Pedestrian Attribute Recognition via Hierarchical Multi-task Learning and Relationship Attention

Lian Gao
Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University.
Beijing, China
gaolian@buaa.edu.cn

Di Huang*
Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University.
Beijing, China
dhuang@buaa.edu.cn

Yuanfang Guo
School of Computer Science and Engineering, Beihang University, Beijing, China.
Beijing, China
andyguo@buaa.edu.cn

Yunhong Wang
Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University.
Beijing, China
yhwang@buaa.edu.cn

ABSTRACT

Pedestrian Attribute Recognition (PAR) is an important task in surveillance video analysis. In this paper, we propose a novel end-to-end hierarchical deep learning approach to PAR. The proposed network introduces semantic segmentation into PAR and formulates it as a multi-task learning problem, which brings in pixel-level supervision in feature learning for attribute localization. According to the spatial properties of local and global attributes, we present a two stage learning mechanism to decouple coarse attribute localization and fine attribute recognition into successive phases within a single model, which strengthens feature learning. Besides, we design an attribute relationship attention module to efficiently capture and emphasize the latent relations among different attributes, further enhancing the discriminative power of the feature. Extensive experiments are conducted and very competitive results are reached on the RAP and PETA databases, indicating the effectiveness and superiority of the proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**.

KEYWORDS

pedestrian attribute recognition, deep learning, multi-task learning and visual attention

*indicates the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351003

ACM Reference Format:
Lian Gao, Di Huang, Yuanfang Guo, and Yunhong Wang. 2019. Pedestrian Attribute Recognition via Hierarchical Multi-task Learning and Relationship Attention. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351003

1 INTRODUCTION

Nowadays, video surveillance systems have been widely employed with different security demands in various public and private facilities and places, including squares, malls, railway stations, airports, residential buildings, libraries, etc. Pedestrians are major targets in surveillance videos and automatic pedestrian analysis is important to many applications, such as key person indexing, criminal trajectory tracking, and abnormal behavior detection, where Pedestrian Attribute Recognition (PAR) plays a fundamental role. PAR aims to predict intrinsic characteristics (e.g. "gender", "age") as well as appearance properties (e.g. "clothes style", "accessory") of persons and has received increasing attentions in recent years.

PAR is a challenging task with a number of intractable problems. On the one hand, it has to handle the common reputed issues in the field of computer vision, involving changes in ambient illumination, camera viewpoint, video resolution, person gesture, and external occlusion. On the other hand, to satisfy diverse requirements, the number of attributes concerned becomes larger and larger. The attributes convey rich semantic information at different levels. In general, local attributes (e.g. "hair style" and "accessory") are related to low-level or mid-level appearance features of certain regions, while global attributes (e.g. "gender") require holistic representation with special areas highlighted (e.g. face, hair, and torso), probably corresponding to some local attributes. This complexity of attribute relationship makes PAR even more difficult. Figure 1 shows some examples of pedestrians and typical attributes.

Early studies on PAR follow the detection pipeline, which firstly extracts handcrafted features of candidate regions and then feeds them into classifiers for prediction, and demonstrate promising results [2, 12]. Unfortunately, they can only handle single or very few similar attributes, as the features used are ad-hoc and not easy to be

Figure 1: Examples of pedestrian attributes including both global and local ones from the PETA and RAP datasets.

generalized to other applications. Recently, with the rapid increase in the number of attributes, PAR has been formulated as a multi-label classification problem, where multiple attributes are inferred on an image patch of a person given by detectors. To address this, many efforts, which attempt to improve the accuracy by enhancing the discriminative power of features, have been made [7, 26, 29]. Although performance gains are delivered, in particular by Convolutional Neural Networks (CNN), these solutions have a serious bottleneck. To be specific, local attributes are position-aware, but limited by current benchmarks (e.g. RAP and PETA), there are no ground-truth locations annotated, making the model built in a weakly-supervised manner. Obviously, if the attribute is located in a wrong place, the feature cannot be learned accurately.

More recently, several works [15, 17, 27, 28] have pointed out the importance of spatial distributions of attributes. They optimize attribute localization by adding auxiliary branches to provide bounding boxes or sparse landmarks of major body parts. However, these constraints are still too coarse and tend to incur much unrelated information. Besides, attribute localization and classification are usually coupled in existing deep learning models, sharing the features from the same backbone. However, as introduced, the attributes are at different semantic levels. Local attributes possess explicit spatial locations, while those of global attributes seem not clear and their prediction probably requires that of local ones. Due to this ambiguous hierarchy, it is problematic to simultaneously learn features of all the attributes in a single network for both localization and classification. It is likely to confuse the model and thus impairs feature distinctiveness.

On the other side, some investigations show that the relationship among attributes can be considered as extra clues to ameliorate PAR results [3, 6, 22, 24, 28]. Common sense says that women are more likely to have long hair and men seldom wear long dresses. Nevertheless, the methods mine such relationship by using expertise or exhaustive search, and the results are either incomplete or inefficient. Latest progress encodes spatial relationships of attributes through Recurrent Neural Networks (RNN), achieving a better trade-off. But the inconvenience in end-to-end CNN and RNN joint training often makes the model not strong enough, leaving room for improvement.

This paper proposes a novel approach to PAR, which addresses the problem by an end-to-end hierarchical deep learning network. Firstly, we introduce human body semantic segmentation into PAR to compose a multi-task learning model, aiming to add pixel-level supervision in attribute location and thus improve feature learning. Secondly, considering the spatial difference between local and global attributes, we construct a hierarchical two-staged deep CNN. The first stage is designed to locate basic body parts that are relevant to some common local attributes at an early time and helps to facilitate subsequent feature learning at more accurate positions. The second stage follows to build more powerful features through latter layers for both global and local attribute prediction. Thirdly, to sufficiently extract semantic relationships among different attributes and further strengthen features, based on the two-staged multi-task framework, we propose an attribute relationship attention module that refines final prediction by additional weight assignment. To validate the proposed method, we carry out extensive experiments on two public datasets, named RAP [16] and PETA [5], and very competitive results are achieved, which verifies its effectiveness.

In summary, our contributions are as below:

(1) We propose an end-to-end deep multi-task learning approach to PAR, which combines semantic segmentation for fine-grained pixel-level attribute localization in feature learning.
(2) We propose a two-staged learning strategy for PAR, which strengthens feature learning by decoupling coarse attribute localization and fine attribute recognition in successive phases within a single model.
(3) We propose an attribute relationship attention module to capture relationships among different attributes, which further enhances the feature for better prediction.

2 RELATED WORK

In this section, we briefly review the methods in PAR, deep multi-task learning, and deep visual attention mechanism, respectively.

2.1 Pedestrian Attribute Recognition

In PAR, many works focus on discriminative feature extraction from input images or videos. [2] represents each input pedestrian image as a collection of patch features, and employs an ensemble learning classifier for gender recognition. To model appearance variations of attributes, [12] builds a rich appearance dictionary of human parts by decomposing image lattice into multi-scale overlapping windows and iteratively refining local appearance templates. Recently, deep learning has drawn extensive attentions due to its powerful representation ability. As PAR is currently a weakly-supervised problem, a number of investigations make use of spatial constraints to improve feature learning. Some approaches use human pose information in PAR. [27] presents an additional pose normalization network, which generates aligned images/frames, to avoid appearance shifts in PAR. [29] divides the pedestrian image into 15 overlapping patches and applies multiple sub-networks to extract deep features. These features are then fused to predict attribute labels. [26] generates a set of mid-level features through max pooling based detection layers and then predicts attribute labels

by regressing detection response magnitudes. More importantly, their network can infer the locations and rough shapes of pedestrian attributes by performing clustering on a fusion of activation maps. [15] combines the features obtained from pose estimation and image-level human body localization, and builds a pose-guided prediction network. [7] constructs a Generative Adversarial Network (GAN) model to handle the images with poor resolutions and strong occlusions. These approaches confirm the importance of spatial properties of attributes and indeed boost the performance. But the spatial information used in them is coarse, quite different from the true pixel-level distribution.

Recent studies explore the relationships among attributes. [23] and [14] propose ACN and DeepMAR, respectively, to jointly predict multiple attributes and achieve superior performance compared to individual attribute predictors. These methods mainly consider the co-occurrence dependencies of the attributes. However, they tend to ignore other high-order correlations among them. There exist some methods which model high-order semantic correlations. Based on graph models, [6], [3] and [22] compute the attribute co-occurrence likelihoods according to conditional random field or Markov random field. These methods usually incur high memory consumptions and computational complexities when the number of attributes is large. [24] briefs a CNN-RNN-RNN architecture to extract the contexts and correlations of the attributes, and achieves the state-of-the-art accuracy. Unfortunately, [24] only individually employs a feature extractor and an attribute predictor without utilizing the end-to-end strategy to jointly optimize the two phases. [28] constructs an end-to-end framework to capture semantic features from certain human body regions and then recognizes the grouped attributes with an RNN model. Although it employs the local patches to directly model the attributes, the generated proposals only provide coarse spatial information, thus leaving space for improvement.

2.2 Deep Multi-task Learning

Multi-Task Learning (MTL) aims to tackle multiple related tasks in a single model, accounting for the correlations among them. Recently, deep MTL has been widely studied. [20] proposes a unified model for face detection, pose estimation, and landmark prediction in wild images. [1] introduces a recurrent network to simultaneously perform road segmentation, car detection, and road classification. [18] designs a cross-stitching network to automatically achieve the best configurations of the shared layers. These approaches usually make use of the low-order correlations among the tasks and tend to ignore the high-order semantic cues. Besides, the weight sharing mechanism is adopted in these methods and it is considered as the only way to model the correlations among the tasks. Unfortunately, the correlations among tasks are not limited to sharing primary processing operations. Therefore, it is hard to well encode high-level cues in current deep MTL models.

In this paper, we model PAR, along with semantic segmentation, as a multi-task learning problem. Semantic segmentation is expected to provide fine-grained spatial information to improve the performance of PAR, where high-level cues between different tasks are captured within the hierarchical two-staged framework by a specially designed attention module.

2.3 Deep Visual Attention Mechanism

Visual attention mechanism is widely used in computer vision tasks for performance refinement via mining potential spatial cues.

Hard attention, which is exploited by [19] and [25], locates the most important region in the image with a binary mask. Since hard attention cannot be optimized in an end-to-end manner, soft attention, which assigns an independent weight to each pixel, has recently become popular. [4] computes multi-scale features based on the soft visual attention mechanism. [17] designs a multi-directional attention network to improve features for PAR and person re-identification. [21] extracts and aggregates visual attention masks at different scales to encode spatial and semantic correlations. Existing methods always regard the attention mechanism as a powerful add-on for feature extraction; however, semantic correlations between intermediate features are not modeled, which may further benefit feature learning.

To better capture the relationships among different attributes in PAR, in this paper, an attribute relationship attention module is proposed to refine attribute predictions.

3 METHODOLOGY

To deal with the unsolved issues, we propose an end-to-end hierarchical deep multi-task learning framework to improve the overall PAR accuracy. Figure 2 shows the entire framework of the proposed method, and it consists of three components, the multi-task learning network (Sec. 3.1), the hierarchical learning mechanism (Sec. 3.2), and the attribute relationship attention module (Sec. 3.3).

3.1 Multi-task Learning

As introduced in Sec. 1, a major difficulty of PAR is induced by the current weakly-supervised learning paradigm, where no accurate locations of attributes provided in feature learning. Although some investigations [26, 28] attempt to alleviate it by exploiting additional cues from bounding boxes and key landmarks, the spatial information is not precise enough. Therefore, we propose to jointly use the local features with respect to the attribute localization maps (i.e. confidence maps), which convey essential spatial priors, to explicitly model the mappings from specific human body parts to the corresponding attributes for classification. For this purpose, we construct an end-to-end deep multi-task learning network with two highly correlated tasks, i.e. PAR and human body semantic segmentation, aiming to reinforce the former by the latter with more accurate pixel-level attribute localizations.

Specifically, our structure starts from the typical single-task PAR network, which takes the ResNet-50 model as the backbone [9]. It consists of 13 ResBlocks (2a-2c, 3a-3d and 4a-4f blocks) of the original ResNet-50 model, and the layers behind ResBlock 4f are abandoned to avoid pedestrian attribute and segmentation feature maps being too small. In a ResBlock, there are typically 3 stacked convolution layers, and each one is followed by a ReLU activation layer. Meanwhile, batch normalization [10] is employed before the activation layers. Element-wise addition operation is performed on the input and output features of each ResBlock. In the ResBlocks which perform down-sampling, a convolution operation is applied to the input features before they are added to outputs.

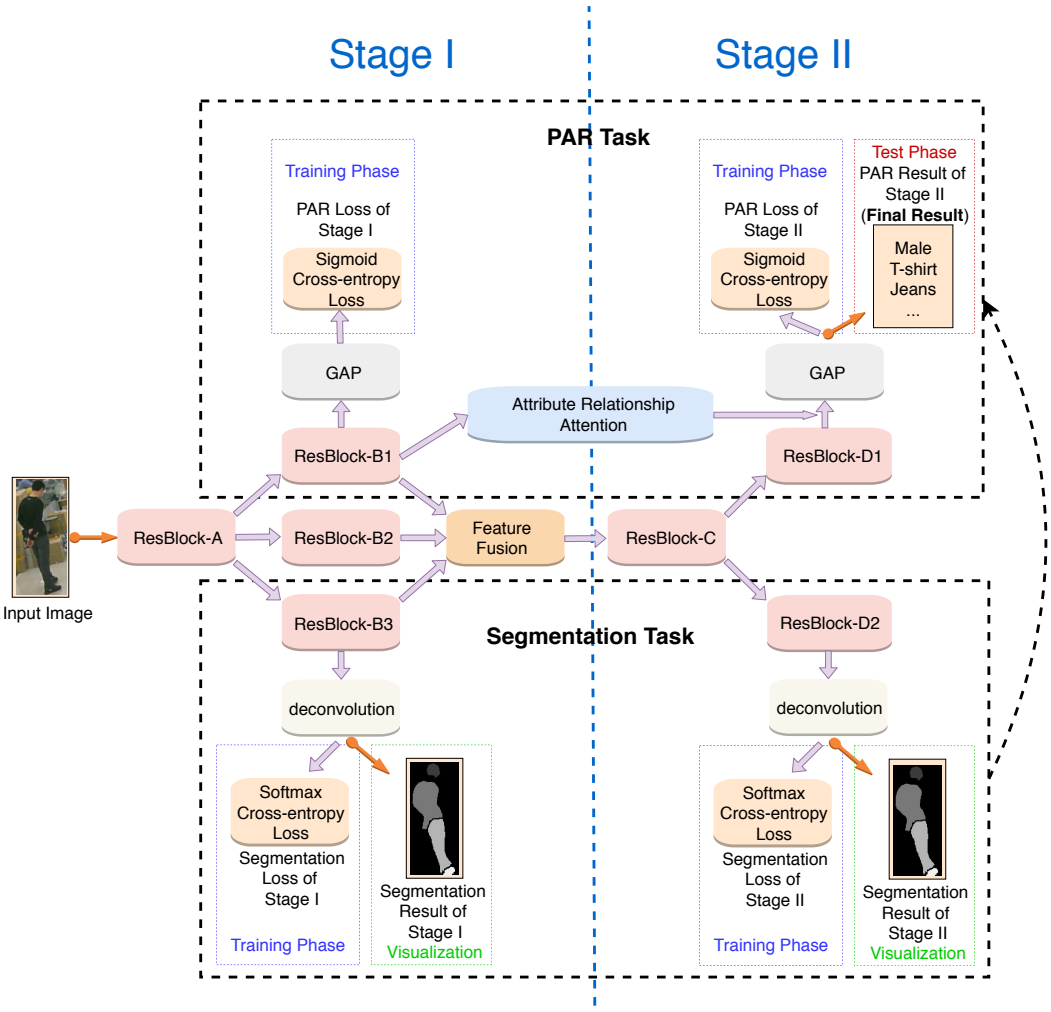

Figure 2: System diagram of the proposed end-to-end hierarchical deep multi-task learning model. It consists of two successive stages. In each stage, there are branches for PAR and human body semantic segmentation. Between the two stages, feature fusion is applied to integrate all the clues and explore the relationships among attributes through an attention module.

3.2 Hierarchical Learning

The multi-task learning network presented in Sec. 3.1 suffers from the same problem as the recent studies do, where the backbone is shared by attribute localization and recognition. Since different attributes have different spatial properties corresponding to different semantics, this structure tends to confuse the model in feature learning. In this case, we update the multi-task learning network to a hierarchical one, where multi-task learning is performed in two successive stages, to decouple feature learning into coarse attribute localization and fine attribute classification. In Stage I, we aim to extract discriminative local features to localize low-level attributes and we thus do not build a very deep architecture. With coarse predictions and local features, we further refine the results in Stage II. Therefore, we consider the two stages to be equally important and we empirically set the depths of the two stages to be approximately equal, as shown in Fig. 2.

In Stage I, the backbone ResBlock-A is utilized to extract local appearance features from the given image patch. Then, the main network is split into three branches. Two of them are specifically trained for segmentation of human parts and recognition of basic attributes, while the other one continues to learn deeper features for the next stage.

In the PAR branch, a Global Average Pooling (GAP) operation is applied to handle the inputs of variable sizes. The sigmoid cross-entropy loss is employed to optimize this branch as

$$J_{1,1}(\theta) = \sum_{i=1}^{I} p^i \cdot \log(\hat{p^i}) + (1 - p^i) \cdot \log(1 - \hat{p^i}), \tag{1}$$

where p^i and $\hat{p^i}$ are the ground-truth and predicted probability of the i-th attribute of a given pedestrian, respectively. Since the PAR problem usually possesses extremely imbalanced numbers of positive samples (e.g. that of certain rare attribute is only 1/20 compared to those of some common ones), the weighted cross-entropy loss, introduced by [26] to handle this, is also exploited:

$$J^*_{1,1}(\theta) = \sum_{i=1}^{I} \frac{1}{2\omega_i} \cdot p^i \cdot \log(\hat{p^i}) + \frac{1}{2(1-\omega_i)} \cdot (1-p^i) \cdot \log(1-\hat{p^i}), \tag{2}$$

where ω_i is the positive sample rate of the i-th attribute. This loss assigns penalties based on classes rather than samples, inducing better performance when the class imbalance problem is severe.

In the human body semantic segmentation branch, a deconvolution layer is employed to upsample the feature map to the same size as the input. Then, our model generates a segmentation result with pixel-level accuracies. A pixel-wise softmax cross-entropy loss is utilized as

$$J_{1,2}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} c_{n,k} \cdot \log(a_{n,k}), \tag{3}$$

where $c_{n,k}$ and $a_{n,k}$ denote the ground-truth and predicted probability of the n-th pixel in the k-th body category, respectively.

Note that Stage I ends in block-3d of the ResNet backbone and Stage II starts with block-4a. The feature maps of the three branches at stage I are further fused by element-wise additions. The local attribute features are combined with the corresponding localization features which contain fine-grained spatial information, and the network at Stage II can thus better predict the global attributes. The fusion process is described as

$$i_{4a} = (o^1_{3d} + o^2_{3d} + o^3_{3d})/3, \tag{4}$$

where $o^i_{3d}(i = 1, 2, 3)$ represents the output of the i-th branch in Stage I and i_{4a} serves as the input of Stage II.

In Stage II, the fused feature is processed by ResBlock-C and split into two branches to perform PAR and semantic segmentation accordingly. The components in these branches and their corresponding loss functions $J_{2,1}(\theta)$ and $J_{2,2}(\theta)$ are similar to Stage I. Then, the final loss function is defined as

$$J(\theta) = \sum_{s=1}^{S} \sum_{t=1}^{T} \lambda_{s,t} \cdot J_{s,t}(\theta), \tag{5}$$

where $\lambda_{s,t}(s, t = 1, 2)$ denotes the factor of learning rate for each loss in each stage. In the training process, the ResNet-50 model, pre-trained on the ImageNet dataset, is employed to initialize the network parameters. Then, the subnetwork in Stage I is optimized by setting $\lambda_{1,1} = 1, \lambda_{1,2} = 1.5e-4, \lambda_{2,1} = 0$ and $\lambda_{2,2} = 0$. At last, we set $\lambda_{1,1} = 1, \lambda_{1,2} = 1.5e-4, \lambda_{2,1} = 1$ and $\lambda_{2,2} = 1.5e-4$ to jointly optimize our end-to-end two-staged hierarchical model with the Adam algorithm [13].

3.3 Attribute Relationship Attention

Intuitively, there exist certain semantically hierarchical correlations among the pedestrian attributes, and local attributes can be utilized to infer global ones. For example, a person, who wears a "skirt", tends to be a "woman" with "long hair". To better exploit such relationships and improve overall performance, we design an unsupervised visual attention based mechanism. Since our end-to-end hierarchical deep multi-task learning model generates multiple PAR results in different stages, the attribute relationships can be emphasized with respect to these results.

As shown in Figure 3, two PAR feature maps are generated by the proposed hierarchical multi-task learning network, and the attribute relationship attention module is constructed between these feature maps, which are obtained before GAP operations in the two stages. Note that the feature map from Stage I is processed by a down-sampling convolution and certain ReLU activation operations. A weight map and a bias map are computed to refine the feature map extracted from Stage II and thus polish the final attribute prediction results. The refining process is formulated as

$$y = b_{att} + w_{att} \cdot x_2, \tag{6}$$

where

$$w_{att} = \varphi(\Theta(x_1, \theta_{att,w})), b_{att} = \Theta(x_1, \theta_{att,b}). \tag{7}$$

where $\Theta(x_1, \theta_{att,w})$ and $\Theta(x_1, \theta_{att,b})$ denote convolutions on the PAR feature map x_1 at stage I, using parameters $\theta_{att,w}$ and $\theta_{att,b}$, respectively. φ is a ReLU activation operation.

The semantic relationships among attributes can be emphasized by applying the weights and bias obtained from the feature maps extracted in Stage I. The sub-network at Stage II then renders more precise prediction with this relationship attention module.

4 EVALUATION

To assess the effectiveness of the proposed approach, we carry out extensive experiments on the RAP and PETA datasets. The datasets, experimental settings, protocols and results, and discussions are subsequently presented.

4.1 Datasets

The RAP (Richly Annotated Pedestrian) dataset [16] is collected by multi-camera surveillance systems for pedestrian analysis. It contains 41,585 samples, each with 72 annotated attributes. To the best of our knowledge, RAP is currently the largest dataset for PAR.

The PETA (PEdesTrian Attribute) dataset [5] is a pioneering and widely employed PAR dataset. It consists of multiple relatively small-scaled person re-identification datasets with totally 19,000 samples annotated with 61 binary attributes and 4 multi-label attributes. The database has multiple interferences, such as illumination, pose and occlusion variations.

4.2 Protocols and Metrics

For the RAP dataset, the pre-defined training protocol [16] is employed, and for the PETA dataset, the protocol in [5] is utilized, to achieve fair comparison with the state of the art.

Although abundant attributes are provided in the two datasets, some of the attributes only appear in very few positive samples (e.g. "beld"). Therefore, in our experiments, we select the attributes whose positive rates are greater or equal to 1/20 as in [5, 16]. Then, 51 attributes in RAP and 35 attributes in PETA are considered.

In our evaluation, the mean accuracy (mAcc) and the sample-based metrics proposed in [16] are used. The mAcc is defined as

$$mAcc = \frac{1}{L} * \sum_{i=1}^{L} (\frac{|TP_i|}{|P_i|} + \frac{|TN_i|}{|N_i|}), \tag{8}$$

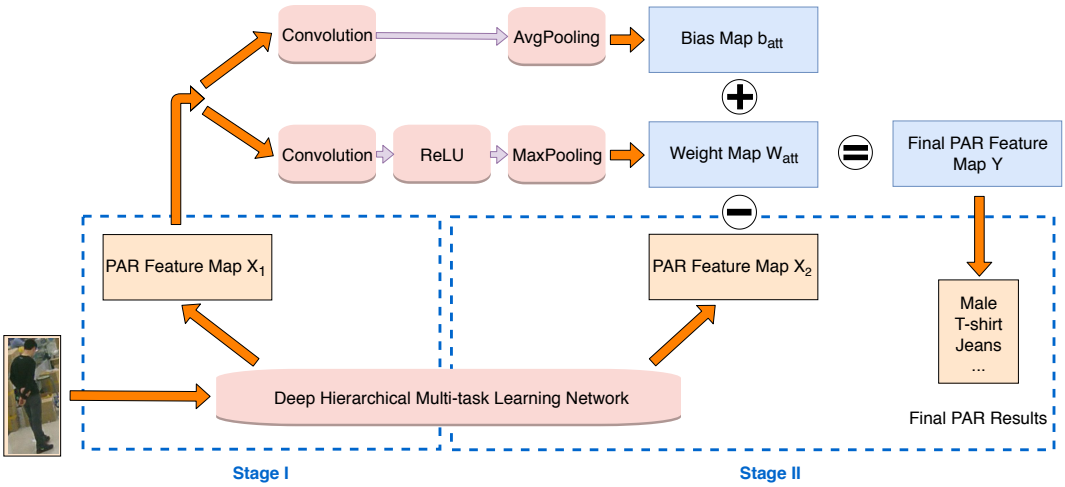

Figure 3: Diagram of the proposed attribute relationship attention module.

Table 1: Ablation study in terms of different metrics using the weighted cross-entropy loss on the RAP and PETA databases.

Method	RAP					PETA				
	mAcc	***Acc***	***Prec***	***Rec***	***F1***	***mAcc***	***Acc***	***Prec***	***Rec***	***F1***
Baseline	81.42	49.03	54.79	79.24	64.78	85.81	76.54	84.23	87.14	85.66
Multi-task	81.38	**49.53**	54.82	78.94	64.71	85.13	76.83	**84.79**	86.82	85.79
MTMS (S1)	75.74	37.00	41.29	77.08	53.77	78.24	69.72	79.81	83.52	81.62
MTMS (S2)	81.91	48.90	54.82	79.96	65.05	**86.32**	77.19	84.61	87.15	**85.86**
MTMS+Att	**82.45**	49.10	**55.00**	**80.44**	**65.33**	86.23	**77.21**	84.52	**87.22**	85.85

Figure 4: Examples of human body semantic segmentation ground-truths on the RAP and PETA datasets.

where L represents the number of attributes; $|TP_i|$ and $|TN_i|$ are the numbers of correctly predicted positive and negative samples; and $|P_i|$ and $|N_i|$ denote the numbers of positive and negative samples. The sample-based metrics are defined as

$$Acc_{sb} = \frac{1}{N} * \sum_{n=1}^{N} \left(\frac{|Y_i \bigcap f(x_i)|}{|Y_i \bigcup f(x_i)|}\right), \tag{9}$$

$$Prec_{sb} = \frac{1}{N} * \sum_{n=1}^{N} \left(\frac{|Y_i \bigcap f(x_i)|}{|f(x_i)|}\right), \tag{10}$$

$$Rec_{sb} = \frac{1}{N} * \sum_{n=1}^{N} \left(\frac{|Y_i \bigcap f(x_i)|}{|Y_i|}\right), \tag{11}$$

$$F1_{sb} = \frac{2 * Prec_sb * Rec_sb}{Prec_sb + Rec_sb}, \tag{12}$$

where N is the total number of samples, Y_i and $f(x_i)$ are the numbers of ground-truth and predicted positive samples, respectively.

4.3 Implementation Details

The proposed model is implemented using the Caffe library [11]. It is trained and tested on a server with a single Nvidia 1080Ti GPU. In the training stage, the batch size is set to 8. The initial learning rate of the proposed model is 3e-4, and it decreases 20% after 2000 epochs. For the ResNet-50 baseline model and the proposed model without the attribute relationship attention module, the initial learning rate is set to 1e-3 and 1e-4, respectively.

For the human body semantic segmentation task, as the RAP and PETA datasets do not provide any pixel-wise segmentation ground-truth labels, a self-trained ResNet-101 model is employed to generate the ground-truths. This model is trained on the Look-Into-Person dataset [8]. Some examples from RAP and PETA as well as their segmentation ground-truths are shown in Figure 4.

4.4 Ablation Study

To evaluate the effectiveness of different components in the proposed approach, an ablation study is performed on both the RAP and PETA datasets. The results are given in Table 1. The baseline results are obtained by using the original ResNet-50 model. The ones of multi-task learning are achieved by the model which integrates PAR and semantic segmentation. Although the scores can be further optimized, we set the depth of the shared layer between the two tasks as that of Stage I (i.e. ResBlock-A) so that fair comparison

Table 2: Comparison with the State-of-the-art methods in the RAP and PETA databases.

Method	*Loss*	RAP					PETA				
		mAcc	***Acc***	***Prec***	***Rec***	***F1***	***mAcc***	***Acc***	***Prec***	***Rec***	***F1***
ELF-mm	original	69.94	29.29	32.84	71.18	44.95	75.21	43.68	49.45	74.24	59.36
FC7-mm	original	72.28	31.72	35.75	71.78	47.73	76.65	45.41	51.33	75.14	61.00
FC6-mm	original	73.32	33.37	37.57	73.23	49.66	77.96	48.13	54.06	76.49	63.35
ACN [23]	original	69.66	62.61	80.12	72.26	75.98	81.15	73.66	84.06	81.26	82.64
DeepMAR [14]	original	73.79	62.02	74.92	76.21	75.56	82.89	75.07	83.68	83.14	83.41
HP-Net [20]	weighted	76.12	65.39	77.33	78.79	78.05	81.77	76.13	84.92	83.24	84.07
WPAL [26]	weighted	81.25	50.30	57.17	78.39	66.12	85.50	76.98	84.07	85.78	84.90
PGDM [15]	original	74.31	64.57	78.86	75.90	77.35	82.97	78.08	**86.86**	84.68	85.76
GAPAR [7]	weighted	79.73	-	76.96	78.72	77.83	-	-	-	-	-
JRL [24]	original	77.81	-	78.11	78.98	78.58	85.67	-	86.03	85.34	85.42
GRL [28]	original	81.20	-	77.70	80.90	79.29	**86.70**	-	84.34	**88.82**	**86.51**
VAA [21]	focal	-	-	-	-	-	84.59	78.56	86.79	86.12	86.46
Ours	original	74.43	**67.09**	**83.13**	76.00	**79.41**	85.72	**78.62**	85.12	87.32	86.21
Ours	weighted	**82.45**	49.10	55.00	**80.44**	65.33	86.23	77.21	84.52	87.22	85.85

is made to the proposed model. MTMS refers to the proposed hierarchical multi-task learning network (S1 and S2 represents Stage I and II, respectively) and MTMS+Att adds the proposed attribute relationship attention module to the main network. All the results are produced with the weighted sigmoid cross-entropy loss.

As can be observed from Table 1, the multi-task model achieves better performance compared to the baseline ResNet-50 based one, indicates the necessity of the semantic segmentation branch. The proposed main network with the same depth (MTMS S2) further improves the precision, which supports our claim that the two-staged pipeline ameliorates the PAR performance. The end-to-end hierarchical deep multi-task learning model with attribute relationship attention module delivers the best scores (except Acc.) on the RAP benchmark, which validate our contributions. Regarding the results on PETA, although the general trend is similar as in RAP, there exist very slight decreases in mean accuracy, precision, and F1 score. The reason lies in that PETA contains less data and larger variety of the attribute combinations, such as "longHair" and "male".

Some visualized results of both the semantic segmentation and PAR tasks from the RAP dataset are shown in Figure 5. For human body semantic segmentation, the ground-truth and predicted segmentation label maps at Stage I and II are displayed in the grayscale form. For the PAR task, the activation maps before GAP are shown in grayscale maps, where a darker color represents that this pixel is less likely to be correlated to the corresponding attribute, and vice versa. When comparing the segmentation results from Stage I and Stage II, we can see that Stage I can generate very good human parts, similar to the ground-truth. It confirms that it makes sense to model coarse attribute localization with a shallower network to facilitate following feature learning. The results also show that the proposed model can successfully locate certain local attributes, such as"long hair", "sports shoes", etc. Note that some attributes may possess strong correlations with other attributes, such as "long hair" and "female", and our method can successfully capture these correlations. On the other hand, some attributes tend to give wider activation areas than our estimation, such as "long hair", because the global semantics appear to affect these attributes and our attribute relationship attention module may give a further boost to this phenomenon. In general, the proposed model can accurately segment the human body and predict the pedestrian attributes simultaneously via a fusion of all the extracted features.

4.5 Comparison to the State-of-the-arts

In this section, the proposed network is compared to the baseline model provide by [16] (ELF-mm, FC7-mm and FC6-mm) as well as 9 state-of-the-art approaches for the PAR task. The compared state-of-the-art techniques include ACN [23], DeepMAR [14], HP-Net [20], WPAL [26], PGDM [15], GAPAR [7], JRL [24], GRL [28] and VAA [21]. For convenience, the results of our method with two different losses are provided. The details are shown in Table 2.

The experimental results on the RAP dataset in Table 2 indicate that our method with the sigmoid cross-entropy loss achieves the state-of-the-art performance in terms of accuracy, precision and F1 score. Meanwhile, our method with the weighted sigmoid cross-entropy loss outperforms the other methods in terms of the mAcc and recall metrics. For the PETA dataset, our networks do not perform as well as those in RAP. They only show the best Acc. score, and the ones of other metrics are slightly inferior but still comparable to the state-of-the-arts. The reason is that segmentation ground-truths in PETA are not as stable as the ones in RAP due to the relatively low image quality.

4.6 Discussion

In the experiments, we employ two different sigmoid cross-entropy losses. The results suggest that the weighted loss performs better in addressing the class imbalance issue. Compared to the standard cross-entropy loss, it assigns penalties to the minority of samples with larger weights and it is thus beneficial to the metrics which emphasize the importance of those samples, such as mAcc and sample-based recall rate. Meanwhile, as the weighted loss assigns smaller weights to the majority of samples, it does not deliver very good scores on the metrics that assess the samples from different classes based on their own distributions.

Due to the multi-task architecture, the computation cost of our model is higher compared to the single-task version. Specifically,

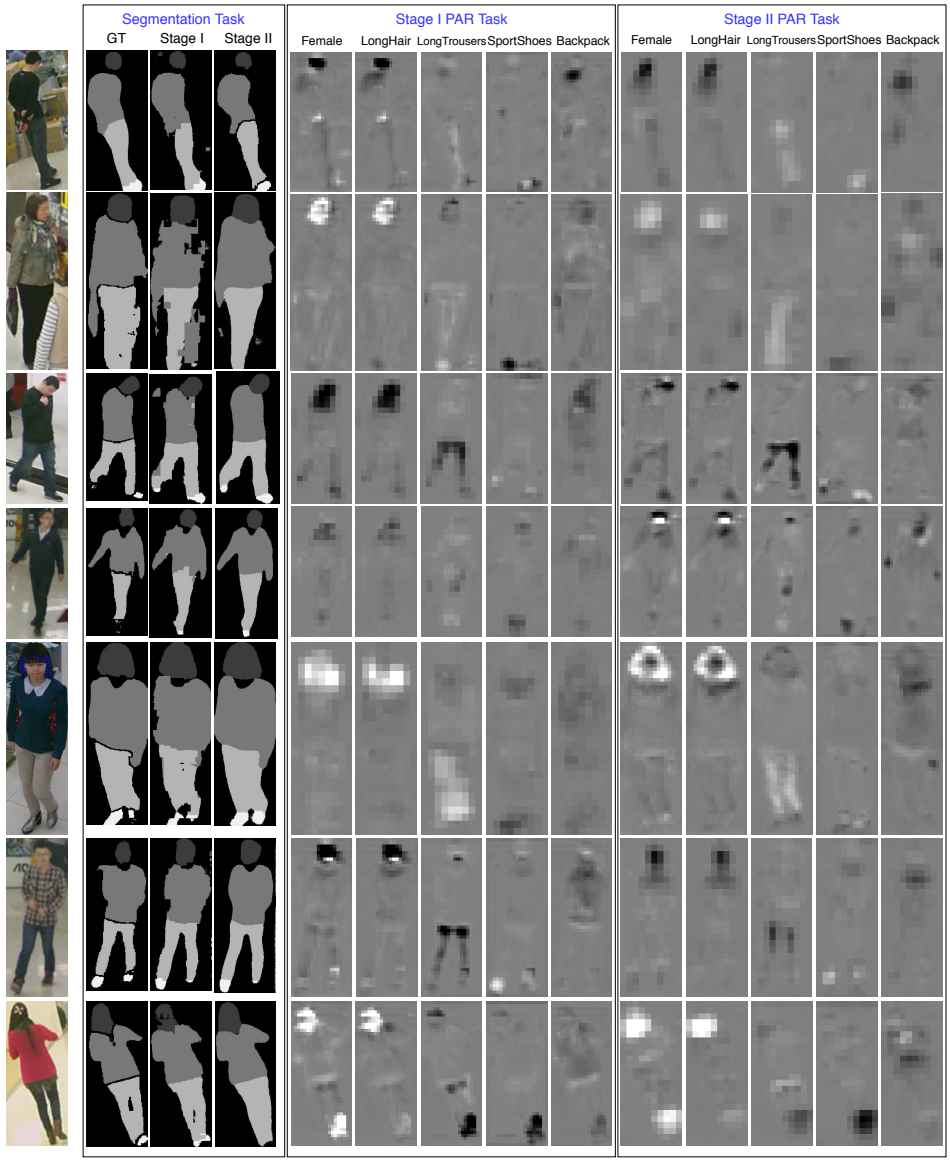

Figure 5: Examples of the human body semantic segmentation and PAR results on the RAP dataset.

the sub-network with the two additional branches (for PAR and Segmentation) in Stage I induces approximately tripled computations while the one with the additional branch in Stage II consumes doubled computations. Besides, the attention module incurs some additional costs as well. But such costs do not significantly increase the running time, because the multi-task architecture handles the two tasks simultaneously. In average, given a pedestrian image patch, with a single 1080Ti GPU, the inference phase spends 17.55ms for the multi-task model and 13.58ms for the single-task one.

5 CONCLUSION

In this paper, an end-to-end hierarchical deep multi-task learning network is proposed to address the PAR problem. It effectively exploits the fine-grained spatial information of the attributes and the abundant relationships among them. The proposed network executes in a two-staged manner to decouple coarse attribute localization and fine attribute recognition into successive phases within a single model, which enhances feature learning. Besides, we propose a novel attribute relationship attention module by adopting the deep visual attention mechanism to better extract and utilize the relationships among the attributes to refine the predictions. Experiments are conducted on the RAP and PETA databases, and the results achieved are competitive, indicating the effectiveness and superiority of the proposed approach.

ACKNOWLEDGMENTS

This work was partly supported by the National Key Research and Development Plan (Grant No.2016YFC0801002) and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] H. Bilen and A. Vedaldi. 2016. Integrated perception with recurrent multi-task neural networks. In *Proc. Advances in Neural Inf. Process. Systems*. 235–243.

[2] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. 2008. Gender recognition from body. In *Procs. ACM Multimedia*. 725–728.

[3] H. Chen, A. Gallagher, and B. Girod. 2012. Describing clothing by semantic attributes. In *Proc. European Conf. Computer Vis.* 609–623.

[4] L. C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. 2016. Attention to scale: Scale-aware semantic image segmentation. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 3640–3649.

[5] Y. Deng, P. Luo, C. C. Loy, and X. Tang. 2014. Pedestrian attribute recognition at far distance. In *Procs. ACM Multimedia*. 789–792.

[6] Y. Deng, P. Luo, C. C. Loy, and X. Tang. 2015. Learning to recognize pedestrian attribute. *arXiv e-prints* (2015).

[7] M. Fabbri, S. Calderara, and R. Cucchiara. 2017. Generative adversarial models for people attribute recognition in surveillance. In *Procs. IEEE Int. Conf. Advanced Video and Signal Based Surveillance*. 1–6.

[8] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 932–940.

[9] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 770–778.

[10] S. Ioffe and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint* arXiv:1502.03167 (2015).

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Procs. ACM Multimedia*. 675–678.

[12] J. Joo, S. Wang, and S. C. Zhu. 2013. Human attribute recognition by rich appearance dictionary. In *Proc. IEEE Int. Conf. Computer Vis.* 721–728.

[13] D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980 (2014).

[14] D. Li, X. Chen, and K. Huang. 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proc. Asian Conf. Pattern Recognit.* 111–115.

[15] D. Li, X. Chen, Z. Zhang, and K. Huang. 2018. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *Procs. IEEE Int. Conf. Multimedia Expo*. 1–6.

[16] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. 2016. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint* arXiv:1603.07054 (2016).

[17] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, and X. Wang. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proc. IEEE Int. Conf. Computer Vis.* 350–359.

[18] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 3994–4003.

[19] V. Mnih, N. Heess, and A. Graves. 2014. Recurrent models of visual attention. In *Proc. Advances in Neural Inf. Process. Systems*. 2204–2212.

[20] R. Ranjan, V. Patel, and R. Chellappa. 2019. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 41, 1 (2019), 121–135.

[21] N. Sarafianos, X. Xu, and I.A. Kakadiari. 2018. Deep imbalanced attribute classification using visual attention aggregation. In *Procs. European Conf. Computer Vis.* 680–697.

[22] Z. Shi, T. M. Hospedales, and T. Xiang. 2015. Transferring a semantic representation for person re-identification and search. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 4184–4193.

[23] P. Sudowe, H. Spitzer, and B. Leibe. 2015. Person attribute recognition with a jointly-trained holistic cnn model. In *Proc. IEEE Int. Conf. Computer Vis. Workshops*. 87–95.

[24] J. Wang, X. Zhu, S. Gong, and W. Li. 2017. Attribute recognition by joint recurrent learning of context and correlation. In *Proc. IEEE Int. Conf. Computer Vis.* 531–540.

[25] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 842–850.

[26] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang. 2016. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint* arXiv:1611.05603 (2016).

[27] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. 2014. Panda: Pose aligned networks for deep attribute modeling. In *Proc. IEEE Int. Conf. Computer Vis. Pattern Recognit.* 1637–1644.

[28] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin. 2018. Grouping attribute recognition for pedestrian with joint recurrent learning. In *Procs. Int. Joint Conf. Artificial Intelligence*. 3177–3183.

[29] J. Zhu, S. Liao, Z. Lei, and S. Z. Li. 2017. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing* 58 (2017), 224–229.