Improving Fashion Landmark Detection by Dual Attention Feature Enhancement

Ming Chen, Yingjie Qin, Lizhe Qi, Yunquan Sun
Fudan University
{mingchen18, yjqin18, qilizhe, sunyunquan}@fudan.edu.cn

Abstract

Fashion landmark detection is a fundamental problem in visual fashion analyze, which aims at locating the precise coordinates of functional key points defined on clothes. Dozens of deep learning-based methods are proposed to address this problem. How to extract adequate and effective features is a critical point for this challenging task. In this paper, we propose the Dual Attention Feature Enhancement(DAFE) module, which strengthens the extracted features by adaptively reusing low-level image details and emphasizing informative parts. First, DAFE enhances the pixel-wise information through capturing the spatial details from low-level features by the guidance of attention matrix, which is generated from high-level ones. Second, DAFE emphasizes task-related features by modeling long-range relationships between channels. Experimental experiments on Deepfashion and FLD datasets demonstrate that our method achieves state-of-the-art performance, and our approach also achieves competitive results on Deepfashion2 Landmark Estimation Challenge[1].

1. Introduction

The potential value of fashion analysis has attracted a lot of attention in the community. Fashion landmark detection is one of the fundamental yet challenging problems with wide applications in visual fashion analysis, like clothes category classification[12, 6], recommendation[8, 7] and retrieval[10].

With the release of large-scale fashion datasets[12, 13, 5], deep learning-based models have achieved impressive detection performance[12, 13, 18, 16]. Most current approaches pass the input image through a basic feature extraction network and then enhance the features map for this specific task, finally predict coordinates by estimating heatmap for each landmark. However, deep stacked convolution and pooling operations cause spatial detail loss and channel feature redundancy.

Nowadays, a new aspect of the architecture design, named attention has been studied extensively[17, 9, 2]. The attention mechanisms help to strengthen the feature representations by focusing on essential features and suppressing unnecessary ones. These mechanisms have improved performance in many fields. Our work integrates these powerful mechanisms into feature upsampling layers to enhance the feature maps with lightweight computations.

In this paper, we propose the novel Dual Attention Feature Enhancement(DAFE) module, an effective module that compensates for the loss of spatial detail and selects task-related features while recovering the size of feature maps. DAFE mainly contains two parts, Spatial Attentive Upsampling(SAU) block, and Channel-wise Attentive Selection(CAS) block. SAU selects the interesting spatial details in low-level feature maps by the guidance of spatial attention matrix, which is generated by high-level features through modeling long-range dependencies in the spatial axis. Then SAU associates the high-level features with selected low-level ones by skip connections. Moreover, channels of a feature map have different functions for a specific task, both task-related and irrelevant types of features are treated equally in the current works. To address this issue, CAS emphasizes the task-related features and suppress others by explicitly modeling interdependencies between channels. Benefiting from DAFE module, our network generates more adequate and effective features for landmark detection. Quantitive evaluations on Deepfashion and FLD datasets show the superiority of our model. Furthermore, the model is extended to Deepfashion2 Landmark Estimation Challenge and again achieves good performance.

In summary, there are three main contributions: First, we propose a simple yet effective network for fashion landmark detection based on Feature Pyramid Network. Second, we design the Dual Attention Feature Enhancement(DAFE) module to enhance the feature representations while recovering the size of feature maps. Third, the proposed model performs well on three large-scale fashion datasets.

[1]The Deepfashion2 Challenge website is : https://codalab.lri.fr/competitions/564

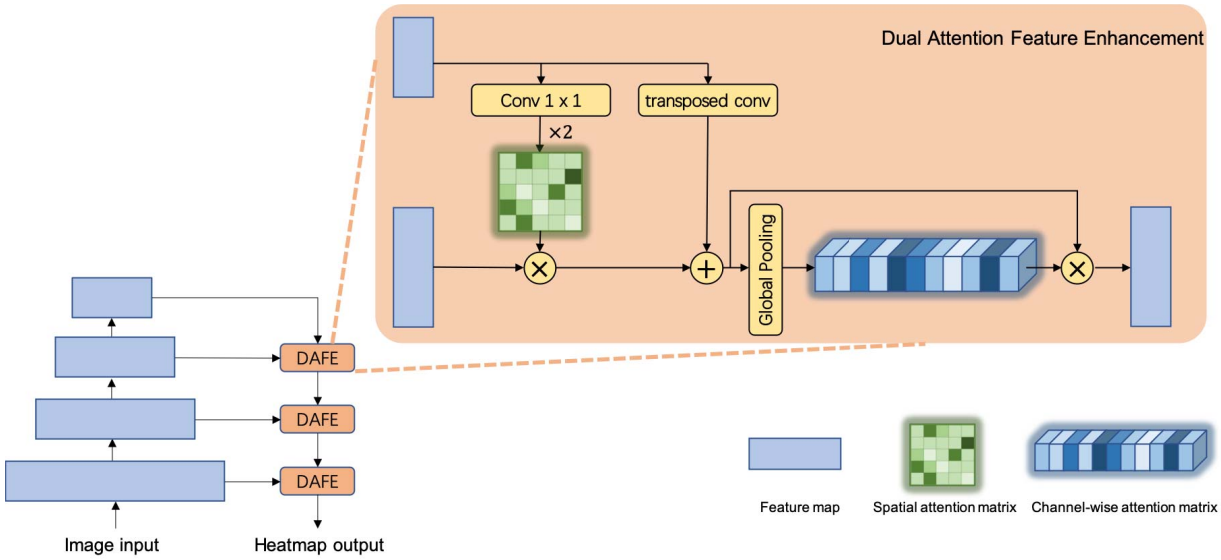

Figure 1. Illustration of our model that incorporates basic convolutional network for features extraction and stacked DAFE modules for feature enhancement. DAFE generates spatial attention matrix and channel-wise attention matrix to select spatial details and task-related feature channels respectively.

2. Related work

2.1. Fashion landmark detection

Extensive research efforts have been devoted exclusively to fashion landmark detection and achieve excellent performances. Liu et al.[12] first introduced the neural network to the fashion landmark detection. They formulated the detection as a regression task and designed FashionNet to regress landmark coordinates directly. Liu et al.[13] designed pseudo-labels to enhance in-variability of fashion landmark. Inspired by the attention mechanism, Wang et al.[16] propose an attentive grammar network with high-level human knowledge to predict the positions of landmarks globally. Simultaneously, [16] indicates that the regression of the fashion landmark is highly non-linear and very difficult to learn directly. Therefore, they learn to predict a confidence map of positional distribution for each landmark. We also adopt this method to detect fashion landmarks.

2.2. Attention mechanism

Since the attention mechanism has widely applied in natural language processing[14, 15], it also achieved good performance in the computer vision tasks of object detection[1], image recognition[3] and pose estimation[4]. In these applications, attention mechanisms act the role of enabling the neural network to focus more on useful information and ignore the useless parts. In this way, the network tilts the limited computational resources towards concerned information. Especially in the field of computer vision, Hu et al.[9] through the squeeze and excitation mechanism to learn global information among the feature channels and perform feature recalibration. Respectively, Wang et al.[17] proposed a generic Non-Local(NL) block that can capture long-range dependencies directly between two distance-independent image or video positions. Cao et al. [2] simplified the NL block and proposed a global context block combining the simplified NL block with SE block[9], which is more lightweight and effective.

3. Our approach

In this section, we first present our detection framework for fashion landmarks, then introduce our Dual Attention Feature Enhancement(DAFE) module in detail.

3.1. Landmark detection framework

The goal of fashion landmark detection is predicting the locations of n functional key points from an RGB image($H \times W \times 3$). We adopt the most widely-used framework to tackle this problem, which estimates n key points confidence maps(heatmaps) for n landmarks labeled in the datasets and then chooses the locations with the highest values as the predicted key points.

As shown in Fig.1, we build the fashion landmark detection network following the intuition of the Feature Pyramid Network(FPN)[11]. First, we use the ResNet-50 to capture multi-scale feature maps of the input image. Then, we recover the size of feature maps by transposed convolutions. To compensate for the loss of spatial detail and select task-related features, we design the Dual Attention Feature Enhancement(DAFE) module, which mainly contains two parts, Spatial Attentive Upsampling(SAU) block, and

Table 1. The normalized error on Deepfashion-C dataset

Methods	*L.Collar*	*R.Collar*	*L.Sleeve*	*R.Sleeve*	*L.Waist*	*R.Waist*	*L.Hem*	*R.Hem*	*Avg.*
FashionNet[12]	.0854	.0902	.0973	.0935	.0854	.0845	.0812	.0823	.0872
DFA[13]	.0628	.0637	.0658	.0621	.0726	.0702	.0658	.0663	.0660
DLAN[18]	.0570	.0611	.0672	.0647	.0703	.0694	.0624	.0672	.0643
AFGN[16]	.0415	.0404	.0496	.0449	.0502	.0523	.0537	.0551	.0484
Ours	**.0295**	**.0297**	**.0363**	**.0361**	**.0311**	**.0313**	**.0394**	**.0402**	**.0342**

The best results are marked in **bold**.

Table 2. The normalized error on FLD dataset

Methods	*L.Collar*	*R.Collar*	*L.Sleeve*	*R.Sleeve*	*L.Waist*	*R.Waist*	*L.Hem*	*R.Hem*	*Avg.*
FashionNet[12]	.0784	.0803	.0975	.0923	.0874	.0821	.0802	.0893	.0859
DFA[13]	.048	.048	.091	.089	-	-	.071	.072	.068
DLAN[18]	.0531	.0547	.0705	.0735	.0752	.0748	.0693	.0675	.0672
AFGN[16]	.0463	.0471	.0627	.0614	.0635	.0692	.0635	.0527	.0583
Ours	**.0366**	**.0369**	**.0587**	**.0573**	**.0485**	**.0478**	**.0504**	**.0497**	**.0482**

The best results are marked in **bold**.

Channel-wise Attentive Selection(CAS) block. We stack three DAFE modules to generate final feature maps. Finally, the enhanced features are utilized to estimate heatmaps.

3.2. Dual Attention Enhancement(DAE) module

3.2.1 Spatial attentive upsampling block

We design the Spatial Attentive Upsampling(SAU) block to recover the image size and spatial details based on the NL block[17] and skip connections. SAU integrates the low-level details into final feature maps, which from shallow layers in the feature extraction network. Moreover, it utilizes the high-level feature maps to generate spatial attention matrices to select informative spatial details. The spatial attention computes the response at a position as the importance of each region in the feature maps. Specifically, We define a feature map extracted by the network as $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, and $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$ is the corresponding low-level feature map in the backbone network. The spatial attention matrix $\mathbf{M}$ is generated by a 1×1 convolutional operation $\mathbf{C}$ followed by a sigmoid function, as adopted in [2]:

$$\mathbf{M} = Sigmoid(\mathbf{CX}) \tag{1}$$

Assuming $\widetilde{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$ denotes the output of the SAU block. The enhanced feature map can be expressed as:

$$\widetilde{\mathbf{X}} = U_t(\mathbf{X}) \oplus U_b(\mathbf{M}) * \mathbf{Y} \tag{2}$$

where $U_t(\mathbf{X})$ denotes transposed convolution operation, $U_b(\mathbf{X})$ denotes bilinear upsampling operation and $\oplus$ denotes broadcast element-wise addition.

3.2.2 Channel-wise attentive selection block

The Channel-wise Attentive Selection(CAS) block is designed to emphasize informative features and suppress useless ones in the dense feature maps. In this way, the network pays more attention to useful information and improves the utilization of computational resources. Specifically, for a feature map after SAU operation, we define it as $\mathbf{X} = \{\mathbf{x_i}\}_{i=1}^{C}$, where C is the channel number and $\mathbf{x_i} \in \mathbb{R}^{H\times W}$ is a feature slice. We use a global average pooling to aggregate the global feature in every feature slice together. The aggregated feature $\mathbf{z} \in \mathbb{R}^{C\times 1\times 1}$ is calculated by:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{x_c}(i, j) \tag{3}$$

where z_c and $\mathbf{x_c}$ are the c-th element of $\mathbf{Z}$ and $\mathbf{X}$.

To compute the importance for each channel, we adopt one 1×1 convolutions $\mathbf{C}_1$, one $ReLU$, one 1×1 convolutions $\mathbf{C}_2$ sequentially, the channel-wise attention map $\mathbf{U} \in \mathbb{R}^{C\times 1\times 1}$ can be expressed as:

$$\mathbf{U} = \mathbf{C}_2 ReLU(\mathbf{C}_1\mathbf{Z}) \tag{4}$$

Given the channel-wise attention map, the enhanced feature map $\widetilde{\mathbf{X}} \in \mathbb{R}^{C\times H\times W}$ is calculated by:

$$\widetilde{\mathbf{X}} = \mathbf{X} \otimes \mathbf{U} \tag{5}$$

where $\otimes$ denotes matrix multiplication.

4. Results

We compare our fashion landmark detection model with four current deep learning-based models on Deepfashion and FLD datasets. Deepfashion[12] is a large-scale clothes dataset, which offers 289222 fashion images with category, attribute, bounding box and landmark annotations. FLD is a subset of Deepfashion, which has large pose and scale variations. For fashion landmark detection, each image is labeled with up to 8 fashion landmarks.

For the training, We first crop input image using labeled bounding boxes and then the cropped image is resized to

320×320. We use the batch size of 64 images on 4 GTX 2080Ti GPUs. And we use Adam to optimize the loss function, the initial learning rate is set to $1e-3$ and decreased by a factor of 0.1 every 10 training epochs. For the testing, We resize the cropped image in the same way as training. Our model generates a $n \times 80 \times 80$ landmark heatmap for a single input image. The locations with the highest values are regarded as the predicted positions.

In Table 1 and 2, we provide the quantitative evaluation results of our proposed method. Our model achieves the state-of-the-art at 0.0342 NE on Deepfashion and 0.0482 NE on FLD. Moreover, our method consistently outperforms other models on all of the fashion landmarks. We also evaluate our model on Deepfashion2 dataset. It is the largest fashion database to date, which contains 491K images of 801K items in total. Given half of the training dataset, we achieve the overall detection AP 54.9 on validation set without any bells and whistles.

5. Conclusion

In this paper, we tackle the fashion landmark detection with the deep convolutional neural networks. The Dual Attention Feature Enhancement module is proposed to capture more spatial details and select task-related features, which contributes to making more accurate and precise landmark detection. Overall, our model achieves state-of-the-art performances on Deepfashion and FLD fashion datasets and also performs well on the Deepfashion2 dataset.

References

[1] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond.

[3] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[4] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.

[5] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping 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*, pages 5337–5345, 2019.

[6] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1463–1471, 2017.

[7] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086. ACM, 2017.

[8] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2018.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015.

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[12] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[13] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision*, pages 229–245. Springer, 2016.

[14] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[16] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun 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*, pages 4271–4280, 2018.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[18] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 172–180. ACM, 2017.