A Multiscale Contextual Technique for Fashion Clothes Landmark Localisation

Shajini Majuran and Amirthalingam Ramanan
shayu.kiri@gmail.com, a.ramanan@univ.jfn.ac.lk

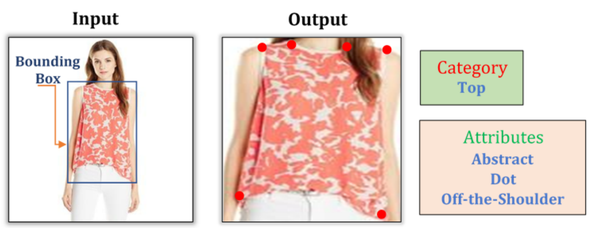
SLSWCS Student Workshop on Computer Science

Introduction

Nowadays, researchers pay attention towards fashion clothes classification using machine learning and deep learning in order to make people's lives better with help of key factors such as image recognition, process of massive data, and enhance the facility of personalisation. In this regard, localising landmarks (collar, sleeves, waistline, and hem) in clothes can be an important attention in clothing classification. *Landmark localisation* involves global integration of information and 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 of feature maps.

Applications: 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 social media.

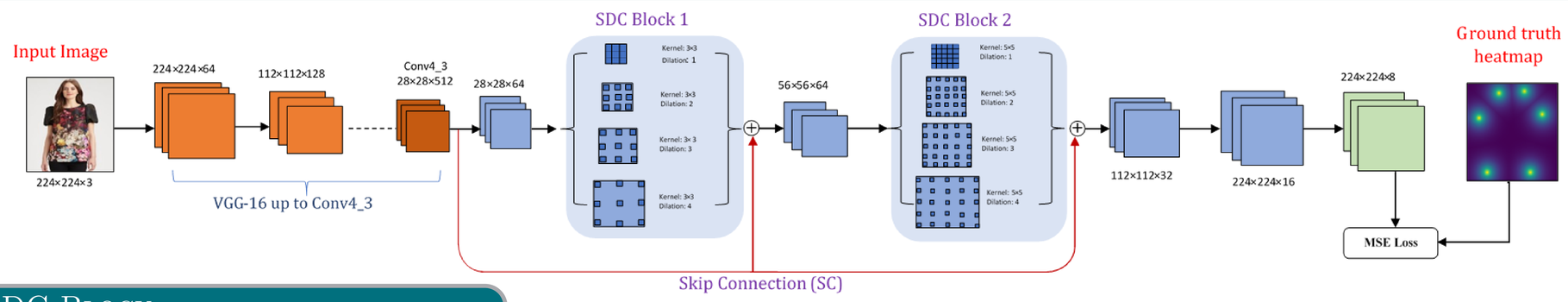
Problem Specification



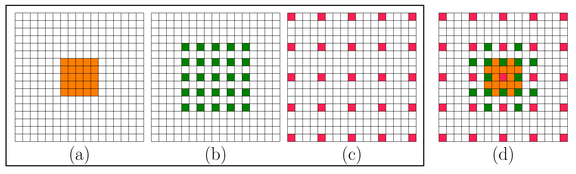
Contribution

1. A scale-driven structure through deep neural network is proposed using *stacked dilated convolutional (SDC) block* which utilises parallel convolutions with different dilation rates.
2. A much simplified framework for better fashion clothing landmark detection is proposed with the help of *upsampling feature maps* which enhance the performance by transposing to high resolution feature maps.

Methodology



SDC Block



Stacked Dilated Convolution (SDC) Block: Dilated convolution is a convolution, applied to the input feature map with designated distance between the kernel points determined by the dilation coefficient $\mathcal{D}$. (a), (b), and (c) show the example operations using 5×5 kernel with $\mathcal{D}$=1, 2, and 4, respectively. (d) is the overall SDC operation on a feature map with receptive field size of 17×17.

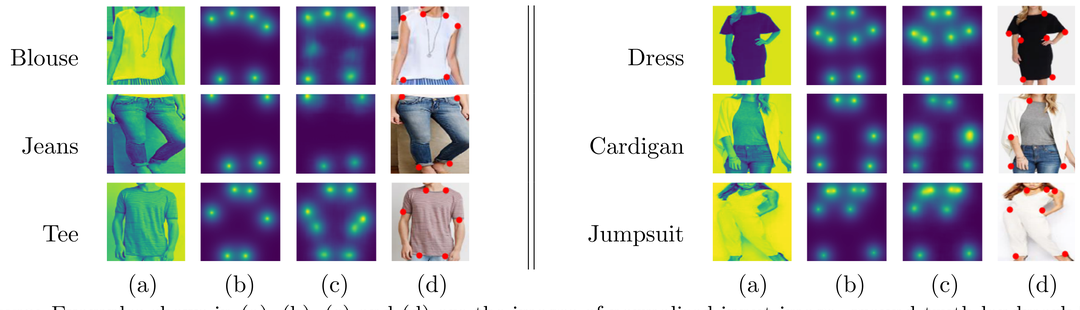


Figure: Examples shows in (a), (b), (c) and (d) are the images of normalised input image, ground truth landmark heatmap, output heatmap, and output landmark locations plotted on original input image for different clothes categories, respectively.

Dataset

DeepFashion-C Dataset [1]

Images: 289,222
Categories: 50
Landmarks: 8
Attributes: 1000

Fashion Landmark Detection (FLD) Dataset [3]

Images: 123,016
Landmarks: 8

Experimental Setup

- The structure is based on *VGG-16* network.
- Each image is cropped using annotated bounding boxes and resized to 224×224.
- (Training,Validation,Testing)=(209222,40000,40000) images for DeepFashion-C, and (83033,19992,19991) images for FLD.
- *Optimiser*: Adam & *Mini-batch size*: 16
- *Learning rate*: 0.0001 & drop by 0.1 while validation plateaus.
- *Loss*: MSE loss
- *Evaluation*: Normalised distance metric

Results

Table I: Experimental reults on the *DeepFashion-C* dataset for landmark localisation in normalised distance metric

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waist	R.Waist	L.Hem	R.Hem	Avg.
Liu *et al.*[1]	0.0854	0.0902	0.0973	0.0935	0.0854	0.0845	0.0812	0.0823	0.0872
Liu *et al.*[3]	0.0628	0.0637	0.0658	0.0621	0.726	0.0702	0.0658	0.0663	0.0660
Yan *et al.*[5]	0.0570	0.0611	0.0672	0.0647	0.0703	0.0694	0.0624	0.0627	0.0643
Wang *et al.*[6]	0.0415	0.0404	0.0496	0.0449	0.0502	0.0523	0.0537	**0.0551**	0.0484
Lu *et al.*[7]	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 reults on the *FLD* dataset for landmark localisation in normalised distance metric

Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waist	R.Waist	L.Hem	R.Hem	Avg.
Liu *et al.*[1]	0.0784	0.0803	0.0975	0.0923	0.0874	0.0821	0.0802	0.0893	0.0859
Liu *et al.*[3]	0.0480	0.0480	0.0910	0.0890	-	-	0.0710	0.0720	0.0680
Yan *et al.*[5]	0.0531	0.0547	**0.0705**	0.0735	0.0752	0.0748	**0.0693**	**0.0675**	0.0672
Ours	**0.0419**	**0.0424**	0.0733	**0.0718**	**0.0648**	**0.0652**	0.0738	0.0755	**0.0635**

Discussion and Conclusion

The output feature maps from each dilated operation of SDC block is concatenated together to make the subsequent convolution layer to learn features from different scales. By adding the SDC blocks, we can produce model which can lead to attain commendable regression results. Dilated convolutions show significant increase in performance for fashion landmark localisation. We demonstrate our experiments on two benchmark datasets and our model outperforms recently proposed state-of-the-art techniques.

References

[1] 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), pp. 1096-1104, 2016.
[2] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 2017.
[3] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. "Fashion Landmark Detection in the Wild." In European Conference on Computer Vision (ECCV), pp. 229-245. 2016.
[4] H. Zhang, Q. Li, Z. Sun, and Y. Liu, "Combining Data-driven and Model-driven Methods for Robust Facial Landmark Detection", In IEEE Transactions on Information Forensics and Security, 13(10), pp.2409-2422, 2018.
[5] 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, pp. 172-180, 2017.
[6] W. Wang, Y. Xu, J. Shen, and S. 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), pp. 4271-4280. 2018.
[7] 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), pp. 30-36, 2018.