Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning *

Xin Zhao[1], Liufang Sang[1], Guiguang Ding[1], Yuchen Guo[1] Xiaoming Jin[1]
[1]Beijing National Research Center for Information Science and Technology(BNRist)
School of Software, Tsinghua University, Beijing 100084, China
{zhaoxin19,yuchen.w.guo}@gmail.com, slf12thuss@163.com, {dinggg,xmjin}@tsinghua.edu.cn

Abstract

Pedestrian attributes 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 semantic pedestrian attributes to be recognised tend to show semantic or visual spatial correlation. Attributes can be grouped by the correlation while previous works mostly ignore this phenomenon. Inspired by Recurrent Neural Network (RNN)'s super capability of learning context correlations, this paper proposes an end-to-end Grouping Recurrent Learning (GRL) model that takes advantage of the intra-group mutual exclusion and inter-group correlation to improve the performance of pedestrian attribute recognition. Our GRL method starts with the detection of precise body region via Body Region Proposal followed by feature extraction from detected regions. These features, along with the semantic groups, are fed into RNN for recurrent grouping attribute recognition, where intra group correlations can be learned. Extensive empirical evidence shows that our GRL model achieves state-of-the-art results, based on pedestrian attribute datasets, i.e. standard PETA and RAP datasets.

1 Introduction

Pedestrian attributes, e.g. age, gender, and clothing are humanly searchable semantic descriptions and can be used as soft-biometrics in visual surveillance applications such as person re-identification [Layne *et al.*, 2012; Liu *et al.*, 2012; Peng *et al.*, 2016], face verification [Kumar *et al.*, 2009], and human identification [Reid *et al.*, 2014]. Attributes are robust against viewpoint changes and viewing condition diversity compared to low-level visual features. While pedestrian attribute recognition has been profitably attacked 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 *et al.*, 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 *et al.*, 2015; Fabbri *et al.*, 2017; Liu *et al.*, 2017b], which significantly improve the performance of pedestrian attribute recognition. For example, DeepMar method [Li *et al.*, 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 a success 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 *BlackHair* 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. Existing methods try to mine the correlations of attributes separately but ignore both the intra-group sematic conflicts and the spatial neighborhood relationship of a group of attributes, which can actually improve the performance of pedestrian attribute recognition. There are two

*This research was supported by the National Natural Science Foundation of China (No. 61571269) and National Basic Research Program of China(2015CB352300). Corresponding author: Guiguang Ding. Co first author: Liufang Sang

Attribute	Age	Body Shape
Mutex Co-occur	158	562
None of Any	134	0
Total Error Number	292	562
Total Image Number	8317	8188
Semantic Error Rate	3.51%	6.86%

Table 1: The rate of semantic conflict of *Age* and *Body Shape* in RAP test set in the prediction result of DeepMar

types of semantic conflicts in the attribute prediction result. For example, a person cannot have the attribute *age 16-30* and *age 31-45* at the same time. If this occurs in the prediction result, it is termed as mutex co-occur. A person cannot be neither male or female. If this occurs, it is termed as none of any. Tab.1 shows the rate of semantic conflict of attributes *Age* and *Body Shape* with an existing pedestrian recognition method DeepMar[Li *et al.*, 2015]. Moreover, the attributes are predicted separately with no attention to the spatial local attribute group, which makes it difficult to process spatial neighborhood relationship of attributes.

To address these problems, one idea is to take advantage of the interdependency and correlation among attributes [Chen *et al.*, 2012; Li *et al.*, 2015; Wang *et al.*, 2016; 2017a; Zhu *et al.*, 2017], 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.

In this work, we model both intra-group semantic mutual exclusion and inter-group correlations in an end-to-end recurrent architecture. A Grouping Recurrent Learning (GRL) 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 relationship (including semantic and spatial). A novel grouping attribute recognition network is introduced which is specifically designed for sequential pedestrian attribute prediction by group. This RNN based model, which applies a sequential grouping attribute prediction, differs from the existing CNN based attribute prediction policy [Li *et al.*, 2015; Fabbri *et al.*, 2017; Liu *et al.*, 2017b]. Moreover, it is an end-to-end single-model method with no need for preprocessing, compared to the multi-model Joint Recurrent Learning (JRL) method [Wang *et al.*, 2017b]. More latent intra-group and inter-group dependency among grouped pedestrian attributes can be exploited, therefore the proposed method outperforms existing methods on the pedestrian attribute recognition task. In summary, we make the following contributions in this paper:

- We put forward a novel approach termed as GRL for pedestrian attribute recognition. To the best of our knowledge, it is the first work that predicts attributes group by group via mining both semantic and spatial correlations in attribute groups.
- A single-model end-to-end architecture, which is easier to train, is adopted without much more preprocessing prior to feature extraction and multi-model voting after attribute prediction.
- A recurrent learning method is proposed for mining the inter-group attribute correlations.

2 Related Work

2.1 Pedestrian Attribute Recognition

Semantic pedestrian attributes have been extensively exploited for person identification [Jaha and Nixon, 2014] and re-identification [Layne *et al.*, 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 *et al.*, 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 *et al.*, 2012; Deng *et al.*, 2015; Shi *et al.*, 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 *et al.*, 2015; Sudowe *et al.*, 2015; Fabbri *et al.*, 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 *et al.*, 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 *et al.*, 2017] are proposed to avoid the negative effect of irrelevant image region. Although the CNN based methods learn more more expressive pedestrian representations by using deep convolutional network, they 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. However, semantic mutex constraint and the spatial neighborhood are not taken into account in this method. Additionally, predicting attributes one by one with multi-model voting afterwards is very expensive in computation.

2.2 Body Region Proposal

Body region proposal problem can be considered as an object detection problem. Region-based Convolutional Network (RCNN) methods are proposed in object detection task and achieve success [Girshick, 2015; Ren *et al.*, 2015]. Fast R-CNN method is proposed for a fast object detection by sharing computation of convolutional feature map with ROI(Region of interest) pooling layer. And then a fully-convolutional object detection framework that simultaneously predicts object bounds and objectness scores at each position termed as Region Proposal Networks (RPN) [Ren

Figure 1: Body region proposal pipeline. A pedestrian image is fed into a fully convolutional network (FCN) for joints detection. And the positions of joints are adopted in body region generation for a detection of body sub-region of pedestrian.

et al., 2015] is adopted for real-time object detection, which further improves the detection speed.

Body region feature extraction plays an important role in distinguishing individuals. Local details can be better described by the region features. A complete body region proposal consists of two steps, which are body joint localization and body region proposal. RPN is introduced for body region proposal [Zhao *et al.*, 2017] in a person re-identification task for more accuracy local features, where a fully convolutional network (FCN) is adopted to predict the localization of body joints, and the position of joints is used for body region generation. In this work, we use this body region proposal method to detect body parts in pedestrian images and use the relevant spatial region for grouping attribute recognition. Fig.1 shows the pipeline of body region proposal.

3 Background

3.1 Recurrent Neural Network

RNN is a neural network consisting of an internal hidden state $h \in R^d$ and operating on a variable-length input sequence $X = (x_1, x_2, \dots, x_t, \dots)$. 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 ϕ_θ denotes the non-linear activation function parameterised by θ.

3.2 Long Short-Term Memory(LSTM) Model

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 updates using the input x_t and the LSTM previous status $h_{t-1} \in R^d$, and $c_{t-1} \in R^d$ as:

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

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 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.

4 Grouping Joint Recurrent Learning for Pedestrian Attribute Recognition

4.1 Problem Definition

Grouping pedestrian attribute recognition can be defined as follows. We are given n training images $\{I_1, \dots, I_n\}$ and each image I_m has k_m visual attribute tags. Each visual attribute tag belongs to set $\mathcal{T} = \{T_1, \dots, T_{K_T}\}$, where K_T is the size of $\mathcal{T}$. And $\mathcal{G} = \{G_1, \dots, G_{K_G}\}$ is a set partitioning of $\mathcal{T}$, where $G_i \bigcap G_j = \emptyset (i \neq j)$ 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.

4.2 Network Architecture

The network architecture is shown in Fig.2. For each pedestrian image I_m, we use a fully convolutional network to detect the joints of the body. And then we use a body region proposal network to generate the head, upper body and lower body region of this person. I_m is fed into an inception based CNN model with the results of body region proposal for ROI average pooling. The ROI average pooling operation can extract the features in the particular region from the feature map extracted from inception module, so that we can take advantage of the spatial neighborhood of relevant group more easily. For example, the hair style, glasses and hat are all in the head region, they are in the same group and these attributes are predicted at the same time. There are attributes related to the whole body region with semantic relationships, by which they can be grouped into several groups in order to take advantage of the semantic correlation with each other.

As is shown in Fig.2, all attributes in the same group share the same fully connected feature, and the features of all the groups are fed into a LSTM unit for recurrent grouping attribute prediction. Each output of LSTM is fully connected into a prediction vector which has the same dimension as the number of attributes in the relevant group. The prediction vectors are connected together with a batch normalization layer (BN) afterwards. The batch normalization layer first

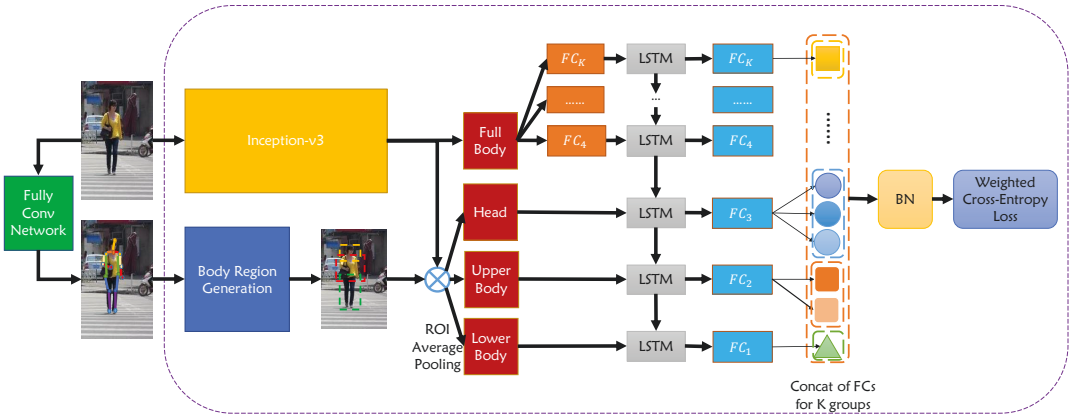

Figure 2: Grouping Recurrent Learning architecture including Body Region Proposal and Recurrent Grouping Attributes Prediction.

normalizes the prediction vector into a vector with zero mean and unit variance, and then scales it and adds a bias in.

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 in Section 4.3.

4.3 Loss Function and Optimization

The sigmoid cross entropy loss, which is defined in Eq.3, is introduced in multi-class classification problem.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T} y_{ij} log(\hat{p}_{ij}) + (1-y_{ij}) log(1-\hat{p}_{ij}) \quad (3)$$

$$\hat{p}_{ij} = \frac{1}{1+exp(-x_{ij})} \quad (4)$$

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 *et al.*, 2015], attributes do not always have uniform distribution, and sometimes it is more like an unbalanced distribution, especially in surveillance scenarios. So we use the weighted sigmoid cross-entropy loss proposed in [Li *et al.*, 2015] as follow to solve this problem.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_T} w_j y_{ij} log(\hat{p}_{ij}) + (1-y_{ij}) log(1-\hat{p}_{ij}) \quad (5)$$

$$w_j = exp(p_j) \quad (6)$$

where p_j is the positive ratio of jth attribute in the training set. w_j donates the learning weights of positive samples to deal with the imbalance label. We train the attribute recognition model with Stochastic Gradient Descent (SGD) algorithm.

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 2: The groups of 35 binary attributes in PETA dataset

Group	Attribute
Gender	male or female
Age	age 16-30, age 31-45, age 45+, gender
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 3: The groups of 51 binary attributes in RAP dataset

5 Experiment

5.1 Datasets

For evaluations, we used 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 *et al.*, 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

Method \ Metric	PETA				RAP			
	mA	precision	recall	F1	mA	precision	recall	F1
ACN (Alexnet) [Sudowe *et al.*, 2015]	81.15	84.06	81.26	82.64	69.66	<u>80.12</u>	72.26	75.98
DeepSar (Alexnet) [Li *et al.*, 2015]	81.30	-	-	-	-	-	-	-
DeepMar (Alexnet) [Li *et al.*, 2015]	82.60	83.68	83.14	83.41	73.79	74.92	76.21	75.56
DeepMar (Inception-v3) [Li *et al.*, 2015]	81.50	**89.70**	81.90	<u>85.68</u>	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 *et al.*, 2017]	-	-	-	-	<u>79.73</u>	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
GRL (ours)	**86.70**	84.34	**88.82**	**86.51**	**81.20**	77.70	**80.90**	**79.29**
JRL* [Wang *et al.*, 2017b]	<u>85.67</u>	<u>86.03</u>	<u>85.34</u>	85.42	77.81	78.11	<u>78.98</u>	<u>78.58</u>

Table 4: Evaluation on PETA and RAP with bold **best** result and underline <u>second best</u> 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.

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.2 while groups of RAP are in Tab.3.

5.2 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 **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 *et al.*, 2015] trains jointly a CNN model for all attributes, and sharing weights and transfer knowledge among different attributes. **DeepSAR** [Li *et al.*, 2015] is a deep model that processes attribute classes individually by training multiple attribute-specific models based on Alexnet. Different with **DeepSAR** [Li *et al.*, 2015], **DeepMar** [Li *et al.*, 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 *et al.*, 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 body region proposal network is trained with **MPII** human pose dataset [Andriluka *et al.*, 2014] as the model stated in [Zhao *et al.*, 2017]. The optimization algorithm used in training the proposed model is SGD. 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.4. The methods in Tab.4 are divided into 4 groups, which are CNN small model, CNN large model, CNN-RNN model and multi-model method. Our **GRL** 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 3.87% and 3.86% in mA and F1 of PETA dataset, while the numbers in RAP dataset are 6.99% and 3.46%. And **GRL** outperforms **JRL*** [Wang *et al.*, 2017b] in mA and F1 score although **JRL*** is a multi-model ensemble method, improving 1.03% and 1.09% in PETA as well as 3.39% and 0.71% in RAP. Compared to large model CNN method, **GRL** is better than **GAPAR** [Fabbri *et al.*, 2017] method in all the metrics in RAP improving 1.47% in mA where **GAPAR** is better than other methods. The **DeepMar** [Li *et al.*, 2015] based on Inception-v3 is better in the instance-centric metric than class-centric metric. **GRL** also achieves little advantage in instance-centric F1 score(0.83% in PETA and 0.99% in RAP). The experiment result shows clearly the benefit of the proposed **GRL** approach in pedestrian attribute recognition. This is mainly due to the capacity of **GRL** in mining both the intra-group and inter-group correlations.

Dataset	Metric / Method	mA	precision	recall	F1
PETA	Baseline	81.50	**89.70**	81.90	85.68
	GRL(no ROI)	85.66	84.98	87.20	86.01
	GRL	**86.70**	84.34	**88.82**	**86.51**
RAP	Baseline	76.10	**82.20**	74.80	78.30
	GRL(no ROI)	79.68	78.60	78.90	78.70
	GRL	**81.20**	77.70	**80.90**	**79.29**

Table 5: The experiment result of Full GRL system and GRL system without ROI pooling from Body Region Proposal compared to the DeepMar (Inception-v3) method as baseline. The **best** result is in bold.

Dataset	Metric / Method	mA	precision	recall	F1
PETA	random order	85.81	84.32	87.49	85.89
	global to local	**86.70**	**84.34**	**88.82**	**86.51**
RAP	random order	80.02	**77.99**	79.21	78.70
	global to local	**81.20**	77.70	**80.90**	**79.29**

Table 6: The experiment result of logical optimized prediction order (global to local) compared to random prediction order of GRL. The **best** result is in bold.

5.3 Further Analysis and Discussions

Effect of Body Region Proposal and Grouping Recurrent Recognition. The improvement of GRL method comes from two aspects which are spatial attention and sematic correlation mining. In this section, we will discuss how much improvement this two aspects bring. First of all, we do not make use of the information of Body Region Proposal. The full body features from the CNN are fed into the LSTM unit directly without ROI average pooling. Compared to the baseline method, which is an inception based DeepMar model, this method only uses Grouping Recurrent Recognition. So the improvement from the baseline donates the effect of Grouping Recurrent Recognition. Compared to the full GRL system, this method is lack of the spatial attention from Body Region Proposal. So the difference from the full GRL system donates the effect of Body Region Proposal. The experiment results are listed in Tab.5.

From Tab.5 we can see that GRL without ROI achieves a gain of 4.16% in mA and 0.33% in F1 score on PETA from baseline, as well as 3.58% and 0.40% on RAP. The full system improves 1.04% in mA and 0.50% in F1 score, while the numbers are 1.52% and 0.59% on RAP. We can see that both of the two components of GRL make sense and GRL achieves more improvement in mA from Grouping Recurrent Recognition than Body Region Proposal.

Effect of Prediction Order of LSTM. The prediction order is an important influence factor for the recognition accuracy because the attributes to be recognised at the very beginning cannot observe much more relevant recognition result. So we should put the global attributes which can be easily recognised without relying badly on others first. For example, gender and age can easily be recognised even though many other attributes are not clear. The recognition of gender is very helpful in predicting some other related attributes

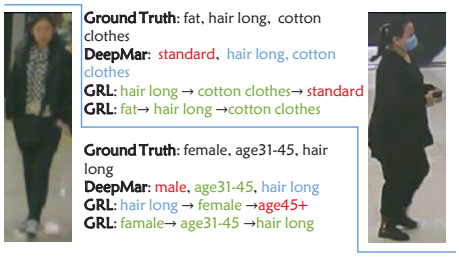

Figure 3: Qualitative analysis of prediction order of attributes with wrong predictions in red, right in green and missed in blue. These two examples are from RAP.

(e.g. clothing and footware). 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.6. The logical optimized prediction order in PETA and RAP is as the order in Tab.2 and Tab.3.

The experiment result listed in Tab.6 confirms our inference that the logical optimized order is better than a random one, which brings an improvement of 0.89% and 0.62% as well as 1.08% and 0.59% in mA and F1 of both datasets.

The effect of attribute correlations are examined more carefully on the GRL model performance. Fig.3 shows two examples from RAP dataset for qualitative analysis which indicates that a proper prediction order is necessary for grouping pedestrian attribute recognition. Non-sequence prediction model DeepMar misses *hair long* and gets wrong prediction of the age for the left person. And it also misses *hair long* and *cotton clothes* for the right one. In contrast, our GRL method gets the right prediction perfectly when the prediction is in a right order. For example, if the model has come to a conclusion that a person is female, the chance she is with long hair is higher. So when predicting the attribute group related to hair after getting gender information, the attribute hair long will be rightly recognised. The prediction for some global attributes (e.g. *age* and *body shape*) is not so much related to the prediction result of other local attributes, so that they should be determined by the overall vision features as much as possible to avoid the misleading of wrong local attribute tags. Attributes like that should be predicted at the beginning of the sequence. GRL gets wrong body shape and age results when predicting in a wrong order as is shown in Fig.3.

6 Conclusion

In this work, we present a novel end-to-end deep Grouping Recurrent Learning (GRL) model for exploring the intra-group relationship including semantic dependency and spatial neighborhood as well as inter-group pedestrian attribute correlations. Spatial attention from Body Region Proposal and Grouping Recurrent Recognition are adopted in the GRL architecture. Our GRL model outperforms a wide range of existing pedestrian attribute recognition methods. Extensive experiments demonstrate the advantages of spatial attention from Body Region Proposal and Grouping Recurrent Recognition on two pedestrian benchmarks. Moreover, a logical optimized prediction order is proved to lead to a better result.

References

[Andriluka *et al.*, 2014] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[Chen *et al.*, 2012] Huizhong Chen, Andrew C. Gallagher, and Bernd Girod. 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*, 2012.

[Deng *et al.*, 2014] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 2014.

[Deng *et al.*, 2015] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning to recognize pedestrian attribute. *CoRR*, 2015.

[Fabbri *et al.*, 2017] Matteo Fabbri, Simone Calderara, and Rita Cucchiara. 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*, 2017.

[Girshick, 2015] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[Jaha and Nixon, 2014] Emad Sami Jaha and Mark S. Nixon. 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*, 2014.

[Kumar *et al.*, 2009] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, 2009.

[Layne *et al.*, 2012] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Person re-identification by attributes. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, 2012.

[Li *et al.*, 2015] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. 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*, 2015.

[Li *et al.*, 2016a] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *CoRR*, 2016.

[Li *et al.*, 2016b] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. 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*, 2016.

[Li *et al.*, 2017] Yao Li, Guosheng Lin, Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Sequential person recognition in photo albums with a recurrent network. In *CVPR*, 2017.

[Liu *et al.*, 2012] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings, Part I*, 2012.

[Liu *et al.*, 2017a] Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang, and Changyin Sun. Semantic regularisation for recurrent image annotation. In *CVPR*, 2017.

[Liu *et al.*, 2017b] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

[Peng *et al.*, 2016] Peixi Peng, YongHong Tian, Tao Xiang, Yaowei Wang, and Tiejun Huang. 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*, 2016.

[Reid *et al.*, 2014] Daniel A. Reid, Mark S. Nixon, and Sarah V. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 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*, 2015.

[Shi *et al.*, 2015] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015.

[Sudowe *et al.*, 2015] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. 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*, 2015.

[Wang *et al.*, 2016] Jingya Wang, Xiatian Zhu, and Shaogang Gong. 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.*, 2016.

[Wang *et al.*, 2017a] Jingya Wang, Xiatian Zhu, and Shaogang Gong. Discovering visual concept structure with sparse and incomplete tags. *Artif. Intell.*, 250, 2017.

[Wang *et al.*, 2017b] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 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*, 2017.

[Zhao *et al.*, 2017] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[Zhu *et al.*, 2015] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics, ICB 2015, Phuket, Thailand, 19-22 May, 2015*, 2015.

[Zhu *et al.*, 2017] Jianqing Zhu, Shengcai Liao, Zhen Lei, and Stan Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image Vision Comput.*, 2017.