ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Appearance-Context aware Axial Attention for Fashion Landmark Detection

Nikhil Kilari, Gaurab Bhattacharya, Pavan Kumar Reddy, J Gubbi and Arpan Pal

Embedded Devices and Intelligent Systems, TCS Research
India

Abstract. Fashion landmark detection is a fundamental task in several fashion image analysis problems. The associated challenges involving non-rigid structures and variations in style and orientation makes it extremely hard to accurately detect the landmarks. In this paper, we propose Appearance-Context network (ACNet), which encapsulates both global and local contextual information extending the axial attention mechanism. We design axial attention augmented local appearance network and introduce a novel Global-Context aware axial attention module which aggregates the global features attending discriminatory cues across height, width and channel axes. The proposed ACNet architecture outperforms existing methods on two large-scale fashion landmark datasets.

1 Introduction

Fashion image analysis has created significant research interest due to its large industrial usage and challenging nature. Fashion landmark detection plays a crucial role in extracting contextual fashion apparel information. These key points serve a pivotal role in clothing category classification, product retrieval and virtual try-on of fashion products. The challenges in this application lie in the large variation and non-rigid structure of the products, which necessitates the incorporation of global and local information of fashion apparels to precisely estimate the landmark location.

In recent years, several large-scale datasets [1, 2] and deep learning models have been proposed to address these challenges [3]-[9]. In [1], authors have considered global and local appearance without any emphasis on important features. Graph-based contextual information has been used in [7], which was improved using dual attention in [8]. Several methods use different form of attention, *e.g.* spatial transformer [3], non-local unit [5, 6], soft attention [4], dual attention [8], *etc.* One common shortcoming to these methods is that none of them is separately learning the local appearance and global context of the fashion apparels. Further, earlier attempts to use of local and global features without appropriate attention modules has led to lack of consistent results across domains [11]. To obtain the contextual features, we should aggregate crucial features across the height, width and channel axes, which is missing in the existing literature.

To address them, we propose ACNet, an appearance-context network for fashion landmark detection. For this, we have created the appearance aware axial attention network for better localization of fashion landmark. Further, we have proposed a novel global-context aware axial attention (GCA) module for

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

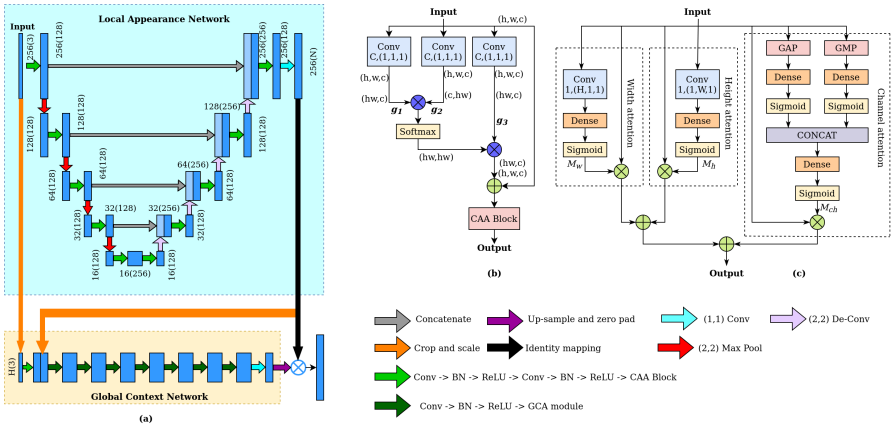

Fig. 1: (a) ACNet architecture; (b) Global-context aware axial attention (GCA) module; and (c) concurrent axial attention (CAA) module. The $Conv\ F, (Y, Z, S)$ represents a *convolution* layer with *kernel size* of (Y, Z), *stride* as S and $filter_{num}$ as F. In (b) and (c), blue multiplication represents vector multiplication, while the green ones represent element-wise multiplication.

the global context network, which can be used flexibly between any intermediate *Conv* layers to extract global context. This block helps in extracting enhanced contextual features with the help of attention across multiple dimensions. Hence, we disentangle global and local features from fashion apparels and capture the landmark locations from local features with the assistance of global features. Our contribution can be noted as follows: (i) a novel architecture (ACNet) is proposed to obtain global and local context of the fashion apparels for effective fashion landmark localization; (ii) we show that the proposed GCA module to extract attention-driven global context leads to better localization of landmarks; and (iii) with extensive experimentation we show that our proposed approach outperforms the state-of-the-methods for the large scale DeepFashion [1] and FLD [2] datasets.

2 ACNet Architecture

Fashion landmarks are correlated to each other with both global and local features being critical for landmark identification. While local features help in understanding the relative positions with respect to other landmarks, global features help in distinguishing the exact location [11]. Therefore, we propose Appearance-Context aware Axial Attention Network (ACNet) which seamlessly disentangles both local and global features that are utilised by the model to identify the landmark location. The local network is a modified version of U-Net [10] in which we stack concurrent axial attention block [12] which extracts features across multiple dimensions using attention. In the global network, we use a series of *Conv* layers each followed by a novel Global Context Aware axial attention

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

(GCA) block, which helps in extracting the global contextual features. The global network takes both input image and output from the composite U-Net to generate the output. The outputs from local network and global network are finally multiplied to generate the final landmark heatmaps. The block diagram of the proposed ACNet is given in Fig. 1 (a).

2.1 Composite U-Net with concurrent axial attention

Extraction of local features plays an important role in identifying the landmarks of fashion apparel. Existing methods attempt to extract local features using spatial transformer [3], local embedding [5] or dual attention [8, 11]. However, these methods do not consider discriminatory cues across height, width and channel axes separately. Contrary to them, we augment the axial attention modules to U-Net [10], popularly used for local feature extraction. For this, we modify the U-Net using the concurrent axial attention (CAA) block [12]. We stack a CAA block after each *Conv* layer in U-Net. This composite U-Net with concurrent axial attention block helps in extracting rich local features of fashion apparel which aid in identifying the fashion landmarks.

2.2 Global context aware axial attention module

The proposed Global context aware axial attention module focuses on extracting key features using self and axial attentions. Consider the input $I(h, w, c)$, the global information is extracted and embedded in $g_1(h, w, c)$ and $g_2(h, w, c)$. We then reshape g_1 and g_2 to (hw, c) and (c, hw), respectively. We multiply the corresponding output to obtain the global spatial correlation between the pixels and activate through softmax, which serves as attention mask for $g_3(h, w, c)$ to extract global features. The extracted features are then added to the input through skip connection. This is fed to the CAA block which in turn extracts key features across height, width and channel dimension using attention.

$$output = \text{CAA}(\text{softmax}(g_1(hw, c) \odot g_2(c, hw)) \odot g_3(hw, c) + I(h, w, c)) \quad (1)$$

2.3 Model architecture

For state-of-the-art comparison with our model, we use U-Net [10] of depth 4, in which concurrent axial attention is employed in local network. In global-context network, similar to earlier work [11], we use 8 convolutional blocks each with 128 filters, with a dilation rate of 1. After each convolutional block in global-context network, we stack a GCA block. Finally we add a (1,1) convolutional block followed by a CAA block. Activation function is ReLU for all convolutional layers.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

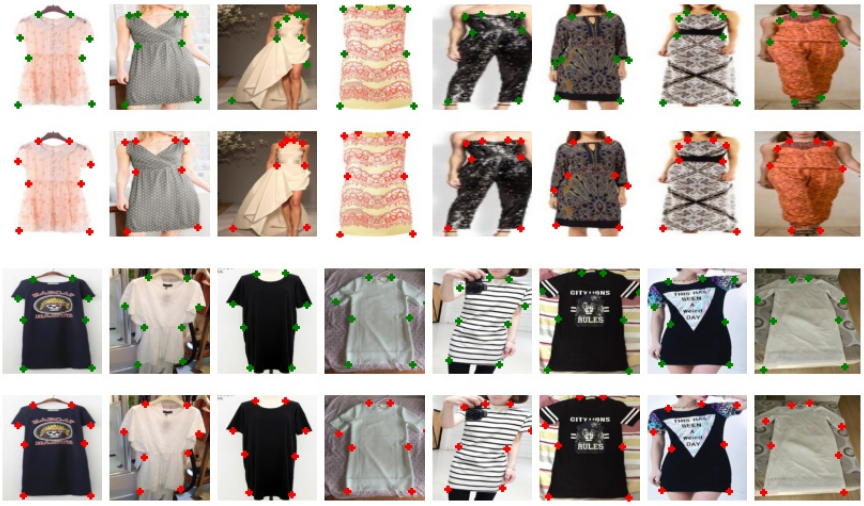

Fig. 2: Visual examples for ACNet architecture using Deep Fashion dataset [1] and FLD [2]. First row contains the images with ground truth landmarks marked with green points for DeepFashion. Second row contains the predicted landmarks with red points for DeepFashion. Third row contains the images with ground truth landmarks marked with green points for FLD. Fourth row contains the predicted landmarks with red points for FLD.

3 Experiments

Training Strategy involves cropping the images according to the bounding box information in the datasets that are then reshaped to (256,256). We use stochastic-gradient descent with a batch size of 5, which are optimised using an Adam optimizer. An initial learning rate of $1.e-5$ is utilised. We use mean-squared-error between the ground truth and generated heatmaps as objective function. While generating the heatmaps, we use a gaussian blur filter with sigma value of 8.

3.1 Comparison with state-of-the-art methods

To compare the performance of proposed ACNet, we have considered two large-scale datasets: DeepFashion [1] and FLD [2]. Our method is compared with several state-of-the-art methodologies, such as DF [1], DFA [2], DLAN [3], FGN [4], GLE [5], SANL [6], LGR [7], AGR [8], and PCL [9], considering normalized error (NE) as performance metric. The results for these two datasets are given in Table 1 and Table 2, respectively. From these results, we observe that our proposed approach outperforms all existing methods in terms of average NE for landmark detection with the benefit of the disentangled local and global features. The visual examples provided in Fig. 2 demonstrate the efficacy of our proposed network. In the third visual example of DeepFashion dataset, the right hemline

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

landmark doesn't match with the ground truth, but the predicted location of the landmark shows that the network is able to localise a meaningful location even when there is a distortion in the fashion apparel.

Table 1: Comparison of normalised errors (NE) on DeepFashion dataset.

Method	L.Cl	R.Cl	L.Sl	R.Sl	L.Wl	R.Wl	L.Hm	R.Hm	Avg
DF [1]	0.0854	0.0902	0.0973	0.0935	0.0854	0.0845	0.0812	0.0823	0.0872
DFA [2]	0.0628	0.0637	0.0658	0.0621	0.0726	0.0702	0.0658	0.0663	0.0660
DLAN [3]	0.0570	0.0611	0.0672	0.0647	0.0703	0.0694	0.0624	0.0627	0.0643
FGN [4]	0.0415	0.0404	0.0496	0.0449	0.0502	0.0523	0.0537	0.0551	0.0484
GLE [5]	0.0312	0.0324	0.0427	0.0434	0.0361	0.0373	0.0442	0.0475	0.0393
SANL [6]	0.0277	0.0282	0.0391	0.0394	0.0297	0.0299	0.0395	0.0401	0.0342
LGR [7]	0.027	0.0116	0.0286	0.0347	0.0307	0.0435	0.0160	0.0162	0.0336
AGR [8]	0.0256	0.0251	0.0318	0.0324	0.0271	0.0286	0.0328	0.0341	0.0297
PCL [9]	0.0203	0.0206	0.0300	0.0304	0.0464	0.0475	0.0150	0.0153	0.0282
Ours	**0.0310**	**0.0215**	**0.0429**	**0.0207**	**0.0317**	**0.0351**	**0.0121**	**0.0164**	**0.0264**

Table 2: Comparison of normalised errors (NE) on FLD dataset.

Method	L.Cl	R.Cl	L.Sl	R.Sl	L.Wl	R.Wl	L.Hm	R.Hm	Avg
DF [1]	0.0784	0.0803	0.0975	0.0923	0.0874	0.0821	0.0802	0.0893	0.0859
DFA [2]	0.0480	0.0480	0.0910	0.0890	-	-	0.0710	0.0720	0.0680
DLAN [3]	0.0531	0.0547	0.0705	0.0735	0.0752	0.0748	0.0693	0.0675	0.0672
FGN [4]	0.0463	0.0471	0.0627	0.0614	0.0635	0.0692	0.0635	0.0527	0.0583
GLE [5]	0.0386	0.0391	0.0675	0.0672	0.0576	0.0605	0.0615	0.0621	0.0568
SANL [6]	0.0296	0.0298	0.0489	0.0471	0.0402	0.0413	0.0546	0.0580	0.0437
LGR [7]	0.0423	0.0152	0.0502	0.0735	0.0195	0.0512	0.0452	0.0393	0.0419
PCL [9]	0.0286	0.0284	0.0501	0.0505	0.0644	0.0628	0.0418	0.0395	0.0458
AGR [8]	0.0257	0.0263	0.0429	0.0431	0.0347	0.0343	0.0458	0.0463	0.0374
Ours	**0.0284**	**0.0256**	**0.0461**	**0.0431**	**0.0514**	**0.049**	**0.0235**	**0.0172**	**0.0355**

3.2 Ablation Study

Different model configurations: Multiple experiments were conducted to check which yields the best possible output. We first find the results for global-local network with CBAM block (ACNet-CBAM) in place of concurrent axial attention block. We then train only the local network (only U-Net) and compile the results. We then check if the resize dimension of the U-Net output (Global_{dim} = 128) before feeding it to the global context network makes any difference in the heatmap generation. To find the best possible configuration for the GCA block in the global network, we train the global-local network by replaced GCA with GCNet [14] (ACNet-GCNet) and non-local block [14] (ACNet-NL). Our method outperforms these variants as shown in Table 3, thereby reinstating the benefit of our design choices.

4 Conclusion

In this paper, we propose a novel Appearance-Context Network (ACNet), that consists of a novel Global Context Aware (GCA) axial attention module which

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Table 3: Ablation study experiments of ACNet using DeepFashion dataset.

Method	L.Cl	R.Cl	L.Sl	R.Wl	L.Wl	L.Hm	L.Hm	R.Hm	Avg
ACNet-CBAM	0.039	0.039	0.069	0.072	0.073	0.074	0.078	0.08	0.065
Only U-Net	0.026	0.027	0.048	0.047	0.048	0.045	0.026	0.028	0.037
$\text{Global}_{dim} = 128$	0.029	0.029	0.056	0.054	0.051	0.050	0.029	0.033	0.041
ACNet-GCNet	0.022	0.023	0.044	0.039	0.039	0.038	0.021	0.021	0.030
ACNet-NL	0.024	0.024	0.042	0.042	0.043	0.043	0.024	0.024	0.033
Ours	**0.031**	**0.021**	**0.042**	**0.020**	**0.031**	**0.035**	**0.012**	**0.016**	**0.026**

seamlessly extracts global context features for fashion landmark detection. The ACNet successfully disentangles fashion apparel features across global and local levels and focus on discriminatory cues to localize the landmark locations. Extensive experimental results on DeepFashion and Fashion Landmark Detection datasets show that our method outperforms the state-of-the-art methods.

References

[1] Z. Liu, P. Luo, S. Qiu, X. Wang and X. Tang, DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *CVPR*, pp. 1096-1104, Las Vegas, NV, 2016.

[2] Z. Liu, S. Yan, P. Luo, X. Wang and X. Tang, Fashion landmark detection in the wild. In *ECCV*, pp. 229-245, Amsterdam (Netherlands), 2016.

[3] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang and X. Tang, Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *ACM MM*, pp. 172-180, Mountain View, CA, 2017.

[4] W. Wang, Y. Xu, J. Shen and S.C. Zhu, Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification. In *CVPR*, 4271-4280, Salt Lake City, UT, 2018.

[5] S. Lee, S. Oh, C. Jung and C. Kim, A Global-Local Embedding Module for Fashion Landmark Detection. In *ICCVW*, pp. 3153-3156, Seoul (S.Korea), 2019.

[6] Y. Li, S. Tang, Y. Ye and J. Ma, Spatial-Aware Non-Local Attention for Fashion Landmark Detection. In *ICME*, pp. 820-825, Shanghai (China), 2019.

[7] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao and L. Lin, Layout-Graph Reasoning for Fashion Landmark Detection. In *CVPR*, pp. 2932-2940, Long Beach, CA, 2019.

[8] M. Chen, H. Ying, Y. Qin, L. Qi, Z. Gan and Y. Sun, Adaptive Graph Reasoning Network for Fashion Landmark Detection. In *ECAI*, pp. 2672-2679, Santiago de Compostela (Spain), 2020.

[9] H. Liu, M. Song, W. Shi and X. Li, Position Constraint Loss For Fashion Landmark Estimation. In *ICASSP*, pp. 1868-1872, Barcelona (Spain), 2020.

[10] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234-241, Munich (Germany), 2015.

[11] P. K. Reddy, A. Kanakatte, J. Gubbi, M. Poduval, A. Ghose and B. Purushothaman, Anatomical Landmark Detection using Deep Appearance-Context Network. In *EMBC*, pp. 3569-3572, Mexico, 2021.

[12] G. Bhattacharya, N. Kilari, J. Gubbi, B. Vasudevan, A. Pal and B. Purushothaman, DAtRNet: Disentangling Fashion Attribute Embedding for Substitute Item Retrieval. In *CVPRW*, pp. 2283-2287, New Orleans, LA, 2022.

[13] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, CBAM: Convolutional block attention module. In *ECCV*, pp. 3-19, Munich (Germany), 2018.

[14] Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In *ICCVW*, pp. 1971-1980, Seoul (S.Korea), 2019.