Electronic Letters on Computer Vision and Image Analysis 20(2):1409, 2021

A Multi-staged Feature-Attentive Network for Fashion Clothing Classification and Attribute Prediction

Shajini Majuran* and Amirthalingam Ramanan*

* *Department of Computer Science, University of Jaffna, Sri Lanka*

Received 08 March 2021; Accepted 08 December 2021

Abstract

In visual fashion clothing analysis, many researchers are attracted with the success of deep learning concepts. In this work, we introduce a multi-staged feature-attentive network to attain clothing category classification and attribute prediction. The proposed network in this work brings out a landmark-free structure, whereas the existing landmark-driven structures take up a lot of manpower for landmark annotation and also suffers from inter- and intra-individual variability. Our focus in this work is to intensify feature extraction by incorporating low-level and high-level feature fusion within a fashion network. The feature fusion helps the network to manifest spatial and rich semantic representation in each level of the network. Besides, the proposed model utilises spatial and channel-wise attention to further enrich the multi-staged features in producing contextual information. Additionally, we enclose a semi-supervised learning approach to escalate the proposed architecture in fashion clothing analysis that utilises collaborative learning using labelled and unlabelled data. The proposed approach is evaluated on large-scale DeepFashion-C dataset while the unlabelled dataset for semi-supervised learning is obtained from six publicly available fashion datasets. Experimental results show that the proposed multi-staged feature-attentive network entailing deep convolutional neural network outperforms the state-of-the-art techniques considerably, in fashion clothing analysis.

Keywords: Feature-attentive network, Fashion clothing analysis, Fashion attribute prediction, Semi-supervised learning, DeepFashion, Landmark-free approach.

1 Introduction

The fast growth in fashion brands and the development in e-commerce giants have led the fashion industry to urge spotting more valuable customers via collecting and analysing large amount of digitalised fashion related data. Artificial intelligence begins to flourish the fashion domain with wide range of applications and innovations through different scenarios such as detection, synthesis, analysis, and recommendation. However, fashion analysis is a challenging task owing to the fact, it has great changes in trends, style and design compared to other general objects. Therefore, numerous researches have been carried out in clothing modelling, recognition, parsing, and retrieval addressing degrees of clothing related challenges [1, 2, 3, 4] in fashion analysis. All these works, relatively accommodate detection of clothes area using bounding box prediction, analysing important landmarks which distinguish various clothing categories, and predicting its attributes. The generic tasks carried out in fashion clothing detection and classification are depicted in Figure 1.

Correspondence to: shayu.kiri@gmail.com. The authors contributed equally.

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1409
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

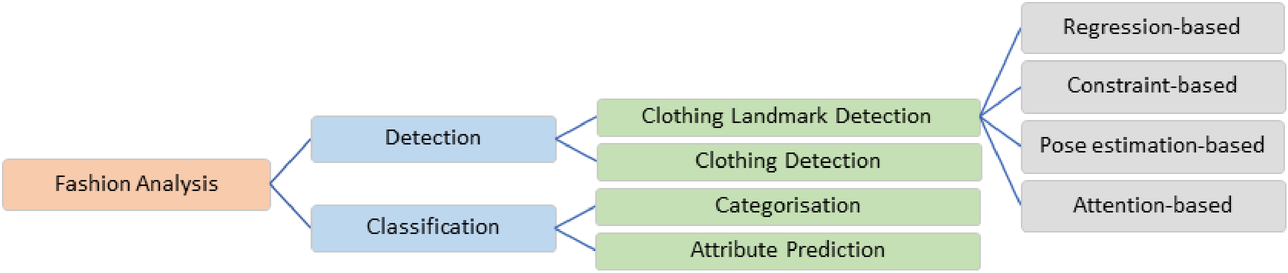


Figure 1: Outline of the generic tasks carried out in fashion clothing detection and classification.

As seen, earlier fashion models relied primarily on handcrafted functions and search for powerful clothing displays such as graphic models, context information, general object suggestions, human parts, bounding-boxes and semantic masks. In these days, many performances in fashion clothing analysis tasks have been repeatedly revealed that deep neural networks can achieve better performance with learning problems using a large-scale labelled data [3, 4, 5, 6, 7, 8, 9]. Inspired by visual attention mechanisms, many researchers have tried to model a soft attentive network to prompt the performance of computer vision tasks [10, 11, 12, 13]. Works that have been reported in the fashion analysis using attention-based deep learning architectures shows that the overall performance is enhanced [1, 3, 5, 6, 7, 14]. In a way, most of the researches focus on the attention obtained from clothing landmark detection to improve category and attribute prediction [3, 4, 5, 7, 14] and various landmark estimation works are reported in supporting the mission in clothing classification [1, 6, 9, 15]. These works appraise the importance of clothing landmarks which are fundamental regional points to express the structure of an clothing item. Though the performance shows improvement, the studies on landmarks are troublesome and time consuming. It also suffers from unique differences in clothing items so that combining or grouping landmarks is a crucial task. On that account, the effective feature engineering which is able to learn discriminative feature representation becomes eminent part in clothing category classification and attribute prediction. In addition, we notice that as in feature pyramid networks (FPN) [16], the fusion of low-level and high-level visual features impel the extraction to different contextual features throughout the network to capitalise attention in helping effective feature representation. This approach is worked on the benefit in identifying salient cloth regions and amplifying their influence, while terminating the irrelevant information in other regions.

Additionally, we address major tasks in fashion analysis using multitask learning technique that extracts feature representation of clothing categories as well as its attributes in semi-supervised manner. Most of the time, the form of an architecture is strongly controlled by labelled data. However, fashion data for clothing analysis are in large-scale, yet, the success on annotated datasets are expensive due to a lot of human effort, pain and/or financial cost in creating such large datasets. Therefore, semi-supervised learning (SSL) has proven to be a powerful paradigm to leverage unlabelled data to mitigate the reliance on large labelled datasets by combining supervised and unsupervised learning [17, 18, 19, 20, 21]. We structure a network which can make use of labelled and unlabelled samples together so that the additional training can be avoided. Inspired from various architectures, minimising the entropy of the prediction function is used which takes the performance to a step forward since the classification cost is not specified for unlabelled samples. As a result, our SSL model shows significant improvement to the clothing analysis architectures.

Hence, the focus of this work is to construct an integrated model which is capable of identifying clothing categories along with its distinctive attributes. Our model comes to grip with the feature attention mechanism for multiscale collaboration and contextual supervision among features from low to high levels of network.

To summarise, our main contribution is three-fold:

- Formalised fashion analysis into a multitask deep neural network for clothing category classification and its attribute prediction,
- Established a multi-staged feature-attentive network through multiscale contextual feature supervision and semantic feature engineering using spatial and channel attention, and
- Experimented a semi-supervised learning approach by integrating collaborative learning architecture for clothing category classification and its attribute prediction using large-scale of unlabelled fashion data.

The rest of the manuscript is structured as follows: Section 2 discusses the previous works carried out in the field of fashion clothing analysis. Section 3 describes the proposed methodology in this work. Section 4 briefly outlines the experimental setup, testing results, findings, and ablation study. Finally, Section 5 concludes this work with future extension.

2 Related Work

In this section, recent works that have been reported in the literature of fashion clothing analysis are summarised based on the tasks: Landmark detection, landmark-driven and landmark-free classification.

2.1 Fashion Landmark Detection

The main task of fashion landmark detection is to recognise and locate the functional points of clothing images such as sleeve-end, collar points, waistline, and hemline. It is the key force to improve other fashion applications such as fashion classification, retrieval, design and recommendation. Fashion landmark detection has been used through various techniques such as regression [4, 22], pose estimation-based methods [23], constraint-based methods [24, 25] and as attentive knowledge for category classification [5, 7, 14, 26, 27]. However, due to various deformations and changes in fashion images, fashion landmark detection is still difficult to apply in actual industrial domains.

2.2 Fashion Clothing Classification

2.2.1 Landmark-driven Approaches

Liu and Lu [5] proposed an attentive fashion network based on VGG-16 by giving knowledge through more accurate landmark localisation by producing high-resolution landmark heatmaps using transposed convolutions. With the help of predicted landmarks, a landmark-driven architecture is proposed to improve the accuracy of fashion category classification and attribute prediction leading to a fully differentiable network that can be trained end-to-end. Wang *et al.* [7] proposed a fashion network based on VGG-16 architecture that employs features from the third layer of fourth convolutional block for landmark-driven feature representation. Authors introduced a bidirectional convolutional recurrent neural network (BCRNN) architecture by processing message passing over grammar arrangements which is flexible to generate more sensible landmark locations. Authors focused on capturing the dependency grammar such as kinematics-like relation and symmetry grammar integrating the bilateral symmetry of clothes. Li *et al.* [26] proposed a two-stream multitask network based on ImageNet pretrained ResNet50 by designing two knowledge-sharing strategies which enables information transfer between tasks and improves the overall performance for clothes landmark detection, category classification, and attribute prediction. Authors also exhibited two awareness methods: Boundary awareness and structural awareness to semantically share representation and aggregate features among different tasks.

Further, Shajini and Ramanan [14] proposed a multitask model based on VGG-16 which focuses more towards the feature attention through predicted landmark points in order to enhance the category classification and attribute prediction. The extracted features from parallel dilated convolutions are then concatenated with dedicated global features and acquire the benefit of transposed convolutions to produce high resolution heatmaps for better localisation of landmarks. The attentive map from landmark localisation branch is then concatenated with the global features of classification branch for clothing classification. Besides, Zhang *et al.* [27] proposed a two-stream fashion network based on VGG-16 incorporating the importance of the landmarks along with impelling effect of the texture and shape to enhance the performance of category classification accompanying attribute prediction based on clothing images. The authors proposed a network combining texture-biased stream using the pretrained model of ImageNet and shape-biased stream using attention through localised clothing landmarks along with its visibility features.

2.2.2 Landmark-free Approaches

Lee *et al.* [8] proposed a landmark-free clothes classification via exploiting feature selective network based on VGG-16. A multitask learning network is divided into attribute prediction and category classification. The attribute prediction was enhanced with the help of class activation map derived by [28] and higher activated feature selection. The average pooling is applied to the activation map and then higher n values are selected to manifest the activated values. Besides, Ferreira *et al.* [3] introduced the relation between attribute localisation and visual appearance by implanting a semantic attention module guided by body pose estimation. The pose estimation is done using the off-the-shelf pose detector OpenPose [29]. Authors designed a visual semantic attention model which uses VGG-16 as the basic network to produce heatmap accompanying combined key joints obtained from OpenPose.

Differently, Corbiere *et al.* [30] proposed a ResNet-50 based model for feature extraction that is directed with a weakly supervised text embedding for fashion images collected from e-commerce websites. For text embedding, labels are predicted from bag-of-words description consisting of probability of each word among the vocabulary. Another work from Cho *et al.* [31] proposed a fashion category classification model that explores hierarchical arrangements of clothing categories. The model consists of two components: Neural network-based image encoder and a hierarchical classifier over a set of annotated categories. The proposed model is implemented on hierarchical multilabel classification networks-feedforward [32].

2.3 Attention

Attention mechanism plays a huge role in both natural language processing and computer vision tasks. Many machine learning approaches absorbed attention-based architectures to perform various tasks such as image recognition [13, 33, 34], visual question answering [35], image captioning [36], and image editing [37]. Squeeze-and-Excitation network [10] is proposed based on the channel attention to weight the relation between the channel and prime information. Non-local neural network [11] computes the response at a position on the feature map as a weighted sum of the features at all positions which assist to a receptive field same as the feature map size. The work in [38] proposed a simple transformer network by assigning both recurrence and convolutions with stacked self-attentions.

In fashion clothing analysis, the importance of the landmark-driven attention in classification has shown an effective enhancement in [6, 7, 14, 15]. Spatial-aware non-local attention (SANL) [6] mechanism shows better performance in landmark detection by expressing the effectiveness of the attention. A landmark-aware attention along with category-driven attention is also reported by Wang *et al.* [7] based on VGG-16 network and top-down approach in category and attribute prediction. Besides, Liu and Lu [5] proposed a landmark-driven network which utilises upsampling technique through basic encoder-decoder architecture and feeding landmark attention map into classification of clothes.

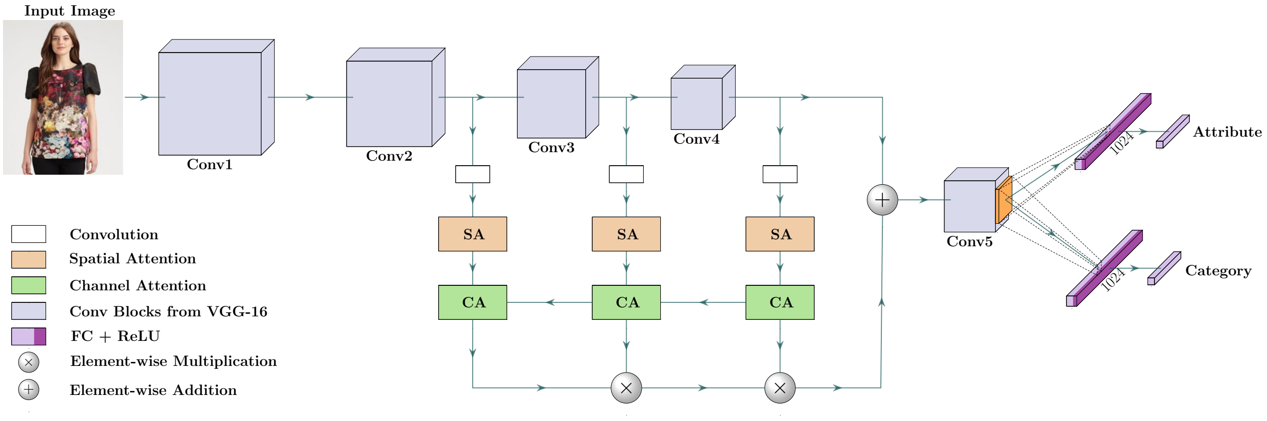


Figure 2: Illustration of the proposed framework. Conv1-5 denote different levels of base network. Lateral connections consist of two attentions: Spatial attention and channel attention.

3 Methodology

3.1 Multi-staged feature attention

Most importantly, huge semantic gap between the low-level features and the high-level semantic features hustle on feature representation in convolutional neural networks. From this inspiration, we construct a multi-staged feature-attentive network based on VGG-16 pretrained model. It is useful to adopt recalibrated multiscale features from multilevel of architecture. This can not only help the network to locate more characteristic and informative features but also the high-level features are able to guide in rectifying low-level features. In a feature pyramid network (FPN) [16], a top-down structure is connected with the bottom-up backbone by lateral connections so that the high-level semantic features can be passed to the low-level feature maps. Furthermore, multiscale predictions are made from the top-down architecture at all scales. Accordingly, we build a top-down feature generation like FPN. However, features are transformed and processed by spatial and channel attentions respectively, before passing them to the next level. Generally, the top-level features of deep neural networks contain rich semantic information while having small resolution with larger receptive field that is useful to recognize patterns in large scale. The features of initial few layers manifest rich spatial information which represent the simple understanding of clothing items by neural networks which have large in resolution and smaller receptive field. Therefore, the larger scale feature is useful to find small patterns. However, not all features are useful to processed further. It is necessary that the informative features provided must be selected which can be executed well focusing on both spatial and channel-wise effective feature rendition.

Attention mechanisms are accompanied by various structures of spatial, semantic and/or channel information and are widely used in computer vision tasks such as medical image segmentation [39], image captioning [36], and regression networks [40]. Generally, important features for precise identification are obtained based on the spatial information so that the spatial attentive features are extracted on salient regions. Similarly, each filter in convolution operation works on a pattern and each channel in the feature map is activated by the response of the corresponding filter. As a result, channel attentive features are constructed on important semantic attributes. Hence, combining both spatial and channel attentions into a network expects to be effective in many cases whereas tiny regions in images are concentrated as areas of interest in fashion classification. In accordance with this ability, instead of direct feed-forward of feature maps in lateral connections of multi-staged feature-attentive network, we concatenated spatial and channel attentions. The proposed architecture is illustrated in Figure 2.

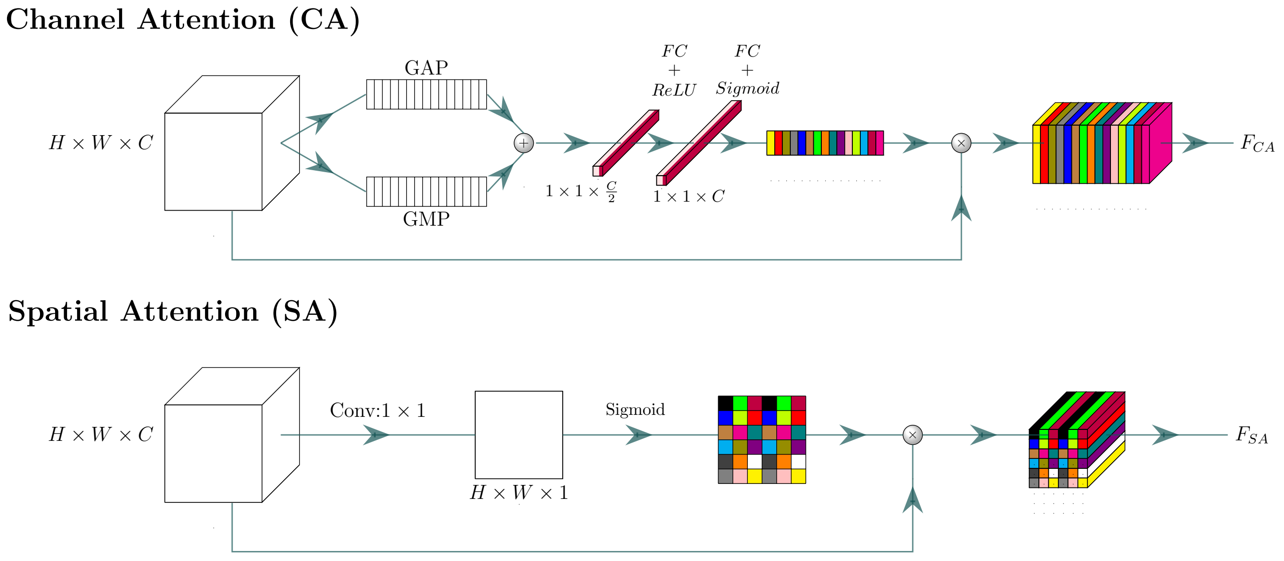


Figure 3: Structure of the spatial and channel attentions used in the lateral connections of multi-staged feature attention. Top block denotes the illustration of channel-wise attention (CA) which is initially achieved by adding Global Average Pooling (GAP) and Global Max Pooling (GMP) to generate channel attentive features (F_{CA}) where bottom block stipulates spatial attention (SA) through a 1×1 convolution to generate spatial attentive features (F_{SA}).

Since different levels of multiscale feature map is passed through network, we fixed the dimension of feature map at each level and all extra convolutional layers as same which can meet the demand of a fixed number of feature map channels, and also reduces the memory consumption while keeping up better performance. Therefore, we first apply a convolution to all stages expect the first stage to set the dimension of feature maps as 512 followed by a spatial attention block to transform the feature representation. Then a channel attention block is incorporated in the gating point, which intakes the concatenated feature maps from different stages of network. This creates a fusion of multilevel of features as an added advantage in representing features effectively. The refined feature maps obtained from all stages are element-wise multiplied. Finally, the multi-staged attentive feature map is added to the features derived from Conv4 block of base network to be followed with the rest of the layers.

3.1.1 Spatial attention

The spatial attention makes the network earn benefits from the features on the most significant areas of fashion clothing items. To create spatial attention map, channel squeeze is implemented with a 1×1 convolution kernel. Then sigmoid function is applied to the convolved features to get a normalised values for feature map. Output of this operation represents the combination of all channel information in corresponding spatial locations. The structure is depicted in Figure 3. The created attention map is then multiplied element-wise with the output feature map from the corresponding level to get the spatial-wise weighted feature maps (F_{SA}). According to the observation, only the cloth area in the image significantly contributes the most on predicting attribute patterns, encoding this knowledge can help the model focus on the target region and learn a better spatial representation.

3.1.2 Channel attention

In this phase, first we concatenate feature map F_{SA} obtained from spatial attention and upsampled feature map from corresponding next level of network. Then the dimension of output feature map is reduced to half the size. To produce channel attention map, initially, we squeeze spatial features by employing global average pooling (GAP) and global max pooling (GMP) simultaneously, then both the squeezed maps values are added together as shown in Figure 3. Though, global average pooling is wisely used to create channel attention maps, we further go into focusing important points to effectively support attribute prediction by selecting the maximum response point in feature maps. Therefore, we use GAP and GMP together to extract spatial-wise contextual information. After that, two fully connected (FC) layers followed with sigmoid activation are utilised to generate channel attention map. The dimensionality reduction ratio is set to 2 in the first FC layer and then it is again resized to original input size for the next FC layer. At this point, residual connection is appended to generate channel-wise weighted feature maps (F_{CA}) from output features obtained from spatial attention instead of concatenated feature input by element-wise multiplying the channel attention map.

4 Experiments

4.1 Dataset

DeepFashion-C dataset[*]. We evaluate our proposed framework for category classification and attribute prediction on this eminent dataset of fashion clothes released in 2016. It contains 289,222 annotated fashion clothing items. The dataset consists of evaluation status for every image as 'train', 'val', and 'test' and is also provided with bounding box of upper left and lower right point coordinates. The dataset statistics is summarised in Table 1. Example images from the dataset for different categories and distinct attribute types are shown in Figure 4.

Table 1: DeepFashion-C dataset statistics

# images	289,222
# categories	50 including upper, lower, and full body clothing items
# attribute types	5 (texture, fabric, shape, part, and style) including 1000 distinct attributes
# landmarks	8 (right/left collar, right/left hem, right/left waistline, and right/left sleeve)
Resolution	The long side of images are resized to 300
Training / validation / testing	209222 / 40000 / 40000 images

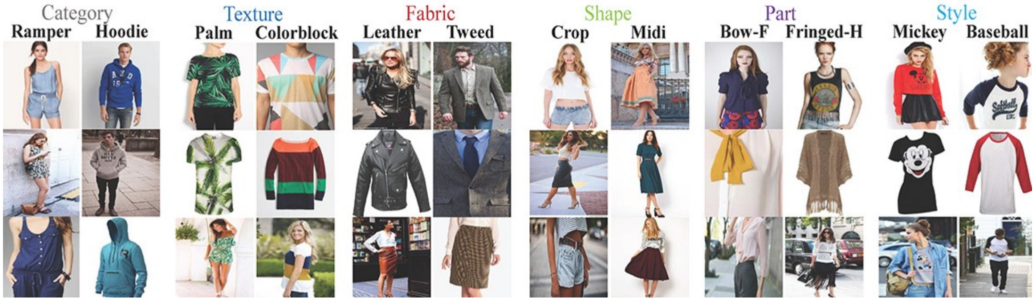


Figure 4: Example images from the DeepFashion-C dataset for different categories and distinct attribute types.

[*] http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

4.2 Network Architecture

Our method is built upon VGG-16 pretrained architecture. Learned feature maps through feature-attentive block incorporated until Conv4 block is then fed into the Conv5 block of VGG-16. We subdivide the fully connected layers as two branches incorporating multitask network with attribute prediction and category classification. Each branch consists of one FC layer with the size of $1 \times 1 \times 1024$. Further, we use standard cross-entropy loss ($Loss_c$) and weighted cross-entropy loss ($Loss_a$) to train category classification and attribute estimation, respectively. To facilitate multitask learning throughout the model, we utilise a weighted loss combination to calculate the total loss which is used to optimise the model weights. The combined loss is as follows,

$$L = w_c \times Loss_c + w_a \times Loss_a \tag{1}$$

where w_c and w_a are the weights for the losses computed from category and attribute branches, respectively. The weighted cross-entropy loss to predict attributes incorporates the weighting factors by the ratio of the numbers of positive and negative samples in the training set as instructed in [4].

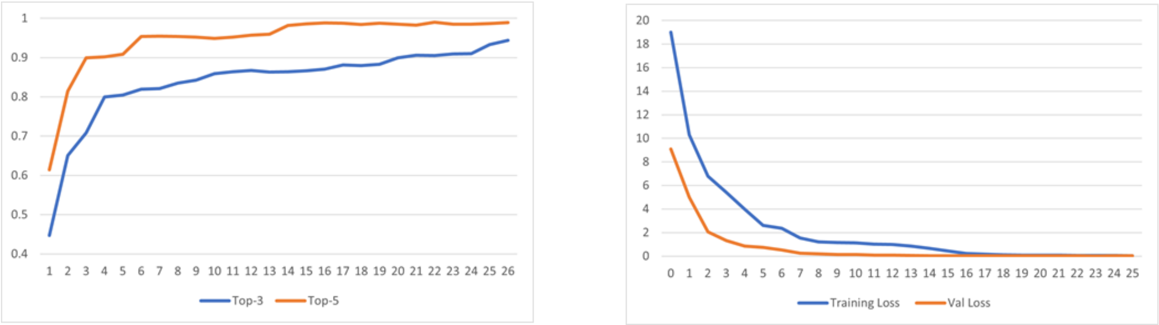
4.3 Quantitative Results

4.3.1 Experimental Setup

Our model is implemented using PyTorch and optimised using Adam optimiser [41] on Tesla P100-16GB system. 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 for every two epochs. We split the training and testing images as presented in the DeepFashion-C dataset where the splits remain as [training/testing/validation: 209,222/40,000/40,000 images]. We crop all the images using the annotated bounding box labels and then resize them into 224×224. The performance of our proposed model in training process is depicted as loss and accuracy curves in Figure 5.

4.3.2 Performance Evaluation

For category classification and attribute prediction, we applied top-*k* classification accuracy and top-*k* recall rate, respectively. We compared the performance with ten recently reported works [4, 5, 7, 8, 14, 27, 30, 31, 42, 43] in fashion analysis. As shown in the Table 2, our model slightly outperforms state-of-the-art approaches using supervised learning in fashion clothing classification. Figure 7 shows some of the example images and corresponding top-5 list of predicted attributes and category. 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 all tables. Our model shows 92.11 and 96.67 for top-*k* accuracy, and for the attribute prediction, overall top-*k* recall achieves 54.92 and 63.18 where k = 3 and 5, respectively. Compared to other works reported in fashion analysis, our model



(a) (b)

Figure 5: (a) Category classification accuracies in top-3 and top-5 for different number of epochs. (b) Training losses for the number of epochs.

Table 2: Performance comparison of category and attributes classification methods on test set using top-*k* accuracies

Methods	**Category**		**Attribute**	
	top-3	top-5	top-3	top-5
Chen *et al.* (2012) [44]	43.73	66.26	27.46	35.37
Huang *et al.* (2015) [42]	59.48	79.58	42.35	51.95
Liu *et al.* (2016) [4]	82.58	90.17	45.52	54.61
Corbiere *et al.* (2017) [30]	86.30	92.80	23.10	30.40
Wang *et al.* (2018) [7]	90.99	95.78	51.53	60.95
Liu and Lu *et al.* (2018) [5]	91.16	96.12	54.69	**63.74**
Lee *et al.* (2019) [8]	91.37	95.26	47.70	57.28
Cho *et al.* (2019) [31]	91.24	95.68	-	-
Zhang *et al.* (2020) [27]	91.99	96.44	50.58	60.43
Shajini *et al.* (2020) [14]	91.02	96.20	51.89	62.04
Ours	**92.11**	**96.67**	**54.92**	63.18

Table 3: Quantitative results for attribute prediction on DeepFashion-C dataset using top-*k* recall

Methods	**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
Chen *et al.* [44]	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68	27.46	35.37
Huang *et al.* [42]	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.14	66.11	71.36	42.35	51.95
Liu *et al.* [4]	37.46	49.52	39.30	49.84	39.47	48.59	44.13	54.02	66.43	73.16	45.52	54.61
Corbiere *et al.* [30]	53.60	63.20	39.10	48.80	50.10	59.50	38.80	48.90	30.50	38.30	23.10	30.40
Wang *et al.* [7]	50.31	65.48	40.31	48.23	53.32	61.05	40.65	56.32	68.70	**74.25**	51.53	60.95
Liu and Lu[5]	56.17	65.83	43.20	53.52	58.28	67.80	46.97	57.42	**68.82**	74.13	54.69	63.74
Lee *et al.* [8]	56.95	66.24	44.03	54.21	56.87	66.25	44.89	55.15	33.98	42.21	47.70	57.28
Shajini *et al.* [14]	56.88	65.16	36.49	44.41	51.88	60.71	47.25	59.97	54.21	67.23	51.89	62.04
Zhang *et al.* [27]	58.52	68.19	**46.44**	**57.02**	61.86	70.81	49.82	60.36	34.40	43.44	50.58	60.43
Ours	**60.02**	**68.84**	43.16	54.69	**61.97**	**71.17**	**49.95**	**60.72**	68.11	73.20	**54.92**	**63.18**

shows top-3 and top-5 accuracy increased by 1% and 0.5%, respectively, in terms of category classification. Similarly, the performance for clothing attribute prediction reveal that our model increases the top-3 recall rate by nearly 3%. The detailed results for each attribute type is summarised in Table 3. Further, the top-5 prediction recall rate for each attribute type with two highest representative attribute values is plotted in Figure 6. Our proposed model shows 36ms of the inference time. We also investigated the impact of semi-supervised learning approach using a knowledge-sharing architecture and the experimental details are discussed in the ablation study.

4.4 Ablation Study

4.4.1 Semi-supervised learning approach

We experiment the major tasks in fashion analysis that extracts feature representation of clothing categories as well as its attributes in semi-supervised manner. The main goal is to structure a framework which can utilise labelled and unlabelled samples together so that we can avoid additional training. The main concern in choosing SSL approach is inspired in various architectures by minimising the entropy of the prediction function that takes the performance to the next level since the classification cost is not defined for unlabelled samples.

Collection of unlabeled dataset. For examining the semi-supervised approach, six publicly available datasets are merged together to create large-scale unlabelled dataset including $400K$ images. We made use of all clothing images from in-shop clothes retrieval and consumer-to-shop datasets which are subsets of large-scale

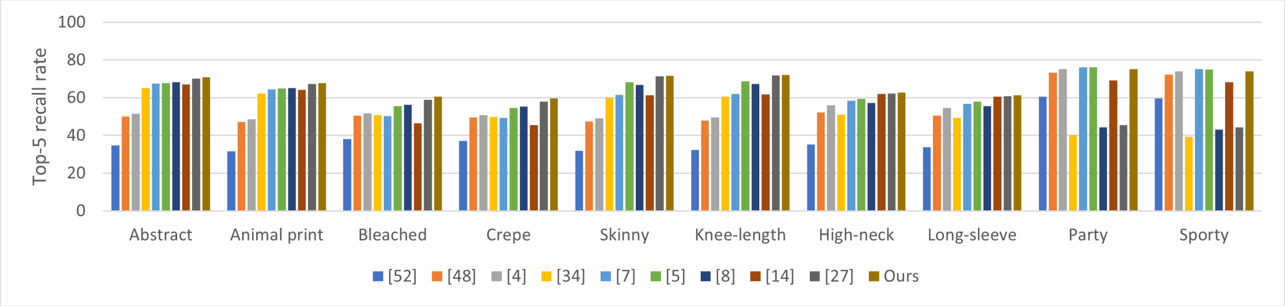


Figure 6: Per-attribute prediction performance: Two representative attribute values for each attribute type in order of texture, fabric, shape, part, and style.

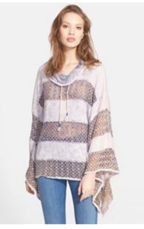
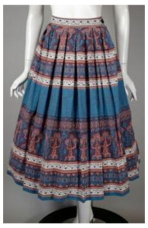
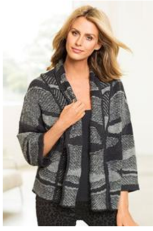
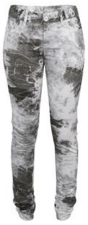
							
Category	Top	Category	Skirt	Category	Jacket	Category	Jeans
Attributes	Boxy Abstract print French terry Dot Knit	Attributes	Southwestern Print Tie-Dye Pleated Scuba	Attributes	Longline Utility Draped Abstract-patterned Open-knit	Attributes	Ankle Bleached Low-Rise Skinny Slim-leg

Figure 7: The results of proposed model which shows the top-5 list of all attributes together with predicted category. If the actual ground truth attribute is listed in top-5, it is marked in green and others in red.

DeepFashion dataset [4] due to its same domain characteristics. We also utilised manually selected samples from the iMaterialist fashion attribute dataset [45], Fashion10000 [46], Fashion data [47], and clothing attribute dataset [44]. The brief statistics of benchmark datasets for fashion analysis is given in Table 4. Besides, Figure 9 shows examples of the clothing images included in the DeepFashion-C dataset and combined unlabelled dataset. According to the decorum of unlabelled dataset, all images are resized to 224×224 relative to their largest center.

We examine a collaborative learning architecture using convolutional neural network to experiment the semi-supervised learning approach with weighted loss minimisation among shared network to learn discriminative clothing representation in utilising unlabelled data. As a result, our SSL model yields significant improvement to the clothing analysis architectures. The semi-supervised framework consists of networks called Teacher-Student (T-S) pair. The way this model selects the pseudo label is defined by the maximum probability score, so that the good teacher model plays a major part in T-S pair model in performance. The proposed multi-staged feature-attentive network is employed as the teacher model. Furthermore, the student model represents straightforward network by which the complexity of the training process is reduced and the unlabelled samples are able to extract the insight of clothing items in an unambiguous way. The student model is constructed by integrating additional spatial-channel attention at the end of residual blocks of Conv3 of pretrained ResNet-18 as shown in Figure 8. In this phase, the student model grasps the structure of spatial and channel attention

Table 4: Summary of datasets used in fashion analysis

Datasets	Authors	# images	# categories
Clothing attribute	Chen *et al.* (2012) [44]	1,856	7
Fashion data	Lukas *et al.* (2013) [47]	590,234	-
Fashion10000	Babak *et al.* (2014) [46]	32,398	470
Fashion Landmark Detection	Liu *et al.* (2016) [22]	123,016	-
DeepFashion-C	Liu *et al.* (2016) [4]	289,222	50
Fashion200K	Xintong *et al.* (2017) [48]	209,544	5
Unconstrained landmark	Yan *et al.* (2017) [49]	30,000	-
CatalogFashion-10x	Heilbron *et al.* (2019) [50]	1,000,000	43
iMaterialist fashion attribute	Sheng *et al.* (2019) [45]	1,000,000	105
DeepFashion2	Ge *et al.* (2019) [51]	491,000	13

similar to teacher model but the channel attention only uses the global average pooling to generate attention map. The student model is modified further by adding two output layers for the multitask learning of category classification and attribute prediction leading with a FC layer size of $1 \times 1 \times 1024$. Equal proportion of labelled and unlabelled samples are taken into the teacher model and pseudo labels are picked by the teacher model for unlabelled samples which gives maximum probability among category scores. It is commendable that the ratio of labelled and unlabelled samples taken for training to be equal so that it creates the effective balance of semi-supervised learning in multitask. Further, instead of selecting top *k* labels introduced in [20], we select the best one due to the concrete performance of selected teacher model. Simultaneously, teacher model calculates the losses $Loss_l$ and $Loss_{ul}$ for labelled and unlabelled samples, respectively. Then the unlabelled samples with the predicted pseudo labels are fed into the student model for training and is tuned to optimise the student model weights.

We experimented the weighted minimisation by utilising task-dependant uncertainty of each paired models which helps to find the optimum balance between losses. It is inspired from the work reported in [52] where the classification likelihood extends to a scaled version of the model output for balancing losses. Let the loss for cross-entropy for y be,

$$L(w) = -log(\text{Softmax}(y, f^w(x), \sigma)) \tag{2}$$

and optimise with respect to w as well as the noise parameter σ. In the main transition of loss function, an explicit assumption is made as,

$$\frac{1}{\sigma^2} \sum exp(\frac{1}{\sigma^2} f^w(x)) \approx (\sum exp(f^w(x))^{\frac{1}{\sigma^2}} \tag{3}$$

The weighting coefficients are turned in to trainable parameters using this uncertainty measure where they are also optimised during training time. This tends to the definition of combined loss as follows,

$$L = \frac{1}{\sigma_1^2} \times (Loss_l + Loss_{ul}) + \frac{1}{\sigma_2^2} \times Loss_{st} + \log \sigma_1 + \log \sigma_2 \tag{4}$$

where, σ_1 and σ_2 are the noise parameters of supervised and unsupervised losses, respectively. We used the cross-entropy and weighted cross-entropy losses for clothing category classification and attribute prediction, respectively, in which the $Loss_l$ and $Loss_{st}$ are together used to minimise the error in joint learning. In addition, we define the losses $Loss_{ul}^c$ and $Loss_{ul}^a$ for unlabelled data samples as an entropy of the prediction

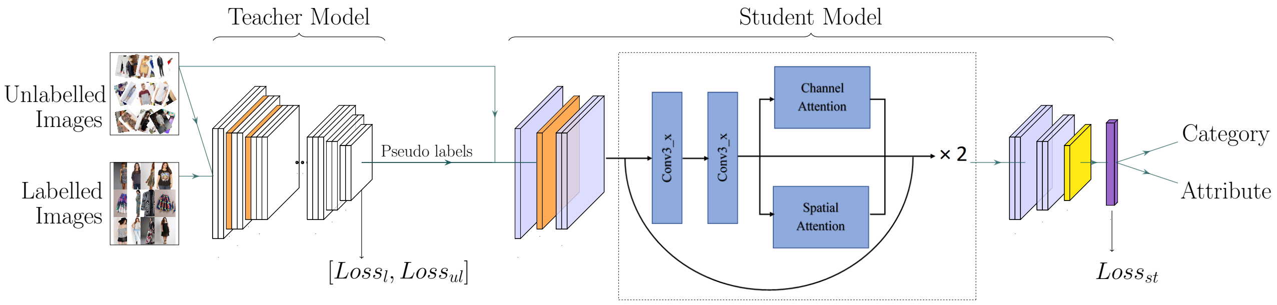


Figure 8: The illustration of the proposed semi-supervised architecture. The predicted labels with high score from teacher model (block in left) will be assigned for the unlabelled samples to further train the student model (block in right).

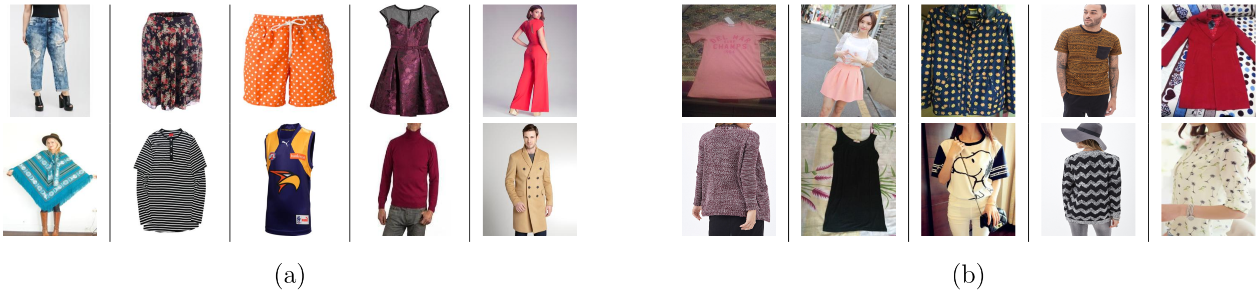


Figure 9: Starting from left to right, the sub parts shows: (a) example labelled images, and (b) example images from the unlabelled dataset for some different categories.

functions $h(x)^c$ and $h(x)^a$ which quantifies the level of uncertainty of the network on unlabelled samples, respectively. The notation c denotes category classification where a denotes attribute prediction.

$$Loss_{ul}^{c} = -\frac{1}{M}\sum_{k=1}^{M} h(x)_k^c \times log(h(x)_k^c + 1e^{-5}) \tag{5}$$

$$Loss_{ul}^{a} = -\frac{1}{M}[\sum_{n=1}^{M} w_p \times h(x)_n^a \times log(h(x)_n^a + 1e^{-5}) \quad + \quad w_n \times (1 - h(x)_n) \times log(1 - h(x)_n + 1e^{-5})] \tag{6}$$

where M is the number of training samples along with the weights w_p and w_n for the ratio of positive and negative samples in attribute prediction, respectively. For category classification and attribute prediction, we

Table 5: Experimental results for category and attributes classification using supervised/semi-supervised learning approach on test set. The results are top-k accuracies and top-k recall, respectively

Methods	Category		Attribute	
	top-3	top-5	top-3	top-5
Supervised (ours)	**92.11**	96.67	**54.92**	**63.18**
Semi-supervised [53]	91.06	96.35	51.22	61.63
Semi-supervised (ours)	91.89	**96.71**	54.66	62.43

applied top-k classification accuracy and top-k recall rate, respectively. We separately evaluated the performance of the student model in fashion clothes category classification using supervised learning. The learning rate is initially set to 0.0001 then reduced by factor of ten when validation loss plateau. It achieves 90.07% and 93.95% for top-k accuracies where k=3 and 5, respectively for category classification. At the end of the training process, the teacher and student models have uncertainty measures of 2.48 and 19.21 for the value of σ_1 and σ_2 which results in effective weighting of losses approximately 1:0.16 in collaborative learning. As shown in Table 5, the proposed semi-supervised learning approach called T-S pair model with our proposed teacher architecture achieves commendable results. Compared to supervised learning approach, the performance increases by 0.04% of top-5 accuracy for category classification and for attribute prediction it shows the relative performance. Besides, compared to [4, 5, 7, 8, 14, 27, 30, 31, 42, 43, 53], the T-S pair model outperforms in top-5 accuracy for category classification and shows relative performance in top-3 and top-5 recall rate for attribute prediction. For the experiments, we used Tesla P100-16GB system and the model shows 56ms of inference time.

4.4.2 Evaluation on multi-staged feature extraction

In these experiments, we performed an extensive study on components of the proposed MHGAN. The proposed framework mainly relies on baseline which is a pretrained VGG-16 model. We further studied the impact of the combination of multilevel features and spatial-channel-wise information to effectively classify clothing items and predict their attributes.

In order to improve the performance, an attention mechanism is indulged in between the Conv4 and Conv5 blocks of the baseline model. The spatial attentive features (F_{SA}) are then fed into channel attention to extract channel attentive features (F_{CA}). The structure of both attentions increases the performance. The spatial attention (SA) together with channel attention (CA) makes full use of the characteristics of CNN and can produce compelling image features, thus the performance has been improved [14, 36, 40]. Generally, important features for precise identification are obtained based on the spatial information so that the spatial attentive features are constructed on salient regions. Therefore, combining both spatial and channel attentions into a network is expected to be effective in many cases, whereas tiny regions in images are concentrated as regions of interest in fashion classification.

The focus is then turned into a multilevel feature enhancement as the introduction of FPN and its effective performance in computer vision tasks attracted the way of fashion analysis. In these experiments, we extend the attention mechanism into a multi-staged feature extraction architecture accompanying experiments on the use of spatial and channel-wise attentions. The SA block pays attention to the global area related to semantic information of clothing items and supports the overall performance in classification. In accordance with the ability of SA and CA blocks, instead of direct feed-forward of feature maps in lateral connections of multi-staged feature-attentive network, we concatenated spatial and channel attentions for the study. Output of this operation represents the combination of all channel information in corresponding spatial locations. 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 for every two epochs. The experimental results are summarised in Table 6. We evaluate the overall in-depth performance of our framework structure in significant ways:

(i) Training the baseline model which is a pretrained VGG-16 network,

(ii) Training the baseline model with spatial and channel-wise attentions structured in parallel at a confined level,

(iii) Training the model with spatial and channel-wise attentions structured in multilevel where channel attention takes input from GAP, and

(iv) Training the model with spatial and channel-wise attentions structured in multilevel where channel attention takes input from combined GAP and GMP.

Table 6: Comparison of the ablation study on category and attribute prediction tested on DeepFashion-C dataset using top-k accuracies and top-k recall, respectively

Methods	**Category**		**Attribute**	
	top-3	top-5	top-3	top-5
Baseline (VGG-16)	82.64	89.08	35.96	54.07
Baseline + SA + CA (single-staged)	84.80	91.47	36.43	55.41
Baseline + SA + CA (multi-staged, GAP)	91.12	94.34	50.49	59.18
Baseline + SA + CA (multiple-staged, [GAP, GMP])	**92.11**	**96.67**	**54.92**	**63.18**

5 Conclusion

In this paper, we presented a multi-staged feature-attentive model for learning robust fashion classification. The effectiveness of the proposed architecture has been verified by experiments, indicating clothing classification and prediction of its corresponding attributes that are more robust by learning feature representation through combining low-level and high-level features of convolutional neural network. This approach recalibrates the feature extraction by focusing on contextual information from different levels of network through more attention on both spatial regions and channel values obtained using different filter operations. We also examined a semi-supervised teacher-student pair learning approach to escalate the performance where the annotation of fashion clothing images became more dreadful. Our experiments show qualitative performance in category classification and attribute prediction that slightly outperforms many approaches reported in recent works.

References

[1] M. Chen, Y. Qin, L. Qi, and Y. Sun, "Improving fashion landmark detection by dual attention feature enhancement," in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3101–3104, doi: https://doi.org/10.1109/ICCVW.2019.00374.

[2] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017, pp. 2268–2274, doi: https://doi.org/10.1109/ICCVW.2017.266.

[3] B. Quintino Ferreira, J. P. Costeira, R. G. Sousa, L. Gui, and J. P. Gomes, "Pose guided attention for multi-label fashion image classification," in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3125–3128, doi: https://doi.org/10.1109/ICCVW.2019.00380.

[4] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1096–1104, doi: https://doi.org/10.1109/CVPR.2016.124.

[5] J. Liu and H. Lu, "Deep fashion analysis with feature map upsampling and landmark-driven attention," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 30–36, doi: https://doi.org/10.1007/978-3-030-11015-4_4.

[6] Y. Li, S. Tang, Y. Ye, and J. Ma, "Spatial-aware non-local attention for fashion landmark detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 820–825, doi: https://doi.org/10.1109/ICME.2019.00146.

[7] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, “Attentive fashion grammar network for fashion landmark detection and clothing category classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4271–4280, doi: https://doi.org/10.1109/CVPR.2018.00449.

[8] S. Lee, H. Eun, S. Oh, W. Kim, C. Jung, and C. Kim, “Landmark-free clothes recognition with a two-branch feature selective network,” *Electronics Letters*, vol. 55, no. 13, pp. 745–747, 2019, doi: https://doi.org/10.1049/el.2019.0660.

[9] S. Lee, S. Oh, C. Jung, and C. Kim, “A global-local embedding module for fashion landmark detection,” in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3153–3156, doi: https://doi.org/10.1109/ICCVW.2019.00387.

[10] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141, doi: https://doi.org/10.1109/CVPR.2018.00745.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803, doi: https://doi.org/10.1109/CVPR.2018.00813.

[12] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, “Diversified visual attention networks for fine-grained object classification,” *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017, doi: https://doi.org/10.1109/tmm.2017.2648498.

[13] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” in *Proceedings of Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 2017–2025.

[14] M. Shajini and A. Ramanan, “An improved landmark-driven and spatial–channel attentive convolutional neural network for fashion clothes classification,” *The Visual Computer*, vol. 37, no. 6, pp. 1517–1526, 2020, doi: https://doi.org/10.1007/s00371-020-01885-7.

[15] C.-Q. Huang, J.-K. Chen, Y. Pan, H.-J. Lai, J. Yin, and Q.-H. Huang, “Clothing landmark detection using deep networks with prior of key point associations,” *IEEE Transactions on Cybernetics*, vol. 49, no. 10, pp. 3744–3754, 2018, doi: https://doi.org/10.1109/TCYB.2018.2850745.

[16] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944, doi: https://doi.org/10.1109/CVPR.2017.106.

[17] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, “Virtual adversarial training: a regularization method for supervised and semi-supervised learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[18] L. Samuli and A. Timo, “Temporal ensembling for semi-supervised learning,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017, pp. 1–6.

[19] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 1195–1204.

[20] I. Z. Yalniz, H. Jgou, K. Chen, M. Paluri, and D. Mahajan, “Billion-scale semi-supervised learning for image classification,” *arXiv preprint arXiv:1905.00546*, 2019.

[21] D.-H. Lee *et al.*, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, pp. 896–912.

[22] L. Ziwei, Y. Sijie, L. Ping, W. Xiaogang, and T. Xiaoou, “Fashion landmark detection in the wild,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 229–245.

[23] Y. Hu, L. Xiao, Y. Wang, and Y. Jin, “Fashion pose machine for fashion landmark detection,” in *Proceedings of the SPIE International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836, 2018, pp. 1–5.

[24] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao, and L. Lin, “Layout-graph reasoning for fashion landmark detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2937–2945.

[25] M. Chen, H. Ying, Y. Qin, L. Qi, Z. Gan, and Y. Sun, “Adaptive graph reasoning network for fashion landmark detection,” in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2020, pp. 2672–2679.

[26] P. Li, Y. Li, X. Jiang, and X. Zhen, “Two-stream multi-task network for fashion recognition,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3038–3042, doi: https://doi.org/10.1109/ICIP.2019.8803394.

[27] Y. Zhang, P. Zhang, C. Yuan, and Z. Wang, “Texture and shape biased two-stream networks for clothing classification and attribute recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 538–13 547.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[29] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[30] C. Corbire, H. Ben-Younes, A. Ram, and C. Ollion, “Leveraging weakly annotated data for fashion image retrieval and label prediction,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2268–2274, doi: https://doi.org/10.1109/ICCVW.2017.266.

[31] H. Cho, C. Ahn, K. M. Yoo, J. Seol, and S. Lee, “Leveraging class hierarchy in fashion classification,” in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3197–3200, doi: https://doi.org/10.1109/ICCVW.2019.00398.

[32] J. Wehrmann, R. Cerri, and R. Barros, “Hierarchical multi-label classification networks,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 5075–5084.

[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458, doi: https://doi.org/10.1109/CVPR.2017.683.

[34] D. Fan, W. Wang, M. Cheng, and J. Shen, “Shifting more attention to video salient object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8546–8556, doi: https://doi.org/10.1109/CVPR.2019.00875.

[35] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, “ABC-CNN: An attention based convolutional neural network for visual question answering,” *arXiv preprint arXiv:1511.05960*, 2016.

[36] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "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)*, 2017, pp. 6298–6306, doi: https://doi.org/10.1109/CVPR.2017.667.

[37] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2019, doi: https://doi.org/10.1109/TPAMI.2018.2840724.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[39] Y. Chen, K. Wang, X. Liao, Y. Qian, Q. Wang, Z. Yuan, and P.-A. Heng, "Channel-unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation," *Frontiers in genetics*, vol. 10, pp. 1–13, 2019.

[40] J. Gao, Q. Wang, and Y. Yuan, "SCAR: Spatial-/channel-wise attention regression networks for crowd counting," *Neurocomputing*, vol. 363, pp. 1–8, 2019, doi: https://doi.org/10.1016/j.neucom.2019.08.018.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the Third International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.

[42] 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)*, 2015, pp. 1062–1070.

[43] 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)*, 2015, pp. 3343–3351.

[44] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer Berlin Heidelberg, 2012, pp. 609–623.

[45] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, H. Adam, M. R. Scott, and S. Belongie, "The iMaterialist fashion attribute dataset," in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3113–3116.

[46] B. Loni, L. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson, "Fashion 10000: An enriched social image dataset for fashion and clothing," in *Proceedings of the 5th ACM Multimedia Systems Conference (MMSys)*, 2014, pp. 41–46, doi: https://doi.org/10.1145/2557642.2563675.

[47] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer Berlin Heidelberg, 2013, pp. 321–335.

[48] 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 International Conference on Computer Vision (ICCV)*, 2017, pp. 1472–1480, doi: https://doi.org/10.1109/ICCV.2017.163.

[49] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Unconstrained fashion landmark detection via hierarchical recurrent transformer networks," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 172–180.

[50] F. C. Heilbron, B. Pepik, Z. Barzelay, and M. Donoser, "Clothing recognition in the wild using the amazon catalog," in *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3145–3148, doi: https://doi.org/10.1109/ICCVW.2019.00385.

[51] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5337–5345.

[52] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[53] M. Shajini and A. Ramanan, "A knowledge-sharing semi-supervised approach for fashion clothes classification and attribute prediction," *The Visual Computer*, pp. 1–11, 2021.