24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Adaptive Graph Reasoning Network for Fashion Landmark Detection

Ming Chen [12], Hang Ying [3], Yingjie Qin [12], Lizhe Qi [12*], Zhongxue Gan [12*], Yunquan Sun [12]

Abstract. In this paper, we address the fashion landmark detection task by enforcing structural fashion layout relationships among landmarks based on Graph Convolutional Networks (GCNs). Unlike previous works that detect each fashion landmark separately and ignore the rich semantic layout relation among different landmarks, we propose an Adaptive Graph Reasoning Network (AGRNet) to integrate the convolutional features with the human commonsense knowledge and make detected fashion landmarks be coherent with clothes layouts from a global perspective. Specifically, we design the Adaptive Graph Reasoning (AGR) module and stack it on top of Fully Convolutional Networks (FCNs), which enforces fashion layout constraints and semantic relations of fashion landmarks on deep representations. AGR maps the convolutional features into structural graph node representations and performs adaptive reasoning according to the correlation matrix, which is adaptively generated from defined basic fashion layout and confidence maps of all landmarks. The graph-based reasoning evolves the cloth node representations to achieve global layout coherency and then the evolved graph nodes are mapped back to enhance convolutional feature representations. Furthermore, we design the Dual Attention Upsample (DAU) module on each decoder layer to emphasize the spatial detailed and task-related features by modelling the semantic interdependencies in spatial and channel dimensions respectively. We achieve new state-of-the-art detection performance on two challenging fashion landmark datasets, i.e., Deepfashion and FLD dataset. In particular, a Normalized Error (NE) score of 0.0297 on the Deepfashion test set is achieved without any additional annotations.

1 INTRODUCTION

Detecting fashion landmarks from an RGB image is a fundamental and practical task, whose goal is to predict the positions of functional keypoints defined on the fashion items, such as the corners of the neckline, hemline, and cuff. The study of this task can be applied to comprehensive high-level fashion applications, such as clothes category classification [6, 24, 13, 1], recommendation [19, 16, 14] and retrieval [18, 21, 12]. With the release of large-scale fashion datasets[24, 25, 11], convolutional neural networks based models [24, 25, 30, 28, 22, 31] have achieved impressive detection performance. To accomplish the task of fashion landmark detection effectively, we need to deal with arbitrary clothing appearances, diverse styles, and part occlusion. For example, clothes on people have arbitrary deformation, sleeves have different lengths and styles, and the collars of clothes often obscured by long hair. The local convolutional features may lead to the detection results of ambiguous landmarks and unreasonable landmark layouts. Therefore, it is necessary to endow the deep network with the capability of structure graph reasoning for structure-consistent landmark detection.

(a) full body. (b) upper body. (c) lower body.

Figure 1. The goal of fashion landmark detection is to recognize and locate the functional key points defined on clothes, such as the corners of neckline, hemline, and cuff. The arbitrary appearances, diverse styles and occlusion of clothes make it challenging to detect each landmark accurately.

Recent methods [24, 25, 30, 22] based on Fully Convolutional Networks (FCNs) [26] extract the image features by a deep convolutional network and then enlarge the resolution of the feature maps by bilinear upsample or transposed convolution, finally estimate confidence maps for each landmark separately regarding the fashion landmark detection as an end-to-end regression problem and ignoring the layout constraints between different landmarks. Although the deep-stacked convolutions help to capture more semantic relations, it can not leverage the relationship between landmarks in a global view,

[1] Shanghai Engineering Research Center of AI & Robotics, Fudan University, China, email: mingchen18@fudan.edu.cn. * Corresponding authors.
[2] Engineering Research Center of AI & Robotics (Fudan University), Ministry of Education, China
[3] Hangzhou Normal University, China

which is essential to fashion landmark detection.

Some state-of-the-art methods [28, 31] inject human commonsense knowledge into the detection model. For example, Wang et al. [28] propose a fashion grammar model for visual fashion analysis and use a bidirectional recurrent neural network for message passing over fashion grammar. Yu et al. [31] define a complicated fashion layout-graph and propagate the information between correlated landmarks to update the feature representations. Intuitively, the information from incorrect landmarks is not as important as information from well-detected ones for landmark detection. The well-detected landmarks could help the detection of other landmarks, the incorrect landmarks instead may deteriorate. However, these methods share information of different landmarks equally by the fixed edges in the prior knowledge fashion grammar or layout-graph regardless of whether or not the landmarks are detected correctly.

To address the above problems, we propose a novel framework, called Adaptive Graph Reasoning Network (AGRNet), for fashion landmark detection. It introduces graph-based reasoning to adaptive enforce structural layout constraints among landmarks on the deep representations. We define a basic fashion layout that encodes the human commonsense knowledge (e.g. symmetry relations, kinematics relations) and constructs graph node representations for all landmarks from convolutional features. To propagate the information and perform structural graph reasoning among landmarks, we need to generate a correlation matrix of the layout graph nodes. Instead of using a fixed correlation matrix, we introduce an adaptive way to generate the correlation matrix for each image which may face different deformation and occlusion. The positions which have the highest responses in confidence maps are considered as the detected landmarks since the values of each position represent the occurrence possibility of landmarks. Inspired by this, we generate a weight vector from the basic confidence maps and then operate a recalibration on the basic correlation matrix by the guidance of the weight vector to generate the adaptive correlation matrix. In this way, the information from landmarks that have high confidences will be shared extensively for aiding the detection of other landmarks and the information from landmarks that have low confidences will be suppressed. Given the graph nodes representations of fashion landmarks and correlation matrix, we update the node representations by propagating the information between connected nodes based on Graph Convolutional Networks (GCNs). GCNs perform graph layout reasoning to make detected fashion landmarks be coherent with clothes layouts from a global perspective. Then the evolved graph node representations are mapped back to enhance the convolutional features.

Moreover, the Dual Attention Upsample (DAU) module on each decoder layer is proposed to further enhance feature representations, which emphasizes the spatial detailed and task-related features by modelling the semantic interdependencies in spatial and channel dimensions respectively. Spatial Attention (SA) block captures the interesting spatial details in low-level feature maps by the guidance of high-level ones. Meanwhile, Channel Attention (CA) block emphasizes the task-related features and suppresses useless ones by capturing channel dependencies between any two channel maps.

We empirically show the performance of the proposed method on two fashion landmark detection datasets. Additionally, ablation studies indicate the effectiveness of our proposed modules. In summary, our main contributions are threefold as follows: i) We propose an Adaptive Graph Reasoning Network (AGRNet) to enforces fashion layout constraints and semantic relations of fashion landmarks for fashion landmark detection. ii) We design a Dual Attention Upsample (DAU) module on each decoder layer to further enhance the feature representations by emphasizing the spatial detailed and task-related features. iii) Combining graph reasoning and attention upsample, we achieve new state-of-the-art results on Deepfashion and FLD benchmarks.

2 RELATED WORK

2.1 Fashion Landmark Detection

Extensive research efforts have been devoted to fashion landmark detection and achieved excellent performances. Liu et al. [24] first introduce the neural network to the task of fashion landmark detection. They formulate the detection as a regression task and design FashionNet to regress landmark coordinates directly. Liu et al. [25] design pseudo-labels to enhance in-variability of fashion landmark. Yan et al. [30] combine selective dilated convolution and recurrent spatial transformer for localizing cloth landmarks in unconstrained scenes. The methods mentioned above almost estimate landmarks for each landmark separately and thus may detect ambiguous and structure-inconsistent landmarks. Wang et al. [28] propose an attentive grammar network with high-level human knowledge to predict the positions of landmarks globally. Simultaneously, Wang et al. [28] indicate 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. The more current method [31] define a complicated fashion layout-graph and propose to model the structural layout relationships among landmarks. However, they propagate the information according to a fixed layout-graph and cannot deal with the diverse deformation or occlusion. Furthermore, all the works suffer from loss of image detailed information and can only locate the fashion landmarks roughly.

2.2 Graph-based Reasoning

Graph-based methods have been very popular in recent years and shown to be an efficient way of relation reasoning. CRFs [5] and random walk networks [2] are proposed based on the graph model for effective image segmentation. Recently, Graph Convolution Networks (GCNs) [20] are proposed for semi-supervised classification, and Chen et.al [8] propose to use GCNs to capture relations between objects in the large-scale object detection task, which poses severe challenges due to long-tail data distributions, heavy occlusions, and class ambiguities. Reddy et.al. [27] predict 2D and 3D locations of occluded key points for objects using graph reasoning in a largely self-supervised manner. We adopt the reasoning power of graph convolutions to build a global reasoning module for reasoning between correlated fashion landmarks.

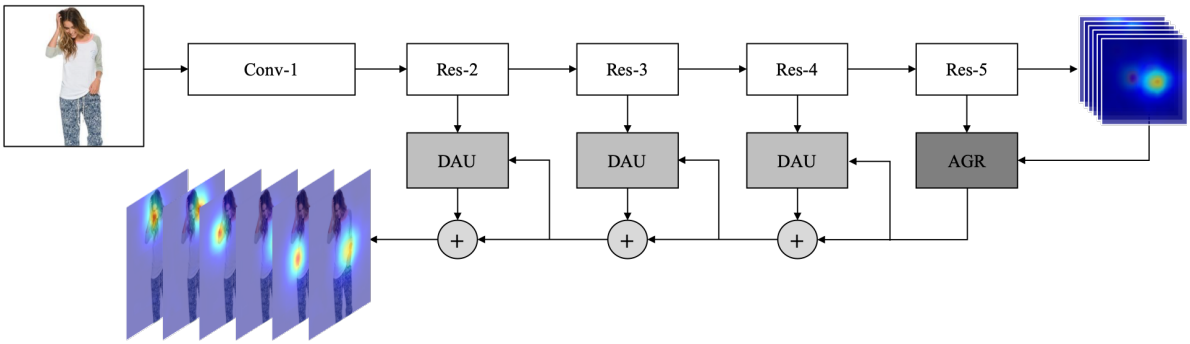


Figure 2. Illustration of our model that incorporates basic convolutional network for features extraction, Adaptive Graph Reasoning (AGR) module for adaptive structural layout reasoning and stacked Dual Attention Upsample (DAU) modules for attentive feature enhancement. AGR module enforces fashion layout constraints and semantic relations of fashion landmarks on deep representations. DAU module generates spatial attention matrix and channel attention matrix as the guidance to adaptively enhance spatial detailed and task-related channel features respectively.

2.3 Attention Mechanism

The attention mechanisms have achieved excellent performance in the many computer vision tasks [3, 7, 9, 10]. In these applications, attention mechanisms act the role of enabling the neural network to focus more on useful information and ignore the useless parts. Especially, Hu et al. [17] through the Squeeze and Excitation (SE) mechanism to learn global information among the feature channels and perform feature recalibration. Wang et al. [29] propose a generic Non-Local (NL) block that can capture long-range dependencies directly between two distance-independent image or video positions. Cao et al. [4] simplify the NL block and propose a global context block combining the simplified NL block with SE block [17], which is more lightweight and effective. In the paper, we design a novel attention block, which combines the advantages of these attention mechanisms and further performs feature refinement across the multi-scale feature maps.

3 METHODS

3.1 Landmark Detection Framework

The task of fashion landmark detection aims at predicting the locations of n functional key points from an RGB image ($H \times W \times 3$), such as the corners of the neckline, hemline, and cuff. We learn to estimate n key points confidence maps (heatmaps) for n landmarks labelled in the datasets and then choose the coordinates with the highest values as the locations of predicted key points. As shown in Figure 2, we build the fashion landmark detection network following the intuition of the Feature Pyramid Network (FPN) [23]. First, we use the ResNet [15] to capture multi-scale feature maps of the input image and generate the basic confidence maps of landmarks. We use the Adaptive Graph Reasoning (AGR) module on the top of ResNet to model the dependencies of different landmarks and enforce the detected fashion landmarks to be coherent with structural fashion layouts from a global perspective. Then, we upsample the feature maps and fuse the multi-scale feature maps. Furthermore, to strengthen the spatial detailed and task-related features globally, we design the Dual Attention Upsample (DAU) module which emphasizes

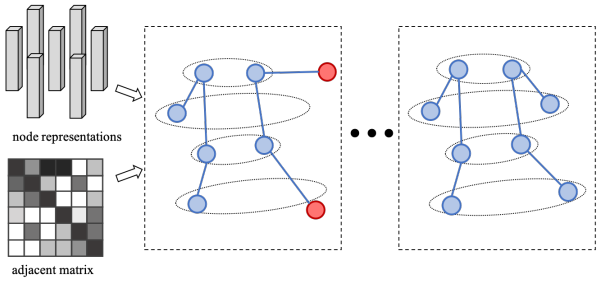


Figure 3. Illustration of Adaptive Graph Reasoning (AGR) module, where blue circles indicate reasonable landmarks and red circles indicate unreasonable landmarks.

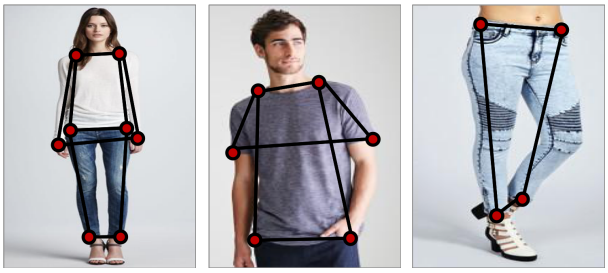
the informative features through modelling long-range dependencies in the spatial and channel dimensions. Finally, a 1×1 convolution with a sigmoid activation function is utilized to estimate final landmark confidence maps.

3.2 Adaptive Graph Reasoning Module

The AGR module aims to enhance convolutional features by adaptive graph reasoning among landmarks. As shown in Figure 3, through propagating information between correlated fashion landmarks guided by high-level human commonsense knowledge, AGR enforces fashion layout constraints and semantic relations of fashion landmarks on deep representations and make the detected fashion landmarks be coherent with structural fashion layouts.

3.2.1 Fashion Layout Definition

For mining semantic correlations and constraints among different fashion landmarks, we first define the general fashion layout that reflects prior knowledge of clothes (e.g. the bilateral symmetric property of clothes, the constraints among kinematically connected clothing parts). Specifically, we define the fashion layout constructed by graph nodes characterizing landmark categories and graph edges representing prior knowledge, which is denoted as $G = (V, E)$. We define the fashion landmark representations as $X \in \mathbb{R}^{n \times d}$, which is generated from convolutional feature maps. The basic landmark



(a) full body. (b) upper body. (c) lower body.

Figure 4. Illustration of our fashion layout, where red circles indicate ground-truth cloth landmarks, black lines indicates the correlations between landmark points.

adjacency weight matrix $A \in \{0,1\}^{n\times n}$ is initialized according to the edge connections in E as shown in Figure 4. n means the number of fashion landmarks and d means the feature dimension of each landmark.

3.2.2 Fashion Graph Node Construction

To reasoning fashion layout in the graph space, we need to map the convolutional feature maps to graph node representations. Given the input convolutional feature maps $F \in \mathbb{R}^{H\times W\times C}$, where H, W and C denote the height, weight, and channel, we first operate a dimension transformation on it, then we get the $F \in \mathbb{R}^{HW\times C}$. The graph node representations $X \in \mathbb{R}^{n\times d}$ for all n landmarks can be obtained by:

$$X = \sigma(\Phi(FW_m)^T FW_t). \tag{1}$$

where $W_m \in \mathbb{R}^{C\times n}$ and $W_t \in \mathbb{R}^{C\times d}$ are trainable sampling matrices. The Φ denotes normalized function *softmax* to sum all rows to one, and the σ denotes non-linear function *Relu*.

3.2.3 Adaptive Layout Reasoning

After creating the graph node representations $X \in \mathbb{R}^{n\times d}$ for all the n landmarks, it is natural to propagate the connected landmarks of X by the edges $A \in \{0,1\}^{n\times n}$ in the prior knowledge graph G. However, we observed that the information propagation may even deteriorate the detection performance when most of the landmarks of a cloth item are not detected correctly.

Intuitively, the information from inaccurate landmarks should be suppressed and the information from well-detected landmarks should be propagated to aid the poor detection of other landmarks. Thus, we attempt to weight the information from different landmarks. We introduce a simple way to generate the weights of fashion nodes to share information better. The positions which have the highest responses in each channel of the confidence map $H \in \mathbb{R}^{H\times W\times n}$ are considered as the detected landmarks since the values of each position represent the occurrence possibility of landmarks. The inaccurate landmarks always have low response scores. We use the response scores $S \in \mathbb{R}^n$ of n landmarks to weight the graph edges. The adaptive adjacency matrix $\widehat{A} \in \mathbb{R}^{n\times n}$ can be generated by:

$$\widehat{A} = \Phi((1 \oplus S) \otimes A), \tag{2}$$

where the Φ denotes normalized function *softmax* to sum all rows to one, $\oplus$ denotes broadcast element-wise addition, and $\otimes$ denotes broadcast element-wise multiplication. Note that the adaptive adjacency matrix $\widehat{A}$ is asymmetric.

Given the graph node representations of fashion landmarks $X \in \mathbb{R}^{n\times d}$ and adaptive adjacency matrix $\widehat{A} \in \mathbb{R}^{n\times n}$. We update the node representations by propagating information between connected nodes based on Graph Convolutional Network (GCN).

Unlike standard convolutions that operate on local Euclidean structures in an image, the goal of GCN is to learn a function $f(\cdot,\cdot)$ on a graph G, which takes feature representations $X^l \in \mathbb{R}^{N\times D}$ and the corresponding correlation matrix $A \in \mathbb{R}^{N\times N}$ as inputs (where D denotes the number of nodes and D indicates the dimensionality of node features), and updates the node features as $X^{l+1} \in \mathbb{R}^{N\times D'}$. Every GCN layer can be written as a non-linear function by

$$X^{l+1} = f(X^l, A). \tag{3}$$

After employing the convolutional operation of [20], $f(\cdot,\cdot)$ can be represented as

$$X^{l+1} = h(AX^lW^l), \tag{4}$$

where $W^l \in \mathbb{R}^{D\times D'}$ is a transformation matrix to be learned and $A \in \mathbb{R}^{N\times N}$ is the correlation matrix, and $h(\cdot)$ denotes a non-linear operation.

Thus, after employing the GCNs on the fashion graph nodes X and adaptive correlation matrix $\widehat{A}$, the information is shared and propagate globally across all n landmarks according to the designed layout constraint by stacking multiple GCN layers.

3.2.4 Feature Enhanced via Graph Reasoning

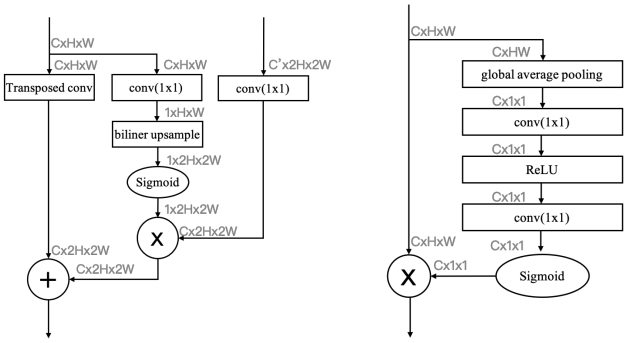
To enhance convolutional features via reasoning between landmarks, we map evolved graph node representations back into convolutional features. Given the input convolutional features F and evolved node representations X, we first perform the dimension transformation for $F \in \mathbb{R}^{HW\times C} \to F \in \mathbb{R}^{HW\times n\times C}$ and $X \in \mathbb{R}^{n\times d} \to X \in \mathbb{R}^{HW\times n\times C}$. Then we concatenated F and X to $X_s \in \mathbb{R}^{HW\times n\times(C+d)}$ for richer feature representations. We can get the enhanced convolutional feature representations $F_r \in \mathbb{R}^{HW\times C}$ by:

$$F_r = \sigma(\Phi(X_a W_{m'}))\sigma(XW_{t'}), \tag{5}$$

where $W_{m'} \in \mathbb{R}^{C+d}$ is a vector with $C+d$ dimension and $W_{t'} \in \mathbb{R}^{d\times C}$ is a trainable sampling matrix.

3.3 Dual Attention Upsample Module

The DAU module is designed to compensate for the loss of spatial details and emphasizes task-related features on the upsample layers. DAU mainly contains two crucial blocks, Spatial Attention (SA) block, and Channel Attention (CA) block, which strengthen the feature representations by modelling long-range interdependencies in spatial and channel dimensions separately.



(a) Spatial Attention. (b) Channel Attention.

Figure 5. Illustration of Spatial Attention (SA) and Channel Attention (CA) blocks in Dual Attention Upsample (DAU) module. Through modelling interdependencies in spatial and channel dimensions, SA enriches the spatial details and CA emphasize the task-related features.

3.3.1 Spatial Attention Block

The spatial details are critical to determining the final location since the rough area of the landmark is detected. We design the SA block to enrich the spatial details on the upsample layers. SA integrates the low-level details into final feature maps, which are from shallow layers in the feature extraction network. Moreover, it utilizes the high-level feature maps to generate spatial attention maps through computing the response at a position as the importance of each spatial detail, the attention maps further help to select informative spatial details and filter out useless parts. Specifically, we define the feature map extracted by the network as $X_h \in \mathbb{R}^{C\times H\times W}$, and $X_l \in \mathbb{R}^{C'\times 2H\times 2W}$ is the corresponding low-level feature map in the backbone network. The spatial attention map M_s is generated by a 1×1 convolutional operation C_1 followed by a sigmoid function:

$$M_s = Sigmoid((U_b(C_1 X_h)), \tag{6}$$

where $U_b(\cdot)$ denotes bilinear upsample operation. Then we fuse the high-level and selected low-level features to generate the enhanced feature map $\widetilde{X} \in \mathbb{R}^{C\times H'\times W'}$. The output of the SA block $\widetilde{X}$ can be expressed as:

$$\widetilde{X} = U_t(X_h) \oplus (M_s \otimes C_2 X_l), \tag{7}$$

where C_2 is a 1×1 convolutional operation, $U_t(\cdot)$ denotes transposed convolution operation, $\oplus$ denotes broadcast element-wise addition and $\otimes$ denotes element-wise broadcast multiplication.

3.3.2 Channel Attention Block

The CA block is designed to emphasize task-related 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 the feature map after SA operation, we define it as $X = \{x_i\}_{i=1}^{C}$, where C is the channel number and $x_i \in \mathbb{R}^{H\times W}$ is a feature slice. We use a global average pooling to aggregate the global feature in every feature slice together. The aggregated feature $Z \in \mathbb{R}^{C\times 1\times 1}$ is calculated by:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j), \tag{8}$$

where z_c and x_c are the c-th element of Z and X.

To compute the importance coefficient for each channel, we adopt one 1×1 convolutions C_1, one $ReLU$, one 1×1 convolutions C_2 sequentially, the channel-wise attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ can be expressed as:

$$M_c = C_2 ReLU(C_1 Z). \tag{9}$$

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

$$\widetilde{X} = (1 \oplus M_c) \otimes X, \tag{10}$$

where $\oplus$ denotes broadcast element-wise addition and $\otimes$ denotes element-wise broadcast multiplication.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metric

4.1.1 Deepfashion

Deepfashion [24] is a large-scale cloth dataset with comprehensive annotations. It offers 289222 fashion images, which are richly annotated with category, attribute, bounding box, landmarks and correspondence of image taken under different scenarios including store, street snapshot, and consumer. For fashion landmark detection, each image is labelled with up to 8 fashion landmarks including left/right collar end, left/right sleeve end, left/right hem and left/right waistline.

4.1.2 FLD

FLD [25] is a fashion landmark dataset with more diverse variations (e.g. pose, scale, background). It contains 123026 images and is divided into five subsets according to the positions and visibility of their ground truth landmarks. For each image, the annotations for 8 landmarks are offered.

4.1.3 Evaluation Metric

We employ the Normalized Error (NE) to estimate our model, which is a comprehensive metric in the task of fashion landmark detection. The NE is defined as the l_2 distance between predicted landmarks and ground truth landmarks in the normalized space. The calculation formula is

$$NE = \frac{\sum_k \left\{ \frac{d_k}{s_k} \delta(v_k = 1) \right\}}{\sum_k \{\delta(v_k = 1)\}} \times 100\% \tag{11}$$

where d_k is the distance between predicted landmark and ground truth landmark, s_k is the size of the image, v_k is the visibility of the landmark, which values 1 when the corresponding landmark is visible. The smaller values of NE indicate better results.

Table 1. Quantitative results for fashion landmark detection on Deepfashion dataset and FLD dataset with NE metric. Lower values are better. The best results are marked in **bold**.

Deepfashion									
Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waist	R.Waist	L.Hem	R.Hem	Avg.
FashionNet [24]	.0854	.0902	.0973	.0935	.0854	.0845	.0812	.0823	.0872
DFA [25]	.0628	.0637	.0658	.0621	.0726	.0702	.0658	.0663	.0660
DLAN [30]	.0570	.0611	.0672	.0647	.0703	.0694	.0624	.0672	.0643
FPN [23]	.0351	.0360	.0563	.0584	.0446	.0468	.0661	.0697	.0518
AFGN [28]	.0415	.0404	.0496	.0449	.0502	.0523	.0537	.0551	.0484
SANL [22]	.0277	.0282	.0391	.0394	.0297	.0299	.0395	.0401	.0342
LGR [31]	.0270	**.0116**	**.0286**	.0347	.0307	.0435	**.0160**	**.0162**	.0336
Ours	**.0256**	.0251	.0318	**.0324**	**.0271**	**.0286**	.0328	.0341	**.0297**

FLD									
Methods	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waist	R.Waist	L.Hem	R.Hem	Avg.
FashionNet [24]	.0784	.0803	.0975	.0923	.0874	.0821	.0802	.0893	.0859
DFA [25]	.048	.048	.091	.089	-	-	.071	.072	.068
DLAN [30]	.0531	.0547	.0705	.0735	.0752	.0748	.0693	.0675	.0672
FPN [23]	.0437	.0441	.0673	.0682	.0710	.0721	.0635	.0618	.0614
AFGN [28]	.0463	.0471	.0627	.0614	.0635	.0692	.0635	.0527	.0583
SANL [22]	.0296	.0298	.0489	.0471	.0402	.0413	.0546	.0580	.0437
LGR [31]	.0423	**.0152**	.0502	.0735	**.0195**	.0512	**.0452**	**.0393**	.0419
Ours	**.0257**	.0263	**.0429**	**.0431**	.0347	**.0343**	.0458	.0463	**.0374**

4.2 Implementation Details

Our fashion landmark model is built upon a ResNet-50 [15] backbone with an AGR module and 3 DAU modules. We crop input images using labelled bounding boxes and resize all the cropped images into 224×224. Thus, our network generates eight 56×56 confidence maps for fashion landmarks. We train the model using stochastic gradient descent with a batch size of 64 images, which is optimized by Adam optimizer with an initial learning rate of $1e-3$ on 4 GTX 2080Ti GPUs. On Deepfashion, we linearly decrease the learning rate by a factor of 10 every 10 epochs. On FLD, we linearly decrease the learning rate by a factor of 10 every 20 epochs. We set the mean squared error (MSE) equation as an objective function between final predicted confidence maps and ground-truth. For the testing, we resize the cropped images in the same way as training. Our model generates eight 56×56 landmark confidence maps for a single input image. The locations with the highest values are regarded as the predicted positions.

4.3 Compare with State-of-the-art

AGRNet achieves a significant improvement over two standard fashion landmark detection benchmarks compared with existing state-of-the-art methods (e.g. FashionNet [24], DFA [25], AFGN [28], LGR [31]) . In Table 1, we provide the quantitative evaluation results of our proposed method and other methods. Our model outperforms the state-of-the-art at 0.0297 NE on Deepfashion and 0.0374 NE on FLD. Compared with traditional FCN models [24, 25, 30, 23, 22], we enforce structural layout constraints among landmarks on the deep representations and generate more structure-constant landmark detection. Compared with AFGN [28] and LGR [31], which define fixed correlation matrices for all samples, we perform adaptive graph reasoning among landmarks and would be inclined to share the information from well-detected landmarks. Thus, unlike LGR only decrease the NE in part of the landmarks on Deepfashion but performs poorly in some hard landmarks (e.g. 0.0347 NE in R.Sleeve, 0.0435 NE in R.Waist), our model consistently decreases the NE and performs well in all landmarks on Deepfashion and FLD. Sampled landmark detection results are presented in Figure 7.

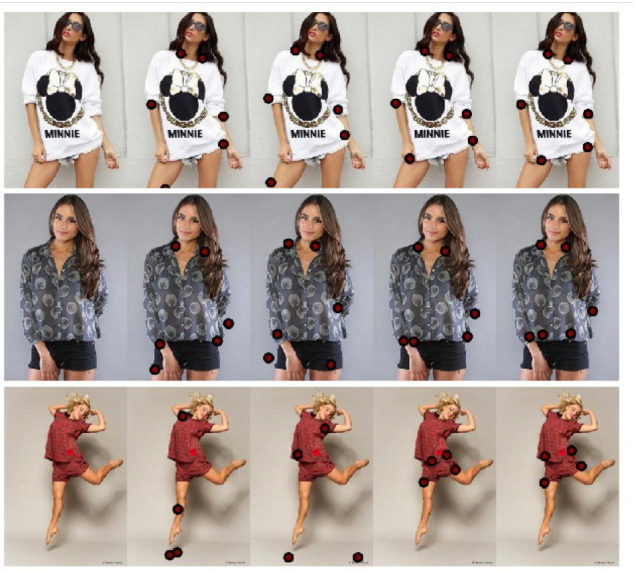


Figure 6. Qualitative results for VGG16, FPN [32] and Ours over DeepFashion dataset.

4.4 Ablation Study

In this section, we perform an in-depth study of the proposed modules in our detection network on Deepfashion dataset.

4.4.1 Effectiveness of Adaptive Graph Reasoning Module

In the first list of Table 2, we explore the effectiveness of different numbers of GCN layers in Adaptive Graph Reasoning (AGR) module. ResNet-50 with three transposed convolution layers baseline achieves 0.0406 NE score, which is the worst

Figure 7. Visualization results of our proposed fashion landmark detection approach. Images on the first row are the results on DeepFashion-C test set, and results on FLD dataset are on the second row.

result in our experiments. Benefiting from graph reasoning, AGR module with two GCN layers achieve the best performance at 0.0335 NE score, with a 17% improvement over the baseline. The performance tends to be destroyed with the depth increasing of GCN layers. Thus, we select two layers in our stand model and apply it to all extensive experiments.

In the second list of Table 2, we build an ablation study of the different correlation matrices in the AGR module. When the AGR module with fixed correlation matrix is used, we get a slightly better detection performance at 0.0374 NE owing to information propagation among landmarks. When we replace the fixed correlation matrix with the proposed adaptive correlation matrix, the NE decreases by 14% achieving better performance at 0.0335 NE. The adaptive correlation matrix guides the GCNs to share information better.

Table 2. Ablation study on Deepfashion dataset. We present the results generated by different numbers of GCN layers in AGR module and compare the effectiveness of the different correlation matrices in the AGR module. The best results are marked in **bold**.

Different Numbers of GCN Layers		
Method	NE score	Δ NE score
baseline	.0406	-
one layer	.0383	.0023
two layers	**.0335**	.0071
three layers	.0397	.0009
four layers	.0452	.0046

Different Correlation Matrices		
Method	NE score	Δ NE score
baseline	.0406	-
fixed matrix	.0374	.0032
adaptive matrix	**.0335**	.0071

4.4.2 Effectiveness of Dual Attention Upsample Module

In the first list of Table 3, we build an ablation study of different stacked DAU modules. We use ResNet-50 with an AGR module and three transposed convolution layers as our baseline model, which achieves 0.0335 NE score. To demonstrate the superior ability of the DAU module, we replace a traditional convolution layer with a DAU module one by one. The performances get better with the number of DAU modules increasing, DAU modules consistently strengthen the feature representations on different scale decoder layers. When three transposed convolution layers are all replaced with the DAU modules, we get the best performance at 0.0297 NE score. Thus, we select three stacks of DAU modules in our final model and apply it to the following experiments.

In the second list of Table 3, we explore the effectiveness of each block in the DAU module. With only Spatial Attention (SA) blocks are used, we can achieve 0.0316 NE. SA blocks enrich the spatial details in the feature maps, which bring large performance improvement. With only Channel Attention (CA) blocks are used, we can achieve 0.0311 NE owing to the emphasis on the informative features. With the complete DAU modules are used, we can achieve 0.0297 NE. Both feature enhancements in spatial and channel dimensions contribute to improving the landmark detection performance.

Table 3. Ablation study of DAU module on Deepfashion. We present the results generated by different stack numbers of DAU modules and compare the effectiveness of the different blocks in the DAU module. The best results are marked in **bold**.

Different Stack Numbers		
Method	NE score	Δ NE score
baseline	.0335	-
one stack	.0318	.0017
two stacks	.0306	.0029
three stacks	**.0297**	.0038

Different Components			
Method	DAU		NE score
	SA	CA	
baseline			.0335
ours	✓		.0316
ours		✓	.0311
ours	✓	✓	**.0297**

5 CONCLUSION

In this paper, we have presented an Adaptive Graph Reasoning Network (AGRNet) for fashion landmark detection, which makes detected fashion landmarks be coherent with clothes layouts from a global perspective. Specifically, we introduce a graph adaptive reasoning module to propagate information between graph node representations of correlated landmarks by the guidance of human commonsense knowledge. Moreover, a dual attention upsample module is proposed to em-

phasize the spatial detailed and task-related features to further improve the detection performance. The ablation experiments show that adaptive reasoning module helps to detect structure-consistent landmarks and the attention upsample module enhance the feature representations. Combining the adaptive reasoning and attention upsample module, our network achieves outstanding performance consistently on two fashion landmark detection datasets.

REFERENCES

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham, 'Learning attribute representations with localization for flexible fashion search', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7708–7717, (2018).

[2] Gedas Bertasius, Lorenzo Torresani, Stella X Yu, and Jianbo Shi, 'Convolutional random walk networks for semantic image segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 858–866, (2017).

[3] 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*, pp. 2956–2964, (2015).

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu, 'Gcnet: Non-local networks meet squeeze-excitation networks and beyond', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (June 2019).

[5] Siddhartha Chandra, Nicolas Usunier, and Iasonas Kokkinos, 'Dense and low-rank gaussian crfs using deep embeddings', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5103–5112, (2017).

[6] Huizhong Chen, Andrew Gallagher, and Bernd Girod, 'Describing clothing by semantic attributes', in *European conference on computer vision*, pp. 609–623. Springer, (2012).

[7] 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*, pp. 3640–3649, (2016).

[8] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta, 'Iterative visual reasoning beyond convolutions', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7239–7248, (2018).

[9] 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*, pp. 1831–1840, (2017).

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, 'Dual attention network for scene segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, (2019).

[11] 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*, pp. 5337–5345, (2019).

[12] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris, 'The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback', *arXiv preprint arXiv:1905.12794*, (2019).

[13] 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*, pp. 1463–1471, (2017).

[14] 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*, pp. 1078–1086. ACM, (2017).

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Jun 2016).

[16] Wei-Lin Hsiao and Kristen Grauman, 'Creating capsule wardrobes from fashion images', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7161–7170, (2018).

[17] Jie Hu, Li Shen, and Gang Sun, 'Squeeze-and-excitation networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, (2018).

[18] 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*, pp. 1062–1070, (2015).

[19] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg, 'Hipster wars: Discovering elements of fashion styles', in *European conference on computer vision*, pp. 472–488. Springer, (2014).

[20] Thomas N. Kipf and Max Welling, 'Semi-supervised classification with graph convolutional networks', in *The International Conference on Learning Representations (ICLR)*, (2017).

[21] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang, 'Fashion retrieval via graph reasoning networks on a similarity pyramid', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3066–3075, (2019).

[22] Y. Li, S. Tang, Y. Ye, and J. Ma, 'Spatial-aware non-local attention for fashion landmark detection', in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 820–825, (July 2019).

[23] 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*, pp. 2117–2125, (2017).

[24] 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*, pp. 1096–1104, (2016).

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

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell, 'Fully convolutional networks for semantic segmentation', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, (2015).

[27] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan, 'Occlusion-net: 2d/3d occluded keypoint localization using graph networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7326–7335, (2019).

[28] 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*, pp. 4271–4280, (2018).

[29] 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*, pp. 7794–7803, (2018).

[30] 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*, pp. 172–180. ACM, (2017).

[31] Weijiang Yu, Xiaodan Liang, Ke Gong, Chenhan Jiang, Nong Xiao, and Liang Lin, 'Layout-graph reasoning for fashion landmark detection', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2019).