dilated convolutions is that when $\mathcal{D} = 1$ then it is a standard convolution operation. If $\mathcal{D} = n$, then $n - 1$ zeros will be added between every neighbouring kernel points in the standard kernel (Fig. 3a). Let $d$ be the dilation rate, and the dilated convolution will be defined as,

$$F_{*d}k(p) = \sum_{p} F(s)k(t) \tag{1}$$

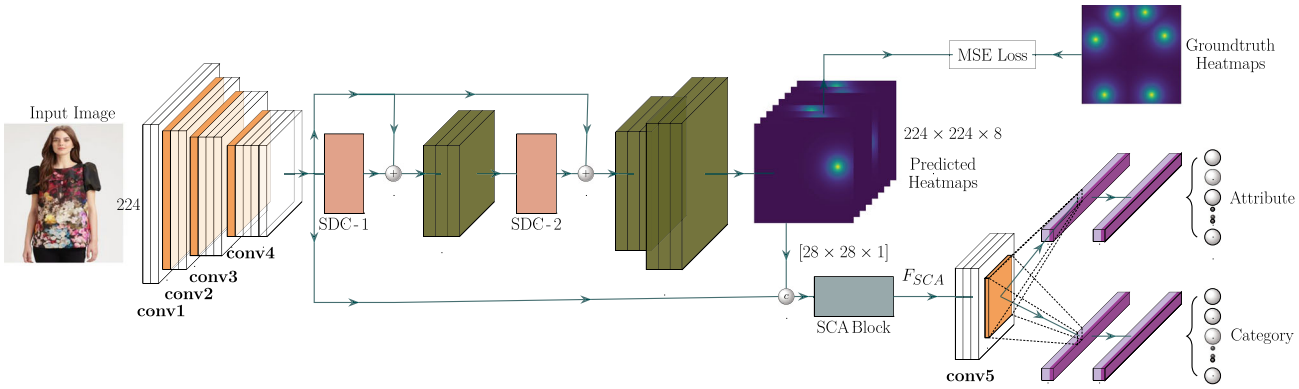


**Fig. 2** The illustration of the proposed framework which consists of three parts: (i) A pair of SDC blocks (pink cubes) are established for extracting multiscale contextual information of the clothing landmarks, (ii) transposed convolutions (green cubes) to produce high-resolution heatmaps, and (iii) SCA block (blue cubes) to capitalise the efficient feature extraction of clothing images

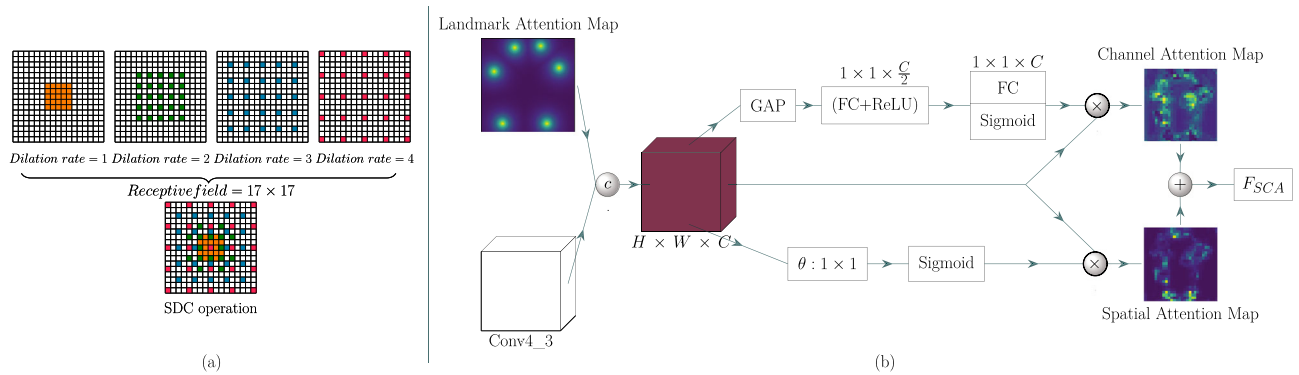


**Fig. 3** Figure (**a**) shows the example operations using $5 \times 5$ kernel with dilation rate=1, 2, 3, and 4, respectively, and the overall effect of SDC operation on a feature map with receptive field size of $17 \times 17$. Figure (**b**) illustrates the structure of SCA pipeline for fashion attribute prediction and clothes classification. Top branch indicates the illustration of channel-wise attention which is initially achieved by global average pooling (GAP) where bottom branch stipulates spatial attention through a $1 \times 1$ convolution

where $*d$ and $k$ are the dilated convolution operator and kernel, respectively. $p$ is the receptive field size and denoted as $s + dt$. Here, we use the stacked dilated convolutions which are constructed in parallel by exponentially increasing dilation coefficients and receptive field to construct a multiscale context aggregated block without losing resolution of the feature maps.

In this landmark branch, we extend the Conv4_3 block of VGG-16 by adding two convolutions to build feature maps with depth size of 64 and then stack two SDC blocks at both end of the first upsampling process (feature maps ended up with the size of $56 \times 56$) where we use $4 \times 4$ transposed convolutions to upsample the feature maps. Transposed convolutions that furnish learnable parameters are more effective than using predefined interpolation methods. At each end of the SDC block, we concatenate the features that emerged from the last convolution layer placed before applying the SDC blocks. Three sets of transposed convolutions followed by $3 \times 3$ convolutions are used to produce heatmaps with the size of input images in order to preserve high resolutions of output. Final output heatmaps from landmark branch are then converted into a $28 \times 28$ landmark attention map by taking maximum pixel value through channels. Global features include representation of outline, shape, and texture features. Combining local features with global features is recommended for classification, as it only represents the texture in an image region [26]. In this regard, the attention map is then concatenated with feature maps produced from Conv4_3 of VGG-16 to pass the global representation of items of clothing in a constructive manner. Unlike the structure constructed in [2] which creates the landmark attention map through feature map upsampling using several convolution blocks, here the attention map is directly concatenated with global features and then fed into spatial- and channel-wise attention pipeline which effectively make the transition of features to classify categories and attributes of clothes.

### 3.2 Spatial–channel attention

Attention mechanisms in accordance with the several implementation using spatial, semantic, and/or channel information influence image captioning [22], medical image segmentation [27], and regression networks [28]. In fashion clothes classification, areas of interest are usually concentrated in smaller areas. At the same time, spatial information generally contains important features for accurate identification, so it should also be used reasonably. It should be noted that each convolution filter acts as a pattern detector, and each feature map channel on convolutional neural network is activated by the response of the corresponding convolution filter. Therefore, applying the attention mechanism in a channel-wise manner can be considered a process of selecting important semantic attributes. In this work, we use the spatial attention together with channel attention which are built in parallel to potentially extract features by collaborating global and landmark-driven features in order to attribute prediction and category classification of fashion clothes.

In this regard, landmark attentive features are then fed into another pipeline called spatial–channel attention (SCA) introduced in [29]. Inspired from squeeze-and-excitation block [30], this attention utilises knowledge of feature responses to take the channel-wise information as well as the spatial-wise information into account (Fig. 3b).

Global average pooling (GAP) is applied to produce spatially squeezed statistics $C$ from feature maps $F = [f_1, f_2, \ldots, f_h] \in \mathbb{R}^{H \times W}$, where $H \times W$ is the size of the feature map, and $h$th element of statistic can be calculated as,

$$C_h = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} f_h(i, j). \tag{2}$$

Calculated statistics $C$ is followed with the utilisation of two fully connected layers aiming at limitation of model complexity with respect to generalisation and passed through sigmoid activation function. Dimensionality reduction ratio in a fully connected layer is set to half the size of dimension of the previous convolutional block. The enhanced values are then combined with Conv4_3 features using Hadamard product $\otimes$. The resulting feature maps consist of reweighed features along with channels. As long as the number of channels is large, we can significantly improve performance by adding channel attention. Thus, the channel-wise attention is integrated in between the place of Conv4_3 and Conv5 of VGG-16 structure. Meanwhile, spatial attention is produced with the reweighted features along with spatial regions. This can be achieved by taking the channel squeeze using $1 \times 1$ convolution which estimates the linearly combination of all $C$ channels in spatial region $H \times W$. Spatially produced weighted map is redeemed through the sigmoid function similar to channel-wise attention. Finally, both attentive feature maps from spatial and channel attentions are added to get the output ($F_{SCA}$) which are then fed into the Conv5 block of VGG-16 to further classification of clothing items and prediction of clothing attributes.

## 4 Experimental set-up and testing results

### 4.1 Dataset

We evaluate the performance of the proposed model using two large-scale benchmark fashion clothes datasets: Deep Fashion-C dataset [6] and fashion landmark detection (FLD) dataset [8].

*DeepFashion-C dataset*

This large-scale dataset is used to evaluate the proposed landmark localisation, category, and attribute prediction. It contains 289,222 annotated fashion clothes images. Each image is labelled with 46 clothing categories, and 1000 attributes which are categorised into five types: Texture, fabric, style, part, and shape. In this dataset, each image is further annotated with eight landmarks (e.g. collar, sleeve, hem, and waistline) and bounding box coordinates.

*FLD dataset*

This dataset is used only to evaluate the performance of fashion clothes landmark localisation as no information about categories of clothing items is provided. It contains 123,016 clothing images with eight annotated landmarks. Annotation also includes bounding box and clothing type (e.g. upper-body, lower-body, and full-body clothes) for each image.

### 4.2 Network architecture

Our method is built upon VGG-16 architecture with dilated convolutions (Sect. 3.1) and spatial–channel attention (Sect. 3.2). The first four blocks of network are similar as VGG-16 up to Conv4_3, and we split rest of the network into two branches, one for landmark localisation and the other branch for attribute prediction and clothes classification. We use $3 \times 3$ kernel followed by $5 \times 5$ kernel to do the dilated convolutions with $\mathcal{D} = 1, 2, 3,$ and 4. Finally, our network would result eight heatmaps for eight landmarks of size $28 \times 28$ through landmark localisation branch. In the other attention pipeline, emerged feature maps ($F_{SCA}$) through SCA block are then fed into the Conv5 block of VGG-16. We subdivide the fully connected (FC) layers as two divisions incorporating with attribute prediction and category classification. Each division consists of two FC layers with the size of $1 \times 1 \times 4096$. We employ mean-squared error for a fair comparison between ground truth heatmaps which are 2D

**Table I** Experimental results on the DeepFashion-C dataset for landmark localisation using normalised distance metrics

| Methods | L.Collar | R.Collar | L.Sleeve | R.Sleeve | L.Waist | R.Waist | L.Hem | R.Hem | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Liu et al. [6] | 0.0854 | 0.0902 | 0.0973 | 0.0935 | 0.0854 | 0.0845 | 0.0812 | 0.0823 | 0.0872 |
| Liu et al. [8] | 0.0628 | 0.0637 | 0.0658 | 0.0621 | 0.0726 | 0.0702 | 0.0658 | 0.0663 | 0.0660 |
| Yan et al. [4] | 0.0570 | 0.0611 | 0.0672 | 0.0647 | 0.0703 | 0.0694 | 0.0624 | 0.0627 | 0.0643 |
| Wang et al. [3] | 0.0415 | 0.0404 | 0.0496 | **0.0449** | 0.0502 | 0.0523 | 0.0537 | **0.0551** | 0.0484 |
| Jingyuan et al. [2] | 0.0332 | 0.0346 | 0.0487 | 0.0519 | 0.0422 | 0.0429 | 0.0620 | 0.0639 | 0.0474 |
| Ours | **0.0323** | **0.0334** | **0.0443** | 0.0472 | **0.0368** | **0.0370** | **0.0533** | 0.0558 | **0.0425** |

**Table II** Experimental results on the FLD dataset for landmark localisation using normalised distance metrics

| Methods | L.Collar | R.Collar | L.Sleeve | R.Sleeve | L.Waist | R.Waist | L.Hem | R.Hem | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Liu et al. [6] | 0.0784 | 0.0803 | 0.0975 | 0.0923 | 0.0874 | 0.0821 | 0.0802 | 0.0893 | 0.0859 |
| Liu et al. [8] | 0.0480 | 0.0480 | 0.0910 | 0.0890 | - | - | 0.0710 | 0.0720 | 0.0680 |
| Yan et al. [4] | 0.0531 | 0.0547 | 0.0705 | 0.0735 | 0.0752 | 0.0748 | 0.0693 | 0.0675 | 0.0672 |
| Wang et al. [3] | 0.0463 | 0.0471 | 0.0627 | **0.0614** | **0.0635** | 0.0692 | **0.0635** | 0.0527 | 0.0583 |
| Ours | **0.0437** | **0.0435** | **0.0610** | 0.0678 | 0.0638 | **0.0618** | 0.0684 | **0.0514** | **0.0577** |

Gaussian centred on the landmarks and predicted heatmaps in training the landmark localisation branch. Further, we use standard cross-entropy loss and weighted cross-entropy loss to train category classification and attribute prediction, respectively.

Our model is implemented using PyTorch [31] and optimised using Adam optimiser [32]. In each iteration we use mini-batch of size 16. Initially, the learning rate is set to 0.0001 which is decreased by a factor of ten when the validation accuracy plateaus. We first pretrained the landmark localisation branch with six epochs and then the entire model is trained with twelve epochs. Some of the extended studies are experimented with/without the modules used in our work, and the results are reported in Sect. 4.4.

### 4.3 Quantitative results

*Experimental set-up*

We divide the training and testing images as mentioned in the DeepFashion-C dataset (209,222 images for training and 40,000 images for testing and validation). We crop all the images using the annotated bounding box labels and then resize them into $224 \times 224$. Following the etiquette of FLD dataset, we split training, validation and testing images as mentioned in the annotation: 83,033 images for training, 19,992 images for validation, and 19,991 images for testing.

*Performance evaluation.*

For landmark localisation, we used normalised distance metrics which refers the Euclidean distance between the predicted heatmaps and ground truth (normalised with respect to the height and the width of the input image) to evaluate the performance and compare it with five deep learning models [2–4,6,8]. Tables I and II state the comparison results of landmark localisation with normalised distance metric.

As seen, the model achieves a high score over almost all landmarks and an average score of 0.0425 for DeepFashion-C dataset and outperforms compared state-of-the-art techniques. The model achieves 0.0577 for FLD dataset and outperforms four delicate techniques on generic set up. Figure 4 shows some example heatmaps predicted from landmark localisation branch along with ground truth heatmaps, and predicted landmark locations plotted on corresponding input images. Besides, Fig. 5 shows some of the examples for the attentive features learned through the spatial- and channel-wise attention pipeline built on category classification and attribute prediction. Output from both attention pipelines shows the clear transition of landmark points and keypoints detection (e.g. belt, print, buttons, clothe label, etc.). For category classification and attribute prediction, we applied top-$k$ classification accuracy and top-$k$ recall rate, respectively, and it is evident that the proposed network achieves superior performance. We compared the performance with seven recently reported works [3,5,6,14,33–35]. The results reported in the literature make use of top-3 and top-5 accuracies. Therefore, our test results are reported using those measures in Table III. Our model shows 91.02 and 96.20 for top-$k$ accuracy, and for the attribute prediction, overall top-$k$ recall achieves 51.89 and 62.04 where $k = 3$ and 5, respectively.

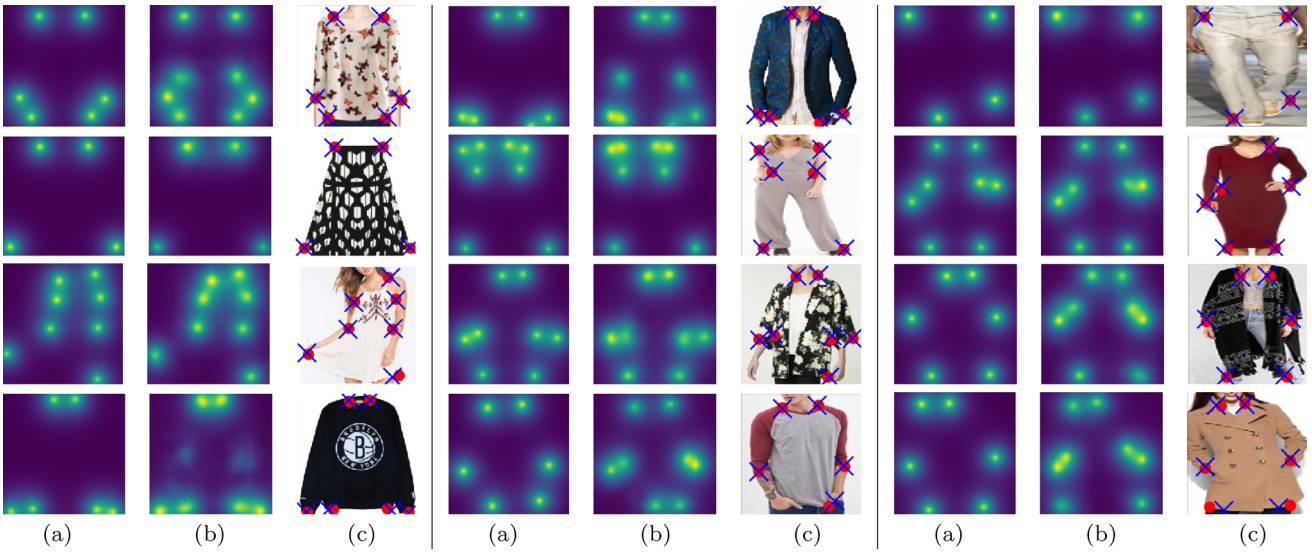


**Fig. 4** Example results of landmark localisation. (**a**) Ground truth heatmaps which are 2D Gaussian centred on the landmark points. (**b**) Heatmaps produced by proposed landmark localisation branch. (**c**) Plotted landmarks on the cropped input images (ground truth in blue cross and predicted in red circle)

**Fig. 5** Example images depict the learned features through spatial–channel attention pipeline for some of the clothing categories. Images at the left are the input images, and images at the right are the attentive feature maps from SCA pipeline

## 4.4 Ablation study

In these experiments, we also performed an extensive study on each component of our fashion clothes classification model. The experimental results are summarised in Table IV.

Additionally, we evaluate the effectiveness of the proposed attention using SDC blocks to localise landmarks in clothes. The landmark localisation component without having the SDC blocks shows an average of 0.0606 normalised distance metric which performs better than the state-of-the-art methods in [4,6,8] evaluated on the DeepFashion-C dataset. For the FLD dataset, it shows an average of 0.0724 normalised distance metric. Receptive fields of feature maps are changed in dilated convolutions based on the empirical study of experiments. The study based on kernel size of SDC

**Table III** Experimental results for category classification and attribute prediction on DeepFashion-C dataset

| Methods | Category | | Attribute | |
|---|---|---|---|---|
| | Top-3 | Top-5 | Top-3 | Top-5 |
| Kiapour et al. [33] | 82.58 | 90.17 | 27.46 | 35.37 |
| Huang et al. [34] | 59.48 | 79.58 | 42.35 | 51.95 |
| Liu et al. [6] | 82.58 | 90.17 | 45.52 | 54.61 |
| Corbiere et al. [35] | 86.30 | 92.80 | 23.10 | 30.40 |
| Wang et al. [3] | 90.99 | 95.78 | 51.53 | 60.95 |
| Lee et al. [14] | **91.37** | 95.26 | 47.70 | 57.28 |
| Hyunsoo et al. [5] | 91.24 | 95.68 | – | – |
| Ours | 91.02 | **96.20** | **51.89** | **62.04** |

**Table IV** Ablation study on category and attribute prediction tested on DeepFashion-C dataset using top-$k$ accuracy

| Methods | Category | | | Attribute | |
|---|---|---|---|---|---|
| | Top-1 | Top-3 | Top-5 | Top-3 | Top-5 |
| Baseline | 64.33 | 82.64 | 89.08 | 35.96 | 54.07 |
| Baseline+ SCA(parallel) | 65.03 | 84.80 | 91.47 | 36.43 | 55.41 |
| Baseline+ LDA | 71.79 | 89.47 | 94.90 | 42.44 | 55.76 |
| Baseline+ LDA+ SCA(spatial–channel) | **72.26** | **90.98** | **95.60** | **49.34** | **58.74** |
| Baseline+ LDA+ SCA(channel–spatial) | 70.28 | 89.03 | 94.78 | 46.58 | 55.29 |

blocks shows that the exponentially increasing size of kernel with respect to the size of feature map influences the performance of keypoint localisation. $3 \times 3$ kernel is helpful in effectively extracting the landmark points from SDC block in standard size of the feature maps of VGG-16 convolutional blocks, while $5 \times 5$ kernel is cooperative with upsampled feature map of size $56 \times 56$. Besides, recurrence of multiple stacked dilated convolutions throughout the landmark localisation branch will mislead the representation of heatmaps at the end of the subnet because of the multiscale feature aggregation, so that the number of stacked blocks which uses dilated convolutions are limited to two in this work from the heuristic study of experiments.

The landmark attention is concatenated channel-wise on the global features, and thereafter it goes through a SCA pipeline to improve the feature learning and feature integration. So, it is expected to formalise the representation in a positive manner to represent the features of clothing items as the SCA pipeline makes full use of the characteristics of CNN and can produce interesting image features in terms of spatial regions and channels; thus, the performance has been improved compared to the state-of-the-art works. Since clothing landmarks localise the functional region of clothes, landmark attention produces the features through a set of landmarks and work as one of the designated feature to next layer. We further studied the impact of the SCA pipeline in learning spatial- and channel-wise information to effectively classifying clothing items and predicting their attributes. Both spatial and channel attention pipelines structured in parallel increase the performance by 1.2% in top-1 accuracy than constructing them in sequence as proposed in [22]. Our proposed framework mainly relies on baseline VGG-16 pretrained model. Furthermore, we evaluate overall in-depth performance of our framework structure in significant ways: (i) training the baseline model with landmark-driven attention (LDA), (ii) training the baseline model with SCA pipeline structured in parallel, (iii) training the model with both LDA and sequentially structured SCA pipeline (spatial–channel), and (iv) training the model with both LDA and swapping the sequentially structured SCA blocks along the pipeline (channel–spatial).

Our proposed model achieved 73.33% for top-1 accuracy which shows an improvement in performance. This advocates that each of the component in our model strongly influences the performance of the fashion clothes landmark localisation and category classification with its attribute prediction.

## 5 Discussion and conclusion

In this paper, we presented an attention-driven and deep learning-based network for fashion clothes classification leading to differentiable network that can be trained end to end. Our model facilitates the stacked dilated convolutional blocks along with upsampling technique which achieves high-resolution heatmaps at the end of fashion clothes landmark localisation. The network also encodes channel- and spatial-wise information through focusing on understanding of what and where the important feature attention exists, respectively, in feature maps. We further utilise the importance of the contextual clues from representation of landmarks in classifying fashion clothes. We demonstrated our experiments on two benchmark datasets and achieved state-of-the-art performances in visual fashion clothes classification and attribute prediction against recently proposed methods.

### Compliance with ethical standards

**Conflict of interest** We declare that we have no conflict of interest.

## References

1. Huang, C.-Q., Chen, J.-K., Pan, Y., Lai, H.-J., Yin, J., Huang, Q.-H.: Clothing landmark detection using deep networks with prior of key point associations. IEEE Trans. Cybern. **49**(10), 3744–3754 (2019)
2. Jingyuan, L., Lu, H.: Deep fashion analysis with feature map upsampling and landmark-driven attention. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 30–36 (2018)
3. Wang, W., Xu, Y., Shen, J., Zhu, S.-C.: Attentive fashion grammar network for fashion landmark detection and clothing category clas-

sification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4271–4280 (2018)
4. Yan, S., Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 172–180 (2017)
5. Cho, H., Ahn, C., Min Yoo, C. , Seol, J., Lee, S. G.: Leveraging class hierarchy in fashion classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
6. Liu, Z., Luo, P., Qiu, S., Wang, X.,Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1096–1104 (2016)
7. Simonyan,K., Zisserman,A., Very deep convolutional networks for large-scale image recognition. In: Third International Conference on Learning Representations (ICLR) (2015)
8. Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X.: Fashion landmark detection in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 229–245. Springer, Berlin (2016)
9. Li, Y., Tang, S., Ye, Y., Ma, J.: Spatial-aware non-local attention for fashion landmark detection. arXiv preprint arXiv:1903.04104 (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
11. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)
12. Cao, Z., Simon, T.,Wei, S.-E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7291–7299 (2017)
13. Li, P., Li, Y., Jiang, X., Zhen, X.: Two-stream multi-task network for fashion recognition. arXiv preprint arXiv:1901.10172 (2019)
14. Lee, S., Eun, H., Oh, S., Kim, W., Jung, C., Kim, C.: Landmark-free clothes recognition with a two-branch feature selective network. Electron. Lett. **55**(13), 745 (2019)
15. Chen, M., Qin, Y., Qi, L., Sun, Y.: Improving fashion landmark detection by dual attention feature enhancement. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
16. Wang, W., Zhang, Z., Qi, S., Shen, J., Pang, Y., Shao, L.: Learning compositional neural information fusion for human parsing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5703–5713 (2019)
17. Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 2017–2025 (2015)
18. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164 (2017)
19. Fan, D.P., Wang, W., Cheng, M.M., Shen,J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 8554–8564 (2019)
20. Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., Nevatia, R.: ABC-CNN: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960 (2015)
21. Shih, K. J., Singh, S., Hoiem, D.: Where to look: focus regions for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4613–4621 (2016)
22. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5659–5667 (2017)
23. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. Pattern Anal. Mach. Intell. **41**(7), 1531–1544 (2018)
24. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5572–5581 (2019)
25. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
26. Kabbai, L., Abdellaoui, M., Douik, A.: Image classification by combining local and global features. Vis. Comput. **35**(5), 679–693 (2019)
27. Chen, Y., Wang, K., Xiangyun, L., Qian, Y., Wang, Q., Yuan, Z., Heng, P.: Channel-Unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation. Front. Genet. **10**, 1110 (2019)
28. Junyu, G., Wang, Q., Yuan, Y.: SCAR: spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing **363**, 1–8 (2019)
29. Shao, J., Qu, C., Li, J., Peng, S.: A lightweight convolutional neural network based on visual attention for SAR image target classification. Sensors **18**(9), 3039 (2018)
30. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7132–7141 (2018)
31. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer,A.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017)
32. Kingma, D. P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the Third International Conference on Learning Representations (2015)
33. Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A. C., Berg, T. L.: Where to buy it: matching street clothing photos in online shops. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3343–3351 (2015)
34. Huang, J., Feris, R. S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1062–1070 (2015)
35. Corbiere, C., Ben-Younes, H., Ramé, A., Ollion, C.: Leveraging weakly annotated data for fashion image retrieval and label prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2268–2274 (2017)
36. Lee, S., Oh, S., Jung, C., Kim, C.: A Global-local embedding module for fashion landmark detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
37. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 401–417 (2018)

**Majuran Shajini** received her B.Sc. Special in Computer Science from the University of Jaffna, Sri Lanka, where she is currently pursuing her PhD degree in Computer Science. Her research interests include image/signal processing, pattern recognition, machine learning, and deep learning.

**Amirthalingam Ramanan** received his B.Sc. Special in Computer Science from University of Jaffna, Sri Lanka, and his PhD in Computer Science from the University of Southampton, UK. He is a Senior Lecturer in the Department of Computer Science, University of Jaffna, where he is heading the department since June 2017. His research interests include image processing, computer vision, pattern recognition, machine learning, and deep learning.