# Recurrent Attention Model for Pedestrian Attribute Recognition*

**Xin Zhao,**[1] **Liufang Sang,**[1] **Guiguang Ding,**[1] **Jungong Han,**[2] **Na Di,**[1] **Chenggang Yan**[3]
[1]Beijing National Research Center for Information Science and Technology(BNRist)
School of Software, Tsinghua University, Beijing 100084, China
[2]School of Computing & Communications, Lancaster University, UK
[3]Institute of Information and Control Hangzhou Dianzi University, Hangzhou, China
zhaoxin19@gmail.com, slf12thuss@163.com, dinggg@tsinghua.edu.cn,
jungong.han@northumbria.ac.uk, dn15@mails.tsinghua.edu.cn, cgyan@hdu.edu.cn

## Abstract

Pedestrian attribute recognition is to predict attribute labels of pedestrian from surveillance images, which is a very challenging task for computer vision due to poor imaging quality and small training dataset. It is observed that many semantic pedestrian attributes to be recognised tend to show spatial locality and semantic correlations by which they can be grouped while previous works mostly ignore this phenomenon. Inspired by Recurrent Neural Network (RNN)'s super capability of learning context correlations and Attention Model's capability of highlighting the region of interest on feature map, this paper proposes end-to-end Recurrent Convolutional (RC) and Recurrent Attention (RA) models, which are complementary to each other. RC model mines the correlations among different attribute groups with convolutional LSTM unit, while RA model takes advantage of the intra-group spatial locality and inter-group attention correlation to improve the performance of pedestrian attribute recognition. Our RA method combines the Recurrent Learning and Attention Model to highlight the spatial position on feature map and mine the attention correlations among different attribute groups to obtain more precise attention. Extensive empirical evidence shows that our recurrent model frameworks achieve state-of-the-art results, based on pedestrian attribute datasets, i.e. standard PETA and RAP datasets.

## Introduction

Pedestrian attributes, e.g., age, haircut, and footware, are humanly searchable semantic descriptions and can be used as the soft-biometrics in visual surveillance applications such as person re-identification (Layne, Hospedales, and Gong 2012; Liu et al. 2012; Peng et al. 2016), face verification (Kumar et al. 2009), and human identification (Reid, Nixon, and Stevenage 2014). Attributes are robust against viewpoint changes and viewing condition diversity compared to low-level visual features. While pedestrian attribute recognition has been profitably tackled from a face recognition perspective, very few works focus on whole people body. There is inherently challenging to recognise pedestrian attributes from real-world surveillance images subject to the poor imaging quality and small training dataset. High imaging quality and large scale training data are not available for pedestrian attributes. For example, the two largest pedestrian attribute benchmark datasets PETA (Deng et al. 2014) and RAP (Li et al. 2016a) contain only 9500 and 33268 training images. Besides, recognising pedestrian attributes has to cope with images with poor quality, imbalance label and complex appearance variations in surveillance scenes.

Attribute recognition methods include hand-crafted feature methods, CNN methods and CNN-RNN methods. Early attribute recognition methods mainly rely on hand-crafted features like colour and texture (Layne, Hospedales, and Gong 2012; Liu et al. 2012; Jaha and Nixon 2014). Recently, deep learning based attribute models have been proposed due to the capacity to learn more expressive representations (Li, Chen, and Huang 2015; Fabbri, Calderara, and Cucchiara 2017; Liu et al. 2017b), which significantly improve the performance of pedestrian attribute recognition. For example, DeepMAR method (Li, Chen, and Huang 2015) utilizes the prior knowledge in the object topology for attribute recognition and designs a weighted sigmoid cross-entropy loss to deal with the data imbalance problem whilst training attribute recognition model. Multi-directional attention modules are applied in an inception based deep model named HydraPlus Network (Liu et al. 2017b) to take the visual attention into consideration. CNN-RNN methods are proved to be successful in multi-label classification task to mine the dependency of labels (Li et al. 2017; Liu et al. 2017a). A recurrent encoder-decoder framework is introduced into pedestrian attribute recognition task (Wang et al. 2017b), which aims to discover the interdependency and correlation among attributes with Long Short-Term Memory (LSTM) model.

Attributes of pedestrian always show semantic or visual spatial correlation by which they can be grouped. For example, *BoldHair* and *LongHair* cannot occur on the same person while they are both related to the head-shoulders region of a person, so they can be in the same group to be recognised together with a specific attention on the head-shoulders region. Existing methods try to mine the correlations of attributes separately but ignore the spatial neighborhood relationship and the semantic similarity of a group

*This research was supported by the National Key R&D Program of China (2018YFC0806900) and the National Natural Science Foundation of China (No. 61571269). Corresponding author: Guiguang Ding.

of attributes, which can actually improve the performance of pedestrian attribute recognition. The attributes are predicted separately with no attention to the spatial local attribute group, which makes it difficult to process the spatial neighborhood relationship of attributes.

To address these problems, one idea is to take advantage of the interdependency and correlation among attributes (Chen, Gallagher, and Girod 2012; Li, Chen, and Huang 2015; Wang, Zhu, and Gong 2016; 2017; Zhu et al. 2017b), while another idea focuses on particular spatial visual region for relevant attributes with intention to avoid the negative influence of the background (Li et al. 2016b; Liu et al. 2017b). However, these two schemes are mostly studied independently in the existing methods. And the attention model proposed for pedestrian attribute recognition did not take the inter-group correlations of attention region into consideration. For example, the head-shoulders region must be upon the upper body, so that the attention regions associated to them must be highly related to each other.

In this work, we model both intra-group attention locality and inter-group correlations in end-to-end recurrent architectures. A Recurrent Convolutional (RC) framework is proposed to mine the attribute correlations from mid-level convolutional feature maps of attribute groups. And a Recurrent Attention (RA) framework is formulated to recognise pedestrian attributes by group step by step in order to pay attention to both the intra-group and inter-group attention relationship. RA can be considered as a particular case of RC model replacing recurrent convolutional subnet with a recurrent residual spatial attention module. A novel Recurrent Spatial Attention model with ConvLSTM operation is proposed to keep both the 2D spatial locality and the correlations among groups in this work. This RNN based attention model, which applies a sequential grouping attention highlight for attribute prediction, differs from the existing CNN based policy (Li, Chen, and Huang 2015; Fabbri, Calderara, and Cucchiara 2017; Liu et al. 2017b). The participation of Convolutional LSTM attention model leads in unique advantage over existing Fully-connected LSTM attention model (Xu et al. 2015) in keeping the spatial locality of attention heat map. Moreover, the proposed RC and RA methods are end-to-end and easy for online prediction. RA model will learn different attention weights for different attribute groups under a recurrent framework. More latent intra-group and inter-group dependency among attention regions of grouped pedestrian attributes can be exploited, therefore the proposed method obtains better attention weights and outperforms existing methods on the pedestrian attribute recognition task. In summary, we make the following contributions in this paper:

- We put forward novel recurrent approaches termed as RC and RA for pedestrian attribute recognition. To the best of our knowledge, RA is the first work that predicts attributes group by group via mining both intra-group attention locality and inter-group attention correlations.
- Single-model end-to-end architectures, which are easier to train, are adopted without much more preprocessing prior to feature extraction and multi-model voting after attribute prediction.
- ConvLSMT is first introduced by the proposed RC and RA models to model the context correlations and keep the 2D spatial locality for visual attention in the meanwhile.

# Related Work

## Pedestrian Attribute Recognition

Semantic pedestrian attributes have been extensively exploited for person identification (Jaha and Nixon 2014) and re-identification (Layne, Hospedales, and Gong 2012; Liu et al. 2012; Peng et al. 2016). Attribute recognition methods include hand-crafted feature methods, CNN methods and CNN-RNN methods. Earlier methods typically model multiple attributes independently and train a separate classifier for each attribute based on hand-crafted features such as color and texture histograms (Layne, Hospedales, and Gong 2012; Liu et al. 2012; Jaha and Nixon 2014). Later on, inter-attribute correlation is considered as an extra information for improving prediction performance, e.g. graph model based methods to capture attribute co-occurrence likelihoods by using conditional random field or Markov random field (Chen, Gallagher, and Girod 2012; Deng et al. 2015; Shi, Hospedales, and Xiang 2015). But existing graph models are expensive to compute when dealing with a large set of attributes. Restricted to the poor discriminability of hand-crafted features, these methods do not work well.

Recently, deep CNN based methods (Zhu et al. 2015; Li, Chen, and Huang 2015; Sudowe, Spitzer, and Leibe 2015; Fabbri, Calderara, and Cucchiara 2017; Liu et al. 2017b) have been adopted in pedestrian attribute recognition task to learn more expressive representations which significantly improve the performance of pedestrian attribute recognition. DeepMar model (Li, Chen, and Huang 2015) utilizes the prior knowledge in the object topology for attribute recognition and designs a weighted sigmoid cross entropy loss to deal with the data imbalance problem while attribute recognition model training. Spatial attention methods (Liu et al. 2017b; Fabbri, Calderara, and Cucchiara 2017) are proposed to avoid the negative effect of irrelevant image region. Although the CNN based methods learn more expressive pedestrian representations by using deep convolutional network, the CNN based methods are always insufficient in mining the correlations of attributes.

A CNN-RNN based encoder-decoder framework is proposed in (Wang et al. 2017b), which aims to discover the interdependency and correlation among attributes with LSTM model. This method intends to apply attention model for every attribute to be recognised with LSTM, which destroys the spatial information because of the fully-connected operation in LSTM unit. Semantic similarity and the spatial neighborhood of attributes are not taken into account in this method. Additionally, predicting attributes one by one with multi-model voting afterwards is very expensive in computation.

## Recurrent Neural Network(RNN)

RNN is a neural network consisting of an internal hidden state $h \in R^d$ and operating on a variable-length input sequence $X = (x_1, x_2, , x_t, )$. At each time step $t$, the RNN takes sequentially an element $x_t$ of $X$ and then updates its hidden state $h_t$ as:

$$h_t = \phi_\theta(h_{t-1}, x_t) \tag{1}$$

where $\phi_\theta$ denotes the non-linear activation function parameterised by $\theta$.

**Long Short-Term Memory(LSTM).** Long range dependency of input sequence can be captured by LSTM (Hochreiter and Schmidhuber 1997) as recurrent neuron for sequential grouping attribute prediction. LSTM is also effective to handle the common gradient vanishing and exploding problems in training RNN. Particularly, at each time step t, the LSTM is updated using the input $x_t$ and the LSTM previous status $h_{t-1} \in R^d$, and $c_{t-1} \in R^d$ as:

$$\begin{aligned}
f_t &= sigmoid(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc} \odot c_{t-1} + b_f)\\
i_t &= sigmoid(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic} \odot c_{t-1} + b_i)\\
o_t &= sigmoid(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc} \odot c_t + b_o)\\
g_t &= tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g)\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t\\
h_t &= o_t \odot tanh(c_t)
\end{aligned} \tag{2}$$

where $sigmoid(\cdot)$ refers to the logistic sigmoid function, $tanh(\cdot)$ refers to the hyperbolic tangent function and the operator $\odot$ refers to the element-wise vector product. The LSTM contains four multiplicative gating units: forget gate $f \in R^d$, input gate $i \in R^d$, output gate $o \in R^d$, input modulation gate $g \in R^d$, with corresponding matrix and bias parameters to be learned. The memory cell $c_t$ depends on the previous memory cell modulated by the forget gate and the current input. Therefore, LSTM learns to forget its previous memory and exploit its current input selectively. And the output gate $o$ learns how to transfer the memory cell $c_t$ to the hidden state $h_t$. These gates learn to effectively modulate the behaviour of input signal propagation through the recurrent hidden states in order to capture long-term dependency in sequence data.

**Convolutional Long Short-Term Memory.** Convolutional Long Short-Term Memory(ConvLSTM) is an extension of LSTM for spatiotemporal sequence forecasting problem, which is first introduced in precipitation nowcasting problem (Shi et al. 2015) and then used in other spatiotemporal tasks such as video sequence prediction (Villegas et al. 2017) and gesture recognition (Zhu et al. 2017a). ConvLSTM network is proved to capture spatiotemporal correlations better and consistently outperform LSTM in spatiotemporal tasks. It replaces the fully-connected operation with convolution operation to keep the 2D spatial information. The formulation of ConvLSTM is as follows:

$$\begin{aligned}
f_t &= sigmoid(W_{fx} * x_t + W_{fh} * h_{t-1} + W_{fc} \odot c_{t-1} + b_f)\\
i_t &= sigmoid(W_{ix} * x_t + W_{ih} * h_{t-1} + W_{ic} \odot c_{t-1} + b_i)\\
o_t &= sigmoid(W_{ox} * x_t + W_{oh} * h_{t-1} + W_{oc} \odot c_t + b_o)\\
g_t &= tanh(W_{gx} * x_t + W_{gh} * h_{t-1} + b_g)\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t\\
h_t &= o_t \odot tanh(c_t)
\end{aligned} \tag{3}$$

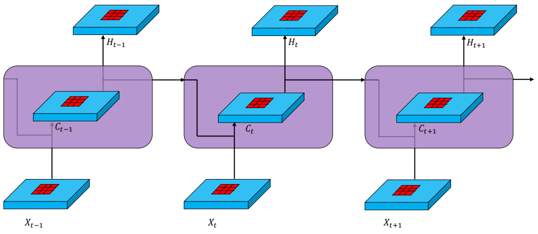


Figure 1: Inner structure of ConvLSTM.

where $*$ denotes the convolution operation and $\odot$ denotes element-wise vector product as before. Fig.1 shows the inner structure of ConvLSTM.

## Attention Model

When facing a complex visual scene, human can efficiently locate region of interest and analyze the scene by selectively processing subsets of visual input. Attention is employed to narrow down the search and speed up the process. Visual attention is a hot topic in computer vision area. It is widely used in object segmentation, object recognition and image caption tasks etc. The visual attention models are mainly categorized into bottom-up models and top-down models.

Bottom-up attention models are based on the image feature of the visual scene. The goal of bottom-up model is to find the fixation points, where it stands out from its surrounding and grabs our attention at first glance. The most representative bottom-up attention model is Faster R-CNN (Ren et al. 2015) for object detection task, while spatial regions are represented as bounding boxes and generated by a region proposal network. And then bottom-up attention methods are applied in general CNN model design, attention mechanisms are used as units in CNN model to improve the fitting ability. Bottom-up attention mechanism units are applied on convolutional feature maps on both spatial (Wang et al. 2017a) and channel-wise (Hu, Shen, and Sun 2017) form.

Top-down attention models are driven by the observer's prior knowledge and current goal. The recurrent attention model (RAM) proposed in (Mnih et al. 2014) simulates the human attention and eye movement mechanism. The basic assumption of a top-down attention model is that there is a bandwidth limit for each glimpse. The model is to predict future eye movements and location to see at next time step. Top-down attention models were first introduced in image caption task by (Xu et al. 2015) and lip reading task by (Chung et al. 2017) and achieved great success. Top-down attention models could deal with a sequence of action with different attention at different time.

# Recurrent Attention Model for Pedestrian Attribute Recognition

## Problem Definition

The definition of grouping pedestrian attribute recognition can be as follows. We are given n images $\{I_1, \ldots, I_n\}$ and each image $I_m$ has $k_m$ visual attribute tags for training. Each

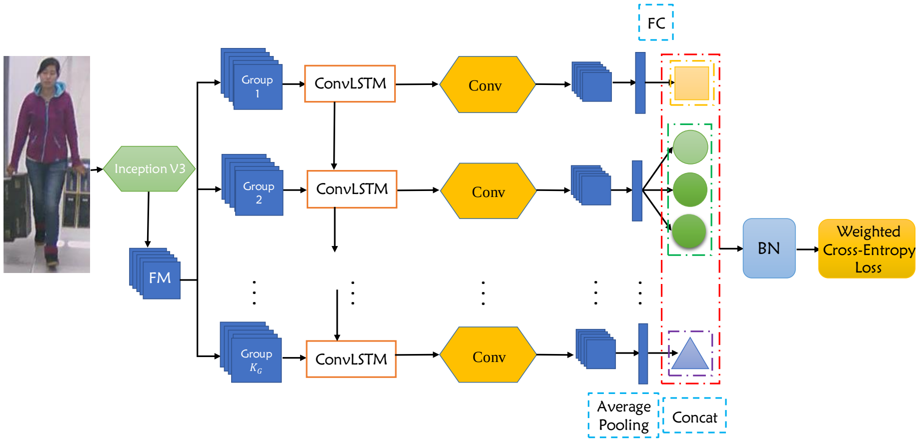


Figure 2: Recurrent Convolutional (RC) Model for Pedestrian Attribute Recognition Framework.

visual attribute tag is in set $\mathcal{T} = \{T_1, \dots, T_{K_T}\}$, where $K_T$ denotes the size of $\mathcal{T}$. And $\mathcal{G} = \{G_1, \dots, G_{K_G}\}$ is a set partitioning of $\mathcal{T}$, where $G_i \bigcap G_j = \o(i \neq j)$ and all combinations of $G_i$ is the entire set $\mathcal{T}$. Tags in the same group are with semantic or spatial constraint with each other. For each image there is a label vector $y_m \in \{0,1\}^{K_T}$ where $y_{mj} = 1$ if $I_m$ has tag $T_j$ and $y_{mj} = 0$ otherwise. We aim to learn attribute recognition models $R^I$: $I \rightarrow \{0,1\}^{K_T}$ to recognise the attributes of image $I_m$.

## Network Architecture

**Recurrent Convolutional Model.** We formulate a Recurrent Convolutional (RC) model to mine the attribute correlations among all the attribute groups. The network architecture is shown in Fig.2. For each pedestrian image $I_m$, we use a CNN to extract the feature map $F$ of it. And then we feed the extracted feature map into a ConvLSTM layer group by group. Next, we calculate the feature map $F_t$ at each time step $t$ by adding a convolutional network after ConvLSTM. And $F_t$ is used for attribute prediction on current attribute group.

**Recurrent Attention Model.** The network architecture is shown in Fig.3. For each pedestrian image $I_m$, we use a CNN to extract the feature map $F$ of it. And then we feed the extracted feature map into the proposed Recurrent Attention Module group by group to calculate the heat map $H_t$ of attention at each time step $t$. The heat map calculated by the attention module is activated by sigmoid activation function. Next, we apply the attention heat map on $F$ with a spatial point-wise multiplication and a residual addition connection to get the highlight feature map $F_t$ for current attribute group as follow

$$F_t = sigmoid(H_t) \otimes F + F \tag{4}$$

where, $\otimes$ denotes spatial point-wise multiplication. Finally, we use the highlight feature map to predict the attribute of this pedestrian image at time $t$.

For both of these two recurrent model above. After all the groups of attribute prediction are finished, we concatenate all of the prediction results together with a batch normalization layer (BN) afterwards. The batch normalization layer first normalizes the prediction vector into a vector with zero mean and unit variance, and then scales it and adds a bias in. The BN operation can be as follow with input $x_i$ and output $y_i$.

$$\begin{aligned} \mu_{\mathcal{B}} &\leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma_{\mathcal{B}}^2 &= \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \\ \hat{x}_i &\leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \\ y_i &\leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i) \end{aligned} \tag{5}$$

If $\beta$ above is zero, the output of BN can be with zero mean. The batch normalize layer is used to balance the positive and negative outputs of this network. The output of batch norm layer is used to compute the weighted sigmoid cross-entropy loss, which will be stated later.

The contrast experiment between DeepMAR (Li, Chen, and Huang 2015) method and RC model shows the effect of recurrent learning with ConvLSTM. And the comparison between RC model and RA model shows the effect of spatial attention.

## Loss Function and Optimization

The sigmoid cross entropy loss, which is defined in Eq.6, is introduced in multi-class classification problem.

$$\begin{aligned} Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T} y_{ij} log(\hat{p}_{ij}) \\ +(1 - y_{ij}) log(1 - \hat{p}_{ij}) \end{aligned} \tag{6}$$

$$\hat{p}_{ij} = \frac{1}{1 + exp(-x_{ij})} \tag{7}$$

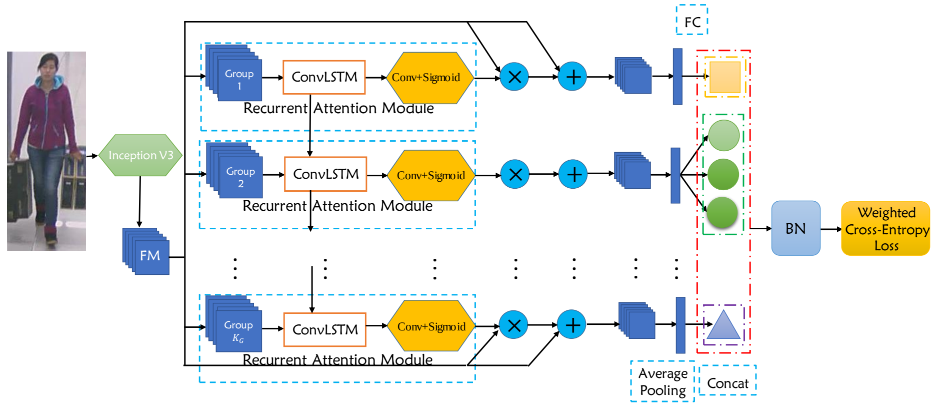


Figure 3: Recurrent Attention (RA) Model for Pedestrian Attribute Recognition Framework. $\otimes$ denotes spatial point-wise multiplication while $\oplus$ denotes point-wise addition.

| Group | Attribute |
|---|---|
| Gender | male or female |
| Age | age 16-30, age 31-45, age 46-60, age 60+ |
| Head | hair length, muffler, hat, glasses |
| Upper Body | clothes style, logo, casual or formal |
| Lower Body | clothes style, casual or formal |
| Footware | footware style |
| Accessories | backpack, messenger bag, plastic bag etc |

Table 1: The groups of 35 binary attributes in PETA dataset

where $\hat{p}_{ij}$ is the output probability for the jth attribute of example $I_i$, $y_{ij}$ is the ground truth label which represents whether $I_i$ has the jth attribute or not, $x_i$ is the output of network fed with $I_i$.

As is stated in (Li, Chen, and Huang 2015), attribute labels do not always have uniform distribution, and sometimes it is more like an unbalanced distribution, especially in pedestrian attribute recognition scenarios. So we use the weighted sigmoid cross-entropy loss proposed in (Li, Chen, and Huang 2015) as follow to solve this problem.

$$\begin{aligned} Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T} w_j y_{ij} log(\hat{p}_{ij}) \\ +(1-y_{ij})log(1-\hat{p}_{ij}) \end{aligned} \quad (8)$$

where $w_j$ is calculated from the positive ratio of jth attribute in the training set as in (Li, Chen, and Huang 2015) which denotes the learning weights of positive samples to deal with the imbalance label. We train the attribute recognition model with Adam algorithm.

# Experiment

## Datasets

For evaluations, we use the two largest publicly available pedestrian attribute datasets: (1) The PEdesTrain Attribute (PETA) (Deng et al. 2014) dataset consists of 19000 person images collected from 10 small-scale person datasets. Each image is labelled with 65 attributes (61 binary + 4 multi-valued). Following the same protocol as (Deng et al. 2015; Li, Chen, and Huang 2015), we divide the whole dataset into three non-overlapping partitions: 9500 for model training, 1900 for verification, and 7600 for model evaluation. And we select 35 attributes from PETA dataset in our experiments. (2) The Richly Annotated Pedestrian (RAP) attribute dataset (Li et al. 2016a) has 41585 images drawn from 26 indoor surveillance cameras. Each image is labelled with 72 attributes (69 binary + 3 multi-valued) as well as viewpoints, occlusions, body parts information. We adopt the same data split as in (Li et al. 2016a): 33268 images for training and the remaining 8317 for test. We evaluate the same 51 binary attributes as (Li et al. 2016a) for a fair comparison. For both datasets, we convert multi-valued attributes into binary attributes. And we group the selected 35 attributes in PETA dataset as follow in Tab.1 while groups of RAP are in Tab.2.

| Group | Attribute |
|---|---|
| Gender | male or female |
| Age | age 16-30, age 31-45, age 45+ |
| Body Shape | slightly fat, standard, slightly thin |
| Role | customer, uniform |
| Head | hair style, hair color, hat, glasses |
| Upper Body | clothes style, clothes color |
| Lower Body | clothes style, clothes color |
| Footware | footware style, footware color |
| Accessories | backpack, single shoulder bag, handbag etc |
| Action | telephoning, gathering, talking, pushing etc |

Table 2: The groups of 51 binary attributes in RAP dataset

## Evaluation

**Metrics.** We use four metrics to evaluate attribute recognition performance. (1) Class-centric: For each attribute tag, we compute the classification accuracy of positive and negative samples respectively, average them to obtain a mean of accuracy termed as **mA**. (2) Instance-centric: For each instance, we measure the attribute prediction **precision** and

| Metric | PETA | | | | RAP | | | |
|---|---|---|---|---|---|---|---|---|
| Method | mA | precision | recall | F1 | mA | precision | recall | F1 |
| ACN (Alexnet) (Sudowe, Spitzer, and Leibe 2015) | 81.15 | 84.06 | 81.26 | 82.64 | 69.66 | 80.12 | 72.26 | 75.98 |
| DeepSar (Alexnet) (Li, Chen, and Huang 2015) | 81.30 | - | - | - | - | - | - | - |
| DeepMar (Alexnet) (Li, Chen, and Huang 2015) | 82.60 | 83.68 | 83.14 | 83.41 | 73.79 | 74.92 | 76.21 | 75.56 |
| DeepMar (Inception-v3) (Li, Chen, and Huang 2015) | 81.50 | **89.70** | 81.90 | 85.68 | 76.10 | 82.20 | 74.80 | 78.30 |
| HydraPlus-Net (Inception-v3) (Liu et al. 2017b) | 81.77 | 84.92 | 83.24 | 84.07 | 76.12 | 77.33 | 78.79 | 78.05 |
| GAPAR (Resnet-50) (Fabbri, Calderara, and Cucchiara 2017) | - | - | - | - | 79.73 | 76.96 | 78.72 | 77.83 |
| CTX CNN-RNN (Li et al. 2017) | 80.13 | 79.68 | 80.24 | 79.68 | 70.13 | 71.03 | 71.20 | 70.23 |
| SR CNN-RNN (Liu et al. 2017a) | 82.83 | 82.54 | 82.76 | 82.65 | 74.21 | 75.11 | 76.52 | 75.83 |
| JRL (Wang et al. 2017b) | 82.13 | 82.55 | 82.12 | 82.02 | 74.74 | 75.08 | 74.96 | 74.62 |
| RC (ours) | 85.78 | 85.42 | 88.02 | **86.70** | 78.47 | **82.67** | 76.65 | **79.54** |
| RA (ours) | **86.11** | 84.69 | **88.51** | 86.56 | **81.16** | 79.45 | **79.23** | 79.34 |
| JRL* (Wang et al. 2017b) | 85.67 | 86.03 | 85.34 | 85.42 | 77.81 | 78.11 | 78.98 | 78.58 |

Table 3: Evaluation on PETA and RAP with bold **best** result and underline second best result. The first group is CNN method with small model such as Alexnet, while the second group is based on larger CNN model(Inception-v3 or Resnet50). The third group is CNN-RNN joint learning method. All above are single model methods, while JRL* uses multi-model ensemble.

**recall** as well as the **F1** score based on precision and recall.

**Competitors.** Our method is compared against 8 state-of-the-art methods including 5 CNN based deep learning attribute recognition methods and 3 CNN-RNN based joint learning models. Attributes Convolutional Network (**ACN**) (Sudowe, Spitzer, and Leibe 2015) trains jointly a CNN model for all attributes, sharing weights and transferring knowledge among different attributes. **DeepSAR** (Li, Chen, and Huang 2015) is a deep model that processes attribute classes individually by training multiple attribute-specific models based on Alexnet. Different with **DeepSAR** (Li, Chen, and Huang 2015), **DeepMAR** (Li, Chen, and Huang 2015) considers additionally inter-attribute correlation by learning all attributes in a single model. We train an inception based **DeepMAR** model for fair comparison. **HydraPlus-Net** (Liu et al. 2017b) is an inception based multi-directional attention network to capture the spatial information of local attribute for better recognition performance. Resnet based Generative Adversarial Models are adopted in pedestrian attribute recognition to improve the accuracy of recognition termed as Generative Adversarial Pedestrian Attribute Recognition (**GAPAR**) (Fabbri, Calderara, and Cucchiara 2017) in this work. Contextual CNN-RNN (**CTX CNN-RNN**) (Li et al. 2017) is a CNN-RNN based sequential prediction model designed to encode the scene context and inter-person social relations for modeling multiple people in an image. Semantically Regularised CNN-RNN (**SR CNN-RNN**) (Liu et al. 2017a) is a state-of-the-art multi-label image classification model that exploits the ground truth attribute labels for strongly supervised deep learning and richer image embedding. Multi-model Joint Recurrent Learning (**JRL**) (Wang et al. 2017b) method is proposed for pedestrian attribute recognition which introduce an encoder-decoder architecture to process image context and attribute correlation.

**Implementation Details.** Our model is trained with tensorflow. And it is finetuned from the Inception-v3 model pretrained from ImageNet image classification task. The optimization algorithm used in training the proposed model is Adam. The initial learning rate of training is 0.1 and reduced to 0.001 by a factor of 0.1 at last.

**Results.** The experiment results of our method and competitors are in Tab.3. The methods in Tab.3 are divided into 4 groups, which are CNN small model, CNN large model, CNN-RNN model and multi-model method. In general, 7 of 8 best results of 4 metrics on RAP and PETA benchmark datasets are located in our **RC** and **RA** models as well as 11 of 16 top-2 results. **RC** model and **RA** model are very close in F1 sore. But **RA** model outperforms **RC** model, improving 0.33% and 2.69% in mA. Our **RA** method outperforms the **SR CNN-RNN** (Liu et al. 2017a) which is the state-of-the-art single model CNN-RNN method in all the four metrics improving 3.28% and 3.91% in mA and F1 of PETA dataset, while the numbers in RAP dataset are 6.95% and 3.51%. And **RA** outperforms **JRL*** (Wang et al. 2017b) in mA and F1 score although **JRL*** is a multi-model ensemble method, improving 0.44% and 1.14% in PETA as well as 3.35% and 0.76% in RAP. Compared to the state-of-the-art large model CNN methods, **RA** is better than **GAPAR** (Fabbri, Calderara, and Cucchiara 2017) method in all the metrics in RAP improving 1.43% in mA where **GAPAR** is better than other methods. The **DeepMAR** (Li, Chen, and Huang 2015) based on Inception-v3 is better in the instance-centric metric than class-centric metric. **RA** also achieves little advantage in instance-centric F1 score(0.88% in PETA and 1.04% in RAP). The experiment result shows clearly the benefit of the proposed **RC** and **RA** approaches in pedestrian attribute recognition. This is mainly due to the capacity of convolutional recurrent learning in mining both the intra-group and inter-group attention correlations.

**Effect of Prediction Order.** The prediction order is an important influence factor for the recognition accuracy because the region of specific attributes to be focused on at the very beginning cannot take advantage of much more relevant recognition result. So we should put the global attributes which can be easily recognised without relying badly on others first. This prediction order is just like first we give a glimpse on a pedestrian image and get the attention region

| Dataset | Metric<br>Method | mA | precision | recall | F1 |
|---|---|---|---|---|---|
| PETA | random order | 85.05 | 85.11 | 87.33 | 86.21 |
|  | global to local | **85.78** | **85.42** | **88.02** | **86.70** |
| RAP | random order | 76.75 | 80.97 | **77.77** | 79.34 |
|  | global to local | **78.47** | **82.67** | 76.65 | **79.54** |

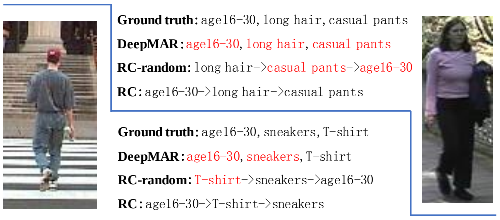
Table 4: The experiment result of logical optimized prediction order (global to local) compared to random prediction order of RC. The **best** result is in bold.

| Dataset | Metric<br>Method | mA | precision | recall | F1 |
|---|---|---|---|---|---|
| PETA | random order | 85.54 | 84.23 | 8807 | 86.11 |
|  | global to local | **86.11** | **84.69** | **88.51** | **86.56** |
| RAP | random order | 80.87 | 78.95 | 79.22 | 79.09 |
|  | global to local | **81.16** | **79.45** | **79.23** | **79.34** |

Table 5: The experiment result of logical optimized prediction order (global to local) compared to random prediction order of RA. The **best** result is in bold.

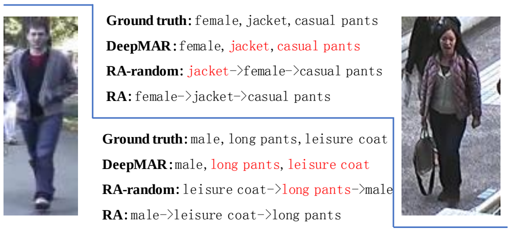
of the whole body and then it gives a summary of global recognition result and then we carefully check each local part to recognise detail attributes. The attention at the beginning is close to a bottom-up attention which is determined by the visual feature of a glimpse and the attention in the end is more likely as a top-down attention which is inferred from previous attention position. In this section, we show the experiment result of logical optimized prediction order from global group to local group and that of a random order which are listed in Tab.4 and Tab.5. The logical optimized prediction order in PETA and RAP is as the order in Tab.1 and Tab.2.

The experiment result listed in Tab.4 and Tab.5 confirms our inference that the logical optimized order is better than a random one, most of the experiment results of logical optimized order are better than the random order. **Visualization of Recurrent Spatial Attention.** Fig.6 shows an example of our proposed Recurrent Attention (RA) model. We take the output of the attention module for 5 different attribute groups, which are head-shoulder, upper-body, lower-body, shoes and accessories. In this example, the corresponding attention regions are highlighted rightly and therefore the local feature can be completely utilized. The effect of attribute correlations are examined more carefully on the RA and RC model performance. Fig.4 and Fig.5 shows attributes prediction examples of RC and RA model from RAP dataset for qualitative analysis which indicates that a proper prediction order is necessary for grouping pedestrian attribute recognition.


**Ground truth:** age16-30, long hair, casual pants
**DeepMAR:** age16-30, long hair, casual pants
**RC-random:** long hair->casual pants->age16-30
**RC:** age16-30->long hair->casual pants

**Ground truth:** age16-30, sneakers, T-shirt
**DeepMAR:** age16-30, sneakers, T-shirt
**RC-random:** T-shirt->sneakers->age16-30
**RC:** age16-30->T-shirt->sneakers

Figure 4: Qualitative analysis of RC prediction order of attributes with wrong predictions in red, right in black from RAP dataset.


**Ground truth:** female, jacket, casual pants
**DeepMAR:** female, jacket, casual pants
**RA-random:** jacket->female->casual pants
**RA:** female->jacket->casual pants

**Ground truth:** male, long pants, leisure coat
**DeepMAR:** male, long pants, leisure coat
**RA-random:** leisure coat->long pants->male
**RA:** male->leisure coat->long pants

Figure 5: Qualitative analysis of RA prediction order of attributes with wrong predictions in red, right in black from RAP dataset.

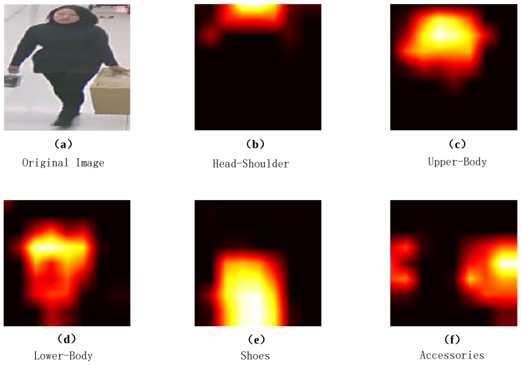


Figure 6: An example of attention heat map of RA model prediction from 5 different attribute groups, which are head-shoulder, upper-body, lower-body, shoes and accessories.

## Conclusion

In this work, we present novel end-to-end deep recurrent models termed as Recurrent Convolutional (RC) model and Recurrent Attention (RA) model for exploring the intra-group spatial neighborhood as well as inter-group pedestrian attribute attention correlations. Convolutional recurrent learning plays a role in both models. And Recurrent Spatial Attention module is adopted in the RA architecture. Our RA model outperforms a wide range of existing pedestrian attribute recognition methods. Extensive experiments demonstrate the advantages of convolutional recurrent learning for mining both semantic and attention correlations on two pedestrian benchmarks. Moreover, a logical optimized prediction order is proved to lead to better results in both RC and RA models.

# References

Chen, H.; Gallagher, A. C.; and Girod, B. 2012. Describing clothing by semantic attributes. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*.

Chung, J. S.; Senior, A. W.; Vinyals, O.; and Zisserman, A. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 3444–3453.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2015. Learning to recognize pedestrian attribute. *CoRR*.

Fabbri, M.; Calderara, S.; and Cucchiara, R. 2017. Generative adversarial models for people attribute recognition in surveillance. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. *CoRR* abs/1709.01507.

Jaha, E. S., and Nixon, M. S. 2014. Soft biometrics for subject identification using clothing attributes. In *IEEE International Joint Conference on Biometrics, Clearwater, IJCB 2014, FL, USA, September 29 - October 2, 2014*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*.

Layne, R.; Hospedales, T. M.; and Gong, S. 2012. Person reidentification by attributes. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016a. A richly annotated dataset for pedestrian attribute recognition. *CoRR*.

Li, Y.; Huang, C.; Loy, C. C.; and Tang, X. 2016b. Human attribute recognition by deep hierarchical contexts. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*.

Li, Y.; Lin, G.; Zhuang, B.; Liu, L.; Shen, C.; and van den Hengel, A. 2017. Sequential person recognition in photo albums with a recurrent network. In *CVPR*.

Li, D.; Chen, X.; and Huang, K. 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, November 3-6, 2015*.

Liu, C.; Gong, S.; Loy, C. C.; and Lin, X. 2012. Person reidentification: What features are important? In *Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings, Part I*.

Liu, F.; Xiang, T.; Hospedales, T. M.; Yang, W.; and Sun, C. 2017a. Semantic regularisation for recurrent image annotation. In *CVPR*.

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017b. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*.

Mnih, V.; Heess, N.; Graves, A.; and Kavukcuoglu, K. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2204–2212.

Peng, P.; Tian, Y.; Xiang, T.; Wang, Y.; and Huang, T. 2016. Joint learning of semantic and latent attributes. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*.

Reid, D. A.; Nixon, M. S.; and Stevenage, S. V. 2014. Soft biometrics; human identification using comparative descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.; Wong, W.; and Woo, W. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 802–810.

Shi, Z.; Hospedales, T. M.; and Xiang, T. 2015. Transferring a semantic representation for person re-identification and search. In *CVPR*.

Sudowe, P.; Spitzer, H.; and Leibe, B. 2015. Person attribute recognition with a jointly-trained holistic CNN model. In *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*.

Villegas, R.; Yang, J.; Hong, S.; Lin, X.; and Lee, H. 2017. Decomposing motion and content for natural video sequence prediction. *CoRR* abs/1706.08033.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017a. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6450–6458.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2017b. Attribute recognition by joint recurrent learning of context and correlation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*.

Wang, J.; Zhu, X.; and Gong, S. 2016. Video semantic clustering with sparse and incomplete tags. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*.

Wang, J.; Zhu, X.; and Gong, S. 2017. Discovering visual concept structure with sparse and incomplete tags. *Artif. Intell.* 250.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2048–2057.

Zhu, J.; Liao, S.; Yi, D.; Lei, Z.; and Li, S. Z. 2015. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics, ICB 2015, Phuket, Thailand, 19-22 May, 2015*.

Zhu, G.; Zhang, L.; Shen, P.; and Song, J. 2017a. Multimodal gesture recognition using 3-d convolution and convolutional LSTM. *IEEE Access* 5:4517–4524.

Zhu, J.; Liao, S.; Lei, Z.; and Li, S. Z. 2017b. Multi-label convolutional neural network based pedestrian attribute classification. *Image Vision Comput.*