Attention-Based Pedestrian Attribute Analysis

Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, *Senior Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*— Recognizing the pedestrian attributes in surveillance scenes is an inherently challenging task, especially for the pedestrian images with large pose variations, complex backgrounds, and various camera viewing angles. To select important and discriminative regions or pixels against the variations, three attention mechanisms are proposed, including parsing attention, label attention, and spatial attention. Those attentions aim at accessing effective information by considering problems from different perspectives. To be specific, the parsing attention extracts discriminative features by learning not only where to turn attention to but also how to aggregate features from different semantic regions of human bodies, e.g., head and upper body. The label attention aims at targetedly collecting the discriminative features for each attribute. Different from the parsing and label attention mechanisms, the spatial attention considers the problem from a global perspective, aiming at selecting several important and discriminative image regions or pixels for all attributes. Then, we propose a joint learning framework formulated in a multi-task-like way with these three attention mechanisms learned concurrently to extract complementary and correlated features. This joint learning framework is named Joint Learning of Parsing attention, Label attention, and Spatial attention for Pedestrian Attributes Analysis (JLPLS-PAA, for short). Extensive comparative evaluations conducted on multiple large-scale benchmarks, including PA-100K, RAP, PETA, Market-1501, and Duke attribute datasets, further demonstrate the effectiveness of the proposed JLPLS-PAA framework for pedestrian attribute analysis.**

***Index Terms*— Pedestrian attribute analysis, attention mechanism, pedestrian parsing.**

I. Introduction

VISUAL analysis of pedestrian attributes such as gender, age, body shape, etc., has become a thriving research field in recent years [1]–[8], on account of its wide range of possible applications, such as person retrieval [9], [10], person re-identification [5], [11], video-based business intelligence [12] and so on. In the area of pedestrian attributes analysis, the related technologies have obtained significant improvement in performance owing to the success of deep learning [13]–[16]. However, hampered by arbitrary human poses, different camera viewing angles, occlusions and background clutter, the automatical pedestrian attribute recognition remains a challenging problem.

How human beings are able to recognize objects effortlessly and efficiently has aroused a great interest. It has been proposed in the literature that these recognizing processes are conducted by exploiting attention mechanisms to access the effective information [17], [18]. To be specific, the attention mechanisms contain many different natural recognition behaviors of human beings. For example, when looking at and analyzing a person, our humans would like to focus on the foreground while ignoring the background. Moreover, we would turn our attention to a specific region when analyzing one particular attribute of a person. A conceivable example is that we tend to focus on one's upper-body region when recognizing his/her clothes type. With the recent rapid development of deep learning, these attention mechanisms originated from the human visual system have gained their popularity for their capability to help the network to focus on the most discriminative features to solve challenging recognition problems, and therefore been extensively studied [19]–[27].

In this paper, we formulate a new framework for pedestrian attributes analysis based on the attention mechanisms. Different from the previous works [22]–[24], [28] that formulate the attention for image-based recognition mainly from a spatial-view, temporal-view and/or channels-view, our attention mechanisms for pedestrian attribute analysis are formulated from a more rich point of view. Specifically speaking, we propose three attention mechanisms for pedestrian attributes analysis including parsing attention, label attention and spatial attention. These three components are then incorporated into a unified network to extract discriminative and correlated

Manuscript received July 10, 2018; revised February 17, 2019; accepted May 4, 2019. Date of publication July 3, 2019; date of current version September 4, 2019. This work was supported in part by the National Key Research and Development Plan under Grant 2016YFC0801002, in part by the Chinese National Natural Science Foundation under Project 61876179, Project 61872367, and Project 61806203, in part by the Science and Technology Development Fund of Macau under Grant 152/2017/A, Grant 0025/2018/A1, and Grant 008/2019/A1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wen Gao. *(Zichang Tan and Yang Yang contribute equally to this work.) (Corresponding author: Jun Wan.)*

Z. Tan is with the Center for Biometrics and Security Research, Chinese Academy of Sciences, Beijing 100190, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zichang.tan@nlpr.ia.ac.cn).

Y. Yang and J. Wan are with the Center for Biometrics and Security Research, Chinese Academy of Sciences, Beijing 100190, China, and also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yang.yang@nlpr.ia.ac.cn; jun.wan@nlpr.ia.ac.cn).

H. Hang is with the Institute of Statistics and Big Data (ISBD), Renmin University of China, Beijing 100872, China.

G. Guo is with the Institute of Deep Learning, Baidu Research, Beijing 100193, China, and also with the National Engineering Laboratory for Deep Learning Technology and Application, Beijing 100193, China (e-mail: guoguodong01@baidu.com).

S. Z. Li is with the Center for Biometrics and Security Research, Chinese Academy of Sciences, Beijing 100190, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China (e-mail: szli@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2019.2919199

complementary information for achieving a more reliable attribute recognition.

The parsing attention in this study is formulated with the assistance of pedestrian parsing. Pedestrian parsing [29], [30] aiming at parsing a pedestrian into different semantic regions, e.g., hair, face and upper body, can provide the location cues to indicate the specific body regions at pixel-level. Therefore, we take full advantage of pedestrian parsing to construct our parsing attention mechanism in a split-and-aggregate way. More specifically, the features from different semantic human regions are first split under the guidance of pixel-level location cues generated by the pedestrian parsing, which leads to more detail and location-oriented features. Then, the attention module learns how to aggregate these split features. Since the features are split and so refined in advance, this mechanism helps us to learn more attention-oriented and therefore discriminative features even when there are large variations in the pedestrian images.

The label attention mechanism is another novel mechanism of conquering the difficulty CNN faces when dealing with multi-attributes analysis. When employing a CNN to extract the features for dozens of attributes analysis, the feature representations may hardly be optimal for all attributes at the same time. This problem comes from the fact that different attributes are often related with different human body regions. For example, we mainly look at a pedestrian's upper body region to judge his/her clothing style while look at the feet to judge his/her shoes color. As a consequence, the performance of features obtained from the overall area taking all attributes into consideration may be less optimal than that of the features obtained from each of the attribute corresponding areas. To overcome this problem, we formulate a label attention mechanism by assigning several attention maps for each label, where the features from the most relevant regions are enhanced while the irrelevant features are suppressed for each attribute. In this attention mechanism, the attention map is learned only under image-level supervisions, which is different from the parsing attention supervised by pixel-level labels and image-level labels.

A spatial attention module is also incorporated into the network which learns to localize the most discriminative image regions for all attributes with only image-level supervisions. We should notify that this attention module is different from the previous two kinds of attentions. Unlike parsing attention which is formed based on the pixel-level indicators generated using an external network, the spatial attention discovers the discriminative regions through self-learning and image-level supervisions. Moreover, different from the label attention that localizes the relevant regions for each attribute, the spatial attention is designed for all classification tasks. It can then be apparently seen that these three attention mechanisms capture the discriminative features from different perspectives and therefore are complementary correlated.

In order to synergize the three different mechanisms more effectively, we formulate a three-branch CNN architecture for pedestrian attributes analysis with each branch corresponding to one attention mechanism. In the literature, a mainstream architecture which integrates different attention mechanisms is to take a series structure [23], [27]. However, the prediction performances of this structure are not always that satisfying, let alone its difficulties in training. One possible explanation for the inaccuracy is that there may be cases where the information suppressed in a certain attention mechanism is actually useful in the next attention mechanism. In other words, the series structure may suppress some information in advance before it can play a role. Nevertheless, our CNN architecture is constructed in a multi-task-like or parallel way where the three branches are synergistically correlated and jointly learned instead of being independent. Under this parallel structure, correlated complementary information can be extracted from different views so as to achieve more reliable predictions.

The main contributions of our work can be summarized as follows: (1) We put forward the novel idea of learning different attention mechanisms concurrently in a multi-task-like way to explore the correlated complementary information. To the best of our knowledge, this is the first attempt to jointly learn multiple attention mechanisms in a multi-task-like learning manner. (2) We propose three kinds of attention mechanisms for pedestrian attribute analysis which are parsing attention, label attention and spatial attention. Parsing attention is formulated under the guidance of pixel-level location cues while label attention and spatial attention aim at selecting the important and discriminative regions for each attribute and all attributes, respectively. The three attention mechanisms considering problems from different perspectives are correlated and complementary. Moreover, as far as we know, these attention mechanisms have not been scrutinized in the literature. (3) We also annotate a new pedestrian parsing dataset with abundant annotations at pixel-level for achieving better parsing performance. This dataset will be released to the community. Extensive comparative evaluations demonstrate the superiority of the proposed method over several benchmark datasets including PA-100K, RAP, PETA, Market-1501 and Duke attribute datasets.

II. Related Work

1) Pedestrian Attribute Recognition: Recent years have seen a substantial application potential of pedestrian attribute analysis in video surveillance system, which also promotes the field to become a hot research topic. Earlier methods shedding light on pedestrian attributes recognition typically model each attribute independently based on hand-crafted features such as Gabor [11], [32] and color [33] with SVM or AdaBoost classifiers. The rapid development of deep learning [14]–[16], [34], [35] lately leads to a great successes in pedestrian attribute analysis and approaches [2]–[4], [6], [36]–[38] modeling a multi-task network to analyze all attributes have been extensively investigated. For example, Zhu *et al.* [2] propose a multi-label CNN to predict multiple attributes together in a unified framework. Li *et al.* [36] propose a deep multi-attribute recognition (DeepMAR) method for pedestrian attributes recognition. They also verify that modeling all attributes together in a single network can explore the correlation between different attributes better and capture more complementary features. Wang *et al.* [3] formulate an end-to-end encoder/decoder recurrent network for pedestrian attribute analysis, which aims

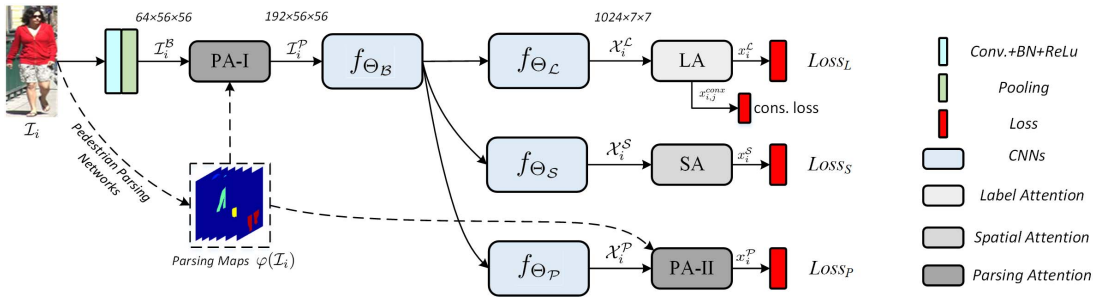


Fig. 1. The overview of the proposed network architecture. The network consists of three branches, where each branch is incorporated with a specific attention mechanism, including parsing attention (PA), label attention (LA) and spatial attention (SA). The network is constructed based on the SE-BN-Inception [24], which is a light CNN architecture in the SE-Net family [24]. More detailedly, the CNNs module $f_{\Theta_{\mathcal{B}}}$ consists of nine inception blocks [31] and nine SE blocks [24] with each inception block followed by a SE block. The module $f_{\Theta_{\mathcal{L}}}$, $f_{\Theta_{\mathcal{S}}}$, $f_{\Theta_{\mathcal{P}}}$ have the same structure, where a inception block followed by a SE block are included in each module. To the end, all three branches are jointly learned concurrently with each branch followed by a loss layer.

to explore attribute context and correlation. Lin *et al.* [5] propose a discriminative CNN embedding for both person re-identification and attributes recognition, yielding promising performance in both tasks. Moreover, Liu *et al.* [6] propose a multi-directional attention mechanism for fine-grained pedestrian analysis. In this study, we establish another attention based method which is different from Liu's work [6]. Furthermore, three different concurrently learned attention mechanisms are proposed to consider the prediction problem from different perspectives.

2) Pedestrian Parsing: Methods [29], [39]–[41] proposed for pedestrian parsing in the early stage rely heavily on training set and lack the ability of accurately fitting object boundaries. However, the present Fully Convolutional Networks (FCNs) which was a category of network architectures has shown its effectiveness and efficiency for segmentation tasks [42]–[47]. When it comes to the specific networks in this category, Long *et al.* [42] first propose a Fully Convolutional Networks (FCNs) for pixel-wise prediction which is originally used in sematic segmentation, and improve the state-of-the-art performance by a big margin at that time. More recently, Zhao *et al.* [45] propose another network in this category named as the Pyramid Scene Parsing Network (PSPNet) for scene parsing which ranks the 1st place in ImageNet Scene Parsing Challenge 2016. Considering how FCNs category achieves a great performance in the segmentation tasks, we are encouraged to construct our pedestrian parsing network based on FCNs. Nevertheless, we are aware of the fact that the low resolution of pedestrian images in surveillance scenes will have a negative influence on the performance of the FCNs. For example, Xia *et al.* [30] employ FCN-32 and FCN-16 [42] for pedestrian parsing while get low performance. As a result, it is inappropriate to directly apply those existing frameworks to pedestrian parsing and some adjustments should be appended to those frameworks.

3) Attention: Attention models [19]–[27], [48] have aroused great enthusiasm in recent years. In the literature, a recurrent attention convolutional neural network architecture to detect the discriminative regions for fine-grained image recognition is proposed by Fu *et al.* [21]. Wang *et al.* [22] propose a residual attention network that is constructed by stacking multiple attention modules. Moreover, a various of experiments are also conducted in their work to show the effectiveness of the proposed network. Li *et al.* [23] propose an attention mechanism that consists of spatial and temporal attention for person re-identification. In more recent work, Hu *et al.* [24] propose Squeeze-and-Excitation Networks (SE-Net) for image classification, where a channels attention mechanism is proposed to recalibrate channel-wise feature responses. With its superior performance, the SE-Net won first place in ILSVRC 2017. Shen *et al.* [48] propose sharp attention networks for person re-identification, and achieve promising performance. Inspired by these works, we establish a new attention network which is expected to achieve better performances. Specially, one of our innovations lies in the adoption of three different kinds of attention mechanisms with two of which are newly developed. The three attention mechanisms are not only incorporated into a unified network and learned jointly, but also, extract the most discriminative features from different views and reach a mutual complementary to obtain better prediction performances.

III. Our Approach

A. The Overall Design

The overview of the proposed network architecture for pedestrian attributes analysis is shown in Fig. 1. The proposed network architecture is constructed based on the SE-BN-Inception [24], which is a light CNN architecture in the SE-Net family. As is presented in the overview, the proposed network architecture utilizes a parallel structure where each of the three branches is incorporated with a specific attention mechanism from parsing attention, label attention and spatial attention. Since different attention mechanisms have different perspectives, they are expected to capture the correlated complementary information and discover optimal per-branch discriminative feature representations. To this end, we formulate a joint learning scheme with the following principles: (1) *low-level features are shared for all branches*. It can be seen from the Fig. 1 that all three branches receive common low-level features before performing the respective CNNs. This shared learning inspired by multi-task leaning [49], [50] can facilitate not only the inter-attention common learning, but also the knowledge transfer between different attentions. Nevertheless, it also helps to reduce the

number of parameters and the risk of overfitting. (2) *all branches are learned concurrently and seperately with the shared label supervisions.* All branches are learned seperately, which means that after receiving the shared label supervisions, the backpropagation training processes of the three branches are parallel. These separate learning processes agree with the nature that perspective of problem analysis varies from one attention mechanism to another, which helps each branch to extract the discriminative features. On the other hand, since different branches are actually synergistically correlated for their sharing of common low-level features, learning concurrently facilitates the network to discover the correlated complementary features. Note that each branch is assigned with a separate objective loss function but all of them, at the same time, share the same label supervisions.

B. End-to-End Network Framework

We assume that a training set with m samples is denoted as $\mathcal{D} = \{\mathcal{I}_i, \{y_i^j\}_{j=0}^{\mathcal{T}-1}\}$, $i = 0, \ldots, m$, where y_i^j indicates the label for j^{th} task of the image $\mathcal{I}_i$ and $\mathcal{T}$ denotes the number of the tasks needed to be analyzed. In the proposed network, a convolutional layer, a BatchNorm layer, a scale layer and a pooling layer are first employed to extract the low-level features $\mathcal{I}_i^{\mathcal{B}}$. Mathematically, this process is represented as:

$$\mathcal{I}_i^{\mathcal{B}} = Conv\left(\mathcal{I}_i\right). \tag{1}$$

After that, the extracted low-level features $\mathcal{I}_i^{\mathcal{B}}$ would be input into the first parsing attention module (PA-I) to obtain the features $\mathcal{I}_i^{\mathcal{P}}$. A parsing attention in low-level layers is implemented here for the reason that the split-and-aggregate structure resided in this module will provide more discriminative and thus effective features for the following network. To be specific, since the low-level features $\mathcal{I}_i^{\mathcal{B}}$ are in large size, e.g. 56×56, it is available to clearly split the features from different semantic human regions to obtain some more detailed and location-oriented features without loss of information before aggregating. This splitting process serves as a pre-refined procedure which helps to provide some more attention-oriented features even when there are large variations in the pedestrian images.

As is shown in Fig. 1, Three more non-linear mappings are further employed to extract high-level features $\mathcal{X}_i^{\mathcal{L}}$, $\mathcal{X}_i^{\mathcal{S}}$ and $\mathcal{X}_i^{\mathcal{P}}$ for label attention module, spatial attention module and the second parsing attention module (PA-II), respectively. The above non-linear mappings are implemented by different CNNs: the bottom CNNs module $f_{\Theta_{\mathcal{B}}}$ shared for all attention modules with parameters $\Theta_{\mathcal{B}}$ and other three CNNs modules $f_{\Theta_{\mathcal{L}}}$, $f_{\Theta_{\mathcal{S}}}$ and $f_{\Theta_{\mathcal{P}}}$ specifically designed for each attention module with parameters $\Theta_{\mathcal{L}}$, $\Theta_{\mathcal{S}}$ and $\Theta_{\mathcal{P}}$. Note that the last three CNNs modules have the same structure but with various parameter values. The above non-linear mappings can be represented as:

$$\mathcal{X}_i^{\kappa} = f\left(\mathcal{I}_i^{\mathcal{P}}; \Theta_{\mathcal{B}}, \Theta_{\kappa}\right) \quad \kappa \in \{\mathcal{L}, \mathcal{S}, \mathcal{P}\} \tag{2}$$

Finally, the high-level features $\mathcal{X}_i^{\mathcal{L}}$, $\mathcal{X}_i^{\mathcal{S}}$ and $\mathcal{X}_i^{\mathcal{P}}$ are imported into the corresponding attention modules to extract the discriminative features for pedestrian attributes recognition. The different attention mechanisms and the objective loss functions for different branches will be introduced in following subsections.

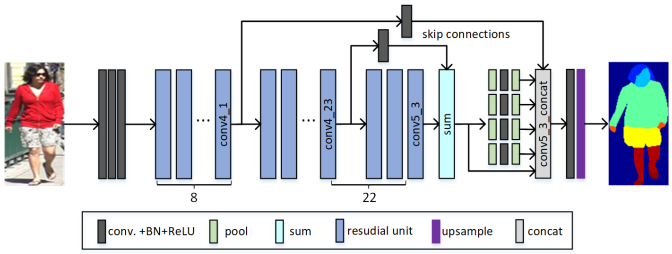


Fig. 2. The network architecture of the pedestrian parsing network.

C. Parsing Attention Mechanism

The parsing attention mechanism aims at learning the discriminative features under the pixel-level supervisions of human body regions, which are generated by a pedestrian parsing network. We give a lucid illustration in this section by first introducing the pedestrian parsing network employed to generate the pedestrian parsing map and then presenting the structure of the proposed parsing attention module.

1) Pedestrian Parsing Network: Our pedestrian parsing network is constructed based on PSPNet [45], which is a popular parsing framework and ranks the 1st place in ImageNet Scene Parsing Challenge 2016. The overall architecture of the proposed Pedestrian Parsing Network (PPN) can be found in Fig. 2, and more detailed information about the architecture can refer to the project.[1] Due to the low resolution of pedestrian images captured in surveillance scenes, however, there is a need for us to make the following proper changes to the existing pedestrian parsing network.

- **Subsampling.** The pedestrian images captured in surveillance scenes are usually in low resolution, e.g., 128×48 in VIPeR dataset, while images in PSPNet [45] usually have a high resolution. Consequently, huge information loss will occur if we simply use the subsampling rate of $\frac{1}{8}$ as in PSPNet. In order to avoid this huge loss, we adopt the subsampling rate of $\frac{1}{2}$ twice ($\frac{1}{4}$ in total) for each image in our network. More specially, the stride of *conv3_1*_3×3 and *conv3_1*_1×1*_proj* is changed from 2 to 1.
- **Skip connections.** Skip connections can merge both low-level and high-level features together to generate more abundant features. There are two skip connections in our pedestrian parsing network as shown in Fig. 2. One skip connects *conv4_1* and *conv5_3_concat* with a convolutional and a BatchNorm layer. The other skip used to connect *conv4_23* and *conv5_3* is also equipped with a convolutional and a BatchNorm layer. The reason why we adopt the skip connection is that it not only helps the information to exchange between different layers, but also is propitious to collect more abundant context information to generate a more reliable parsing map.

The pedestrian parsing network is utilized to generate the parsing maps of pedestrian images for parsing attention

[1] https://github.com/hszhao/PSPNet

modules. For the sake of clarity, we use $\varphi(\mathcal{I}_i)$ to denote the generated parsing map for the image $\mathcal{I}_i$. It is noteworthy that the parsing map $\varphi(\mathcal{I}_i)$ can also be regarded as a probability map which consists of the probabilities of assigning each pixel to different semantic pedestrian regions. Here, we assume that the parsing map contains S pixels and all images are needed to be parsed into C regions. For example, S equals to 861 when the paring map has the size of 41×21, and C equals to 9 if the semantic regions include hair, face, u-clothes (upper clothes), arms, l-cloth (lower clothes), legs, shoes, accessories and background. Then, we rewrite $\varphi(\mathcal{I}_i)$ as a scalar set $\{\varphi(\mathcal{I}_i)_s^c\}, s = [0, S-1], c = [0, C-1]$, where $\varphi(\mathcal{I}_i)_s^c$ denotes the probability of the s^{th} grid belonging to the c^{th} region.

2) Parsing Attention Module: In this paper, we propose two parsing attention modules with one implemented in low-level layers and the other in high-level layers. Both of the parsing attention modules capture the discriminative and attentive features through a *split-and-aggregate* way. To be precise, features from different semantic human regions are first split according to the pedestrian parsing map and then aggregated using convolutional operations. In this way, the module learns where to pay attention to and how to aggregate the features from different regions, which provides more discriminative and more parsing-attention-oriented features for the following work and especially for the final attribute classification. The proposed parsing attention modules are shown in Fig. 3. In both modules, we conduct a probability updating operation and a replicated operation on the parsing map before employing it to split the features. The probability updating operation is shown in Eq. (3).

$$\bar{\varphi}(\mathcal{I}_i)_s^c = \begin{cases} 1 & c = \arg\max_{\varsigma}\{\varphi(\mathcal{I}_i)_s^{\varsigma}, \varsigma = [0, C-1]\} \\ \varphi(\mathcal{I}_i)_s^c & \text{otherwise} \end{cases} \tag{3}$$

If the s^{th} grid, compared to all other regions, is more likely to belong to the c^{th} region, then the $\bar{\varphi}(\mathcal{I}_i)_s^c$ would be set to 1, which can safeguard the information integrity of each semantic region as much as possible in the later attention feature fusion stage. Otherwise, its probability value would remain unchanged, which is used to partially reserve the information of that pixel to reduce the information loss caused by an error prediction. Moreover, the resize and replicated operation $\mathcal{R}(\cdot)$ is then applied to each of the updated parsing maps corresponding to different semantic regions. More specifically, for each semantic region, the operation $\mathcal{R}(\cdot)$ first resizes its parsing map into the size of corresponding features maps through 2D bilinear interpolation. After that, $\mathcal{R}(\cdot)$ further replicates the resized parsing map into parsing maps which have the same channels as input features. Furthermore, it is just by conducting this replication that we can carry out the element-wise production normally to implement the splitting operation. As a result, the splitting features for the c^{th} semantic region in the first attention module can then be obtained via:

$$\phi(\mathcal{I}_i)_{low}^c = \mathcal{I}_i^{\mathcal{B}} \otimes \mathcal{R}(\bar{\varphi}(\mathcal{I}_i)^c) \tag{4}$$

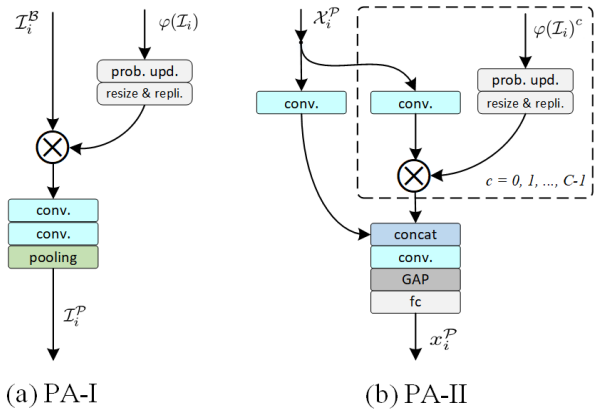


Fig. 3. The network structure of the proposed parsing attention modules: (a) the first parsing attention module (PA-I) and (b) the second parsing attention module (PA-II). GAP denotes the Global Average Pooling, and $\otimes$ represents the element-wise production.

where $\otimes$ denotes the element-wise production. The final splitting features are the ensemble of that of each semantic region, where $\phi(\mathcal{I}_i)_{low} = [\phi(\mathcal{I}_i)_{low}^0, \cdots, \phi(\mathcal{I}_i)_{low}^{C-1}]$.

However, the splitting operation in the second module is different from that of the first module in two aspects. The first difference comes from the fact that feature maps in low-level layers often have a small number of channels (e.g. 64), while feature maps in high-level layers usually contain a large number of channels (e.g. 1024). Therefore, a convolutional layer with only 128 kernels is first conducted on the high-level features $\mathcal{X}_i^{\mathcal{P}}$ to reduce the channel dimension (shown in Fig. 3), which can effectively release the computation burden and reduce the number of parameters in a later aggregating stage. This convolution step can be formulated by

$$\phi(\mathcal{I}_i)_{high}^c = Conv(\mathcal{X}_i^{\mathcal{P}}) \otimes \mathcal{R}(\bar{\varphi}(\mathcal{I}_i)^c) \tag{5}$$

Similarly, the final splitting features are also the concatenation of that of each semantic region, where $\phi(\mathcal{I}_i)_{high} = [\phi(\mathcal{I}_i)_{high}^0, \cdots, \phi(\mathcal{I}_i)_{high}^{C-1}]$. On the other hand, the second difference lies in that we add a global branch to the module. This idea of branch adding originates from an unsatisfying fact that features in high-level are usually in small size (e.g. 7×7) where the boundaries between different regions are ambiguous and features from a single region may be split into different parts. Consequently, the integrity of the feature will be jeopardized. In order to preserve this integrity, a global branch which concatenates the splitting features with the global features is added into the module as is shown in Fig. 3. It then generates some more comprehensive features.

In both low-level and high-level modules, after the splitting process, a few layers like convolutional layers, pooling layers and so on are employed to learn how to aggregate features from different regions, which as a whole is called the *split-and-aggregate*. This framework serves as a very useful tool in providing more attention-oriented features even when there are large discrepancies among different pedestrian images. In particular, the splitting process plays the role of a pre-refined procedure and aggregating integrates these information together for the subsequent network operations. For the

sake of clarity, the detailed structure is provided in Fig. 3. As is shown in the figure, the first parsing attention module generates the discriminative features $\mathcal{I}_i^{\mathcal{P}}$ for high-level layers, and the second parsing attention module produces the discriminative features $x_i^{\mathcal{P}}$ for the subsequent attribute classifiers. Moreover, the second parsing attention module is followed by $\mathcal{T}$ softmax classifiers, and the cross-entropy losses are employed to train the network. The objective loss function in the parsing branch for the j^{th} attribute can be defined by:

$$\mathcal{J}_j^{\mathcal{P}} = -\frac{1}{m}\left[\sum_{i=0}^{m-1}\sum_{k=0}^{K_j-1}\delta(y_i^j,k)\rho_k^j\log(p_k^j(x_i^{\mathcal{P}}))\right] \tag{6}$$

In the above equation, $p_k^j(x_i^{\mathcal{P}}) = \frac{\exp((\theta_{jk}^{\mathcal{P}})^T x_i^{\mathcal{P}})}{\sum_l \exp((\theta_{jl}^{\mathcal{P}})^T x_i^{\mathcal{P}})}$ denotes the probability of assigning the features $x_i^{\mathcal{P}}$ the k^{th} class for the j^{th} attribute with K_j denoting its number of classes. Besides, $\{\theta_{jl}^{\mathcal{P}}\}_{l=1}^{K_j}$ denotes the parameters of the j^{th} softmax classifier; $\delta(q1,q2)$ is the Kronecker delta function, where $\delta(q1,q2)=1$ if $q1=q2$, and $\delta(q1,q2)=0$, otherwise; K_j indicates the number of classes of the j^{th} attribute. Moreover, ρ_k^j is a penalty coefficient used to alleviate the imbalanced data problem in pedestrian attribute classification. In our experiments, we set $\rho_k^j = \sqrt{\frac{1}{2r_k^j}}$, where r_k^j represents the ratio of the k^{th} class in the j^{th} attribute. To be specific, ρ_k^j becomes larger along with r_k^j decreasing, which shifts the bias of the classifier to favor the minority class. The sum of losses for all attributes in the parsing attention branch can be denoted as:

$$\mathcal{J}^{\mathcal{P}} = \sum\nolimits_j \mathcal{J}_j^{\mathcal{P}} \tag{7}$$

D. Label Attention Mechanism

One common strategy of attention in CNN networks is to employ a region selection sub-network which is implemented by generating the attention masks to perform feature recalibration [22], [24]–[26]. Inspired by the above works, we propose a novel label attention mechanism that aims at specially selecting fine-grained discriminative pixels and regions for each attribute. This can be achieved by incorporating a sub-network to generate different selection masks for different attributes. The structure of the proposed label attention module is shown in Fig. 4 (a).

We denote the input features as $\mathcal{X}_i^{\mathcal{L}} \in \mathrm{R}^{n\times h\times w}$, where n, h and w denote the number of channels, height and width of the feature tensor, respectively. In the proposed label attention module, $\mathcal{X}_i^{\mathcal{L}}$ would be first forward propagated into two sub-networks, where one is used to extract higher-level features and also reduce the number of channels for the feasibility of training, and the other one is used to generate the selection masks. The higher-level features can be denoted as:

$$\mathcal{X}_i^{\mathcal{L},fea} = W_2^{\mathcal{L},fea}\cdot ReLU(W_1^{\mathcal{L},fea}\cdot\mathcal{X}_i^{\mathcal{L}}) \tag{8}$$

where $W_1^{\mathcal{L},fea}$ and $W_2^{\mathcal{L},fea}$ are the parameter matrixes of two convolutional layers in the sub-network for feature extraction. The attention masks are produced in a similar way but with an additional softmax function to select important pixels and regions. The attention masks can be generated by:

$$\mathcal{M}_i^{\mathcal{L}} = softmax(W_2^{\mathcal{L},mask}\cdot ReLU(W_1^{\mathcal{L},mask}\cdot\mathcal{X}_i^{\mathcal{L}})) \tag{9}$$

where $W_1^{\mathcal{L},mask}$ and $W_2^{\mathcal{L},mask}$ are the parameter matrixes of two convolutional layers in the sub-network for mask generation. The softmax function is employed to spatially normalize each attention mask, where it is conducted along with the height and width dimension. In our implementation, the two sub-networks have the same structure except for the softmax layer. It ensures the features $\mathcal{X}_i^{\mathcal{L},fea}$ and the attention masks $\mathcal{M}_i^{\mathcal{L}}$ to have the same dimensions, which is necessary in the later fusion stage. In fusion, the features $\mathcal{X}_i^{\mathcal{L},fea}$ are recalibrated by multiplying with the mask $\mathcal{M}_i^{\mathcal{L}}$ element-by-element to produce the attentive features $\mathcal{X}_i^{\mathcal{L},att}$, which can be represented as:

$$\mathcal{X}_i^{\mathcal{L},att} = \mathcal{X}_i^{\mathcal{L},fea}\otimes\mathcal{M}_i^{\mathcal{L}} \tag{10}$$

The attentive features $\mathcal{X}_i^{\mathcal{L},att} \in \mathrm{R}^{(r\times\mathcal{T})\times h\times w}$ have $r\times\mathcal{T}$ channels with each r channels capturing the discriminative features specially for each attribute. For clarity, the attentive features $\mathcal{X}_i^{\mathcal{L},att}$ are further denoted as an ensemble of the features of $\mathcal{T}$ attributes, i.e., $\mathcal{X}_i^{\mathcal{L},att} = [\mathcal{X}_{i,0}^{\mathcal{L},att},\cdots,\mathcal{X}_{i,\mathcal{T}-1}^{\mathcal{L},att}]$ and $\mathcal{X}_{i,j}^{\mathcal{L},att} \in \mathrm{R}^{r\times h\times w}$. To constrain that each r channels generate the attentive features only for a specific attribute, multiple constrained subnetworks are employed. Each of the $\mathcal{T}$ constrained subnetworks consists of a convolutional layer, a global average pooling layer and a classifier corresponding to a specific attribute of the total $\mathcal{T}$ attributes. To mention it, each of the $\mathcal{T}$ constrained subnetworks only connects with different r channels in $\mathcal{X}_i^{\mathcal{L},att}$. In other word, the j^{th} constrained subnetwork is only connected with the features $\mathcal{X}_{i,j}^{\mathcal{L},att}$. In this way, $\mathcal{X}_{i,j}^{\mathcal{L},att}$ is learned with a constrained subnetwork under the supervision of the labels of j^{th} attribute only, which prompts the module to generate the attention-oriented and discriminative features for each attribute specially. However, since employing multiple subnetworks seems to be very tedious, we can replace them by group convolutional layers [13], [51]. For the sake of convenience, we use $x_{i,j}^{cons}$ to denote the features that were generated by the global average pooling layer for the j^{th} attribute, and the cross-entropy loss is employed to train the classifiers. The objective loss function for the j^{th} classifier can be written as:

$$\mathcal{J}_j^{cons} = -\frac{1}{m}\left[\sum_{i=0}^{m-1}\sum_{k=0}^{K_j-1}\delta(y_i^j,k)\rho_k^j\log(p_k(x_{i,j}^{cons}))\right] \tag{11}$$

In the above equation, $p_k(x_{i,j}^{cons}) = \frac{\exp((\theta_{jk}^{cons})^T x_{i,j}^{cons})}{\sum_l \exp((\theta_{jl}^{cons})^T x_{i,j}^{cons})}$ denotes the probability of assigning the features $x_{i,j}^{cons}$ to the k^{th} class for the j^{th} attribute, and $\{\theta_{jl}^{cons}\}_{l=1}^{K_j}$ denote the parameters of the j^{th} softmax classifier. And the sum loss for all attributes also can be represented as $\mathcal{J}^{cons} = \sum_j \mathcal{J}_j^{cons}$.

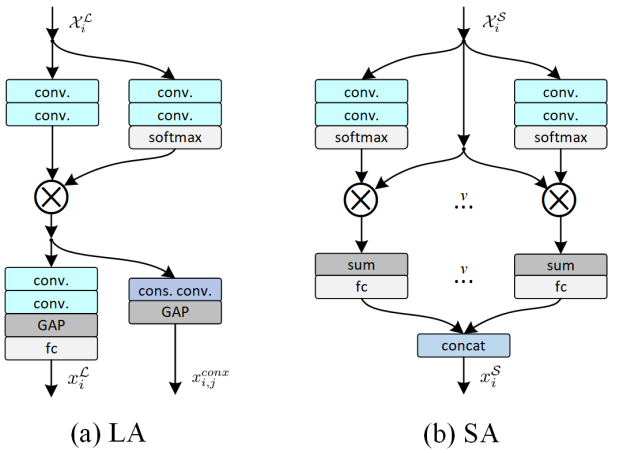


Fig. 4. The network structure of (a) the proposed label attention (LA) where GAP denotes the Global Average Pooling and (b) the spatial attention module (SA) where the global sum layer is abbreviated as sum, and $\otimes$ denotes the element-wise production.

In our implementation, we set $r \times \mathcal{T} \ll n$, which is inspired by the encoder-and-decoder in the work [3], [52]. More specially, the input features $\mathcal{X}_i^{\mathcal{L}}$ are first encoded into the features $\mathcal{X}_i^{\mathcal{L},att}$ with a smaller number of channels (e.g. $1 \times \mathcal{T}$, $2 \times \mathcal{T}$), and then decoded by a subsequent non-linear maps with a large number of channels (e.g. 512), including two convolutional layers, a global average pooling layer and a fully connected layer as shown in Fig. 4 (a). Let $x_i^{\mathcal{L}}$ denote the features of the final fully connected layer, and $\mathcal{T}$ softmax classifiers are employed for attribute classifications. The loss function for j^{th} classifier $\mathcal{J}_j{}^{\mathcal{L}}$ is similar to Eq. (6), and it is omitted here for convenience. The sum of losses for all attributes in the label attention branch can be written by $\mathcal{J}^{\mathcal{L}} = \sum_j \mathcal{J}_j^{\mathcal{L}}$.

E. Spatial Attention Mechanism

The spatial attention mechanism, as shown in Fig. 4 (b), is proposed to recalibrate the feature responses for selecting the important pixels and regions for all attributes. By following the works [23], [24], [28], we employ a small subnetwork to generate the attention mask. However, the truth is that attention masks often only cover small and most discriminative regions of object of interest when supervised by only a classification loss [53]. Therefore, multiple subnetworks are employed to generate different attention masks, which enables the selected features to be more abundant and comprehensive. We denote the input features by $\mathcal{X}_i^{\mathcal{S}}$. Assume we employ v attention masks in total, and the v^{th} attention mask is represented as $\mathcal{M}_{i,v}^{\mathcal{S}}$. In spatial attention mechanism, $\mathcal{M}_{i,v}^{\mathcal{S}}$ is produced by two convolutional layers and a softmax layer:

$$\mathcal{M}_{i,v}^{\mathcal{S}} = softmax(W_{2,v}^{\mathcal{S},mask} \cdot ReLU(W_{1,v}^{\mathcal{S},mask} \cdot \mathcal{X}_i^{\mathcal{S}})) \quad (12)$$

Similar to the previous label attention, the softmax function is also employed to spatially normalize each attention mask. Then, an element-wise multiplication is conducted between $\mathcal{X}_i^{\mathcal{S}}$ and $\mathcal{M}_{i,v}^{\mathcal{S}}$ to generate the attentive features $\mathcal{X}_i^{\mathcal{S},att}$, i.e.

$$\mathcal{X}_{i,v}^{\mathcal{S},att} = \mathcal{X}_i^{\mathcal{S}} \otimes \mathcal{M}_{i,v}^{\mathcal{S}} \quad (13)$$

After importing each $\mathcal{X}_{i,v}^{\mathcal{S},att}$ into its corresponding global sum layer and fully connected layer, we obtain the resulting higher-level features and all these features are then concatenated together. We use $x_i^{\mathcal{S}}$ to denote the concatenated features, and $\mathcal{T}$ softmax classifiers are used for attributes classification. The loss function for the j^{th} classifier $\mathcal{J}^{\mathcal{S}}$ is similar to Eq. (6), and it is omitted here for convenience. The sum of losses for all attributes in this branch can be represented as $\mathcal{J}^{\mathcal{S}} = \sum_j \mathcal{J}_j^{\mathcal{S}}$.

One thing we need to clarify is that though we have not utilized any pooling layers in this module, these spatial attention modules themselves can be actually treated as pooling-like operations. Concretely, when it comes to the global average pooling, we can find that it is conducted by first assigning an equal probability to each pixel and then taking the sum of them. Similarly, the probability assignment in our module is achieved by employing an element-wise production. However, one advantage of our module compared to the global average pooling is that the probability assigned to each pixel of the input feature is different and is learned automatically by the network. As a result, our pooling-like module has the nature of being more adaptive and therefore more effective.

F. The Loss Function

The total loss J is the sum of the loss functions corresponding to all branches, which can be defined by

$$\mathcal{J} = \mathcal{J}^{\mathcal{P}} + \mathcal{J}^{\mathcal{L}} + \lambda \mathcal{J}^{cons} + \mathcal{J}^{\mathcal{S}} \quad (14)$$

where λ is the weighting parameter for the constrained sub-networks and we set $\lambda = 0.2$ in our experiments. Under the guidance of $\mathcal{J}$, the network learns to extract the discriminative features from different views. The three branches are jointly learned concurrently to explore the correlated and complementary information. In the test stage, the average results of three branches are used as the final prediction.

IV. Experiments

A. Datasets and Settings

1) Pedestrian Attribute Datasets: We conduct experiments on pedestrian attribute analysis with five popular datasets: PA-100K [6], RAP [54], PETA [32], Market-1501 attribute [5] and Duke attribute [5] datasets.

PA-100K dataset is the largest dataset for pedestrian attribute classification where 100,000 pedestrian images from 598 outdoor scenes are included in total. 26 commonly used attributes, like gender, age, handbag, upper-clothing, etc., are annotated in this dataset. Following the settings in [6], the dataset is divided into 3 data subsets for evaluation: 80,000 images for training, 10,000 images for validation and 10,000 images for testing.

RAP dataset contains 41,585 pedestrian images captured by indoor scenes of a shopping mall. Each image in this dataset is annotated with 72 attributes, viewpoints, occlusions and body parts. We follow the official protocol provided by Li *et al.* [54], where 51 attributes with a positive label ratio above 1 % are employed for evaluation. Following the work [54], we evaluate our model on this dataset with 5 random splits, where 33,268 images are used for training and

8,317 images are used for testing in each resulting data subset. The final performance is the average over all data subsets.

PETA dataset provides 19,000 images collected from various outdoor scenes. Each image is annotated with 61 binary and 4 multi-class attributes. Following the evaluating protocol [32], 35 binary attributes, including 15 important attributes and 20 difficult yet interesting attributes, are selected for evaluation. The dataset is randomly split into 3 parts, where 9,500 images are used for training, 1,900 images are used for validation and the rest 7,600 images are used for testing.

Market-1501 attribute dataset contains 32,688 images of 1,501 identities. 10 binary attributes (such as gender and sleeve length) and 2 multi-class attributes (e.g. colors of upper body clothes) are annotated for each identity. Note that this dataset is annotated in the identity level. We evaluate our model on this dataset following the work [5] where the whole dataset is split into 751 identities for training and 750 identities for testing.

Duke attribute dataset contains 34,183 images from 1,812 identities. 8 binary attributes and 2 multi-class attributes are annotated for each identity. Similar to Market-1501 attribute dataset, this dataset is also annotated in the identity level. Following the work [5], the dataset is split into two parts where 16,522 images for training and 17,661 images for testing.

2) Pedestrian Parsing Datasets: The existing datasets, e.g., PennFudan dataset [39] and PPSS [29] are not suitable for the task of facilitating attribute analysis. For PennFudan dataset, the amount of images contained in the dataset is too small (only 169 images are included) to be conducted for our cases. As for the PPSS dataset, though many pedestrians in this dataset are occluded, those occluded regions are still annotated based on an approximate estimation, which may be detrimental to our attribute analysis. Most importantly, both datasets do not provide the annotation of accessories, which is also an important part of pedestrian analysis. Due to the above considerations, we decide to collect a new dataset for pedestrian parsing.

a) Image labeling: Most of the annotation tools [55] are based on superpixels, but they are ineffective for the pedestrian images in surveillance scenes, since these images often have low resolutions and occlusions. Here, we propose a useful interactive algorithm for an accurate pixel-level annotation and the labeling pipeline can be found in Fig. 5 (a). Our labeling algorithm mainly includes the following steps:

Step 1: boundary labeling. Manually label the boundary for each semantic region, e.g., hair, face and legs. In this way, each part can be localized precisely. The boundary line should be closed, which is necessary for finding the regions in the next step.

Step 2: region labeling. Find the closed regions according to the boundaries, and then annotate each region with predefined categories, e.g., hair, face and legs. Specially, region labeling can reduce the complexity of pixel-level labeling and accelerate the annotation process.

Step 3: boundary assigning. Assign each boundary pixel to a category according to the k-NN classification algorithm. For each boundary pixel $x_{i,j}$, at first we make the decision

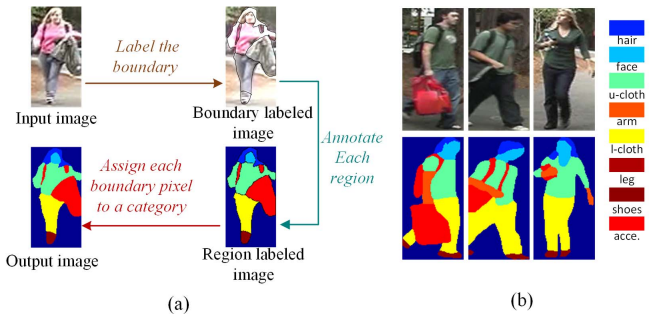


Fig. 5. (a) The labeling pipeline for accurate pixel-level annotation. (b) Some samples from our VIPeR parsing dataset.

with 1-NN, and the label $l_{i,j}$ is assigned by

$$l_{i,j} = \arg\max_{c} N_c \tag{15}$$

where N_c denotes the number of pixels that belongs to the c-th category in 1-nearest grids. If there are two categories with the same maximum number, we would assign the label to pixel $x_{i,j}$ using 2-NN, 3-NN and so on.

Moreover, considering that VIPeR is a very classical dataset for person re-identification and there are many accessories on the pedestrians of that dataset, we annotate VIPeR dataset with the proposed labeling algorithm. There are nine regions annotated in total, including "hair", "face", "u-clothes" (upper clothes), "arms", "l-cloth" (lower clothes), "legs", "shoes", "accessories" and "background". Some sample images are given in Fig. 5 (b).

b) Training protocols: The *VIPeR parsing dataset* contains 1,264 images from 632 pedestrians where each pedestrian has two images. When evaluating on PA-100K and RAP datasets, all images are employed to train the pedestrian parsing network so as to generate a promising parsing map. Considering that there is an overlap between VIPeR and PETA datasets, the images occurring in the test set of PETA therefore would not be used to train the pedestrian parsing network when we conduct experiments on PETA dataset. When it comes to specific steps of our experiment, 476 images are removed from the VIPeR dataset and the rest 788 images are used for training. Moreover, in order to figure out whether the proposed network structure has evident improvements compared to the baseline structure, we are supposed to evaluate the performance of our pedestrian parsing network and the baseline network. To this end, we divide the data into two parts which are the training set with the first 532 pedestrians images and the testing set containing the last 100 ones, respectively.

B. Implementation Details

In our experiments, we first train the pedestrian parsing network, and then keep its weight parameters unchanged to generate the parsing map for attributes inference. The input size of the pedestrian parsing network is 161×81. Due to the lack of training data, we initialize the network with the pretrained model provided by [45], and then pretrain the network with ATR dataset [56], which is the largest clothes

TABLE I

THE PARSING RESULTS ON VIPeR PARSING DATASET. THE RESULTS OF DEEPLAB AND PSPNET ARE RE-IMPLEMENTED BY US

Method	hair	face	u-cloth	arms	l-cloth	legs	shoes	acce.	Avg.
Deeplab [44]	65.48	77.56	91.08	60.05	92.24	87.68	54.20	36.83	70.64
PSPNet [45]	66.09	77.94	92.48	58.19	93.57	86.16	59.28	42.31	72.00
PSPNet† [45]	70.68	82.33	92.80	71.94	95.14	89.49	73.39	50.75	78.32
Attention+SSL [46]	70.74	82.74	93.15	70.64	94.37	89.87	74.26	43.78	77.44
Ours‡	72.70	82.83	93.06	71.92	95.16	88.92	74.11	53.49	**79.02**

† $\frac{1}{4}\times$ subsampling are employed, and ‡ $\frac{1}{4}\times$ subsampling and skip connections are employed.

Fig. 6. Parsing results of PPN on PA100K, PETA, RAP, Market-1501 and Duke datasets. Both good and bad results are given in the figure.

parsing dataset. Finally, the network is finetuned on the VIPeR parsing dataset. The weight decay and the momentum are set to 0.0005 and 0.9, respectively. The learning rate is started with 0.001 and reduced by a factor of 10 along with the number of iterations increases. For pedestrian attribute recognition, the input size of the bottom layers is 224×224. The weight decay and the momentum are set to be the same as above. The learning rate is started with 0.0001 with the consideration that there are dozens of attributes needed to be analyzed, and it is reduced by a factor of 10 along with the number of iterations increases. All models are trained and tested with Caffe [57] on GTX 1080Ti GPU.

C. Evaluation Metrics

For pedestrian attribute classification, when evaluating on PA-100K, PETA and RAP datasets, five criteria are employed following the works [3], [6], [54], including a label-based criteria, mean accuracy (mA), and four instance-based criteria, accuracy (Accu), precision (Prec), recall and F1. When evaluating on Market-1501 and Duke attribute datasets, we employ the accuracy for evaluation according to the work [5]. For pedestrian parsing, the per-pixel accuracy is employed for evaluation according to the previous works [29], [39].

D. Pedestrian Parsing Results

The quantitative results of our pedestrian parsing network on the VIPeR parsing dataset are given in Table I. PSPNet employs $\frac{1}{8}\times$ subsampling in their network for scene parsing. However, due to that the pedestrian images captured in surveillance scenes are usually presented with low resolution, $\frac{1}{8}\times$ subsampling easily results in features vanishing for pedestrian parsing. With this fact taken into consideration, $\frac{1}{4}\times$ subsampling is employed in our network. Moreover, the skip connections are employed in our network to capture the features at multi-level. Compared with the baseline method, PSPNet, the performance can be improved by 6.32 when using $\frac{1}{4}\times$ subsampling. Moreover, the performance can be further improved by 0.70% when skip connections is used. Additionally, compared the basedline PSPNet, we find that the improvements are primarily presented in small categories, e.g., hair, face, arms, legs. It also shows our method helps to avoid the feature vanishing in local regions. Moreover, we also re-implemented two popular methods, Deeplab [44] and Attention+SSL [46], for comparisons. The results are also presented in Table I. And the proposed method also performs better than them. Additional quantitative results on PA-100K, PETA, RAP, Market-1501 and Duke attribute datasets are shown in Fig. 6. Note that three datasets don't provide the ground truth labels for the raw images, so only the raw images and parsing results are listed in Fig. 6.

E. Detailed Analysis on Each Attention Mechanism

1) Analysis on Parsing Attention Mechanism: We propose two parsing attention modules in our network, with one for low layers and another for high layers, which are denoted as PA-I and PA-II, respectively. The plain SE-Net [24] is employed as a baseline method called the pedestrian attribute analysis (denoted as PAA). To notify, for the sake of simplicity, PA-I and PA-II together will be directly denoted by PA in the following work.

At first, we investigate how to set the number of semantic regions C in parsing maps would be better for attribute recognition. We take four different settings ($C = 1, 2, 5, 9$) for experiments, and the experiments are conducted with PA-I module. More specifically, there is only one part and the parsing map is filled with 1 everywhere when $C = 1$. Actually, it is also same as the network PAA without using parsing attention modules. Moreover, the parsing map is split into background and foreground parts when $C = 2$, and the parsing map is split into head, upper-body, lower-body and accessories

TABLE II
THE ANALYSIS OF THE NUMBER OF SEMANTIC REGIONS IN PA-I ON PA-100K DATASET

Num. (C)	mA	Accu	Prec	Recall	F1
1	79.99	77.34	86.10	86.54	86.32
2	80.43	77.44	86.25	86.50	86.38
5	80.20	77.54	86.24	86.58	86.41
8	80.69	77.71	86.10	87.02	86.56

TABLE III
THE ANALYSIS OF THE PARSING ATTENTION ON PA-100K DATASET

Method	mA	Accu	Prec	Recall	F1
PAA	79.99	77.34	86.10	86.54	86.32
PAA + PA-I	80.69	77.71	86.10	87.02	86.56
PAA + PA-II	80.86	77.80	86.10	86.98	86.54
PAA + PA	**81.23**	**78.15**	**86.49**	**87.15**	**86.82**

parts when $C = 4$. When $C = 9$, all the 9 semantic regions that have been mentioned above are employed. The experimental results are shown in Table II. The best performance is achieved by using $C = 9$, which demonstrate that the fine-grained parsing can be a better choice than coarse-grained parsing for assisting attribute recognition. Thus, all the parsing modules in the following experiments would choose fine-grained parsing for experiments.

Here, we analyze those two modules PA-I and PA-II step-by-step, and the experimental results are summarized in Table III. Both two parsing attention modules are useful to extract the discriminative features for pedestrian attribute analysis, where the average performances among five criteria are improved by 0.36% and 0.40%, respectively. When both parsing attention modules are employed, the average performance can be improved by 0.71% compared with the baseline PAA. Thus, we would take both two parsing attention modules together in our following experiments.

2) Analysis on Label Attention Mechanism: The label attention mechanism first encodes the features with $r \times \mathcal{T}$ channels, with each r channels features specially extracted for each attribute. r is a hyperparameter in the mechanism, and we conduct the experiments with various r to search its optimal value. The experiments are also conducted on the PA-100K dataset, and the results are shown in Table IV. When $r = 0$, it denotes the plain network without label attention mechanism used. As is shown in the table, the label attention mechanism really works by extracting discriminative features to achieve better performance. When r is small, the encoded features can't reserve enough information for dozens of attribute analysis. When r is large, the encoded features may contain redundant information, which may be detrimental for attribute analysis. The best results are achieved with $r = 4$, and we also take $r = 4$ in later experiments.

3) Analysis on Spatial Attention Mechanism: The spatial attention mechanism employs multiple attention masks to enable the selected features to be more abundant and comprehensive. Assume that v attention masks are employed in total. The experiments with various v are conducted on PA-100K. Experimental results are shown in Table V. When $v = 0$, it denotes the plain network without label attention mechanism used. In order to facilitate the comparison, the average performance among five criteria is also listed in the Table. The conclusion is similar, where a small number of masks can not reserve enough information and a large number of masks may contain redundant information. The best results are achieved when $v = 6$, and it will be used in later experiments.

TABLE IV
THE ANALYSIS OF THE LABEL ATTENTION ON PA-100K DATASET

Channels (r)	mA	Accu	Prec	Recall	F1
0	79.99	77.34	86.10	86.54	86.32
1	80.75	77.94	86.19	87.02	86.60
2	80.69	77.97	86.21	87.12	86.67
4	**81.32**	**78.10**	**86.35**	87.13	**86.74**
8	81.01	77.98	86.25	**87.14**	86.69

TABLE V
THE ANALYSIS OF THE SPATIAL ATTENTION ON PA-100K DATASET

Num. (v)	mA	Accu	Prec	Recall	F1	Avg.
0	79.99	77.34	86.10	86.54	86.32	83.26
1	80.59	77.07	86.01	86.11	86.06	83.17
2	81.32	77.75	86.32	86.69	86.50	83.71
4	81.19	77.96	86.25	87.07	86.65	83.82
6	81.27	77.98	86.19	87.15	86.87	**83.89**
8	81.10	77.93	86.07	87.23	86.65	83.80

TABLE VI
THE ANALYSIS OF THE JOINT LEARNING OF THREE ATTENTION MECHANISMS ON PA-100K DATASET

Method	mA	Accu	Prec	Recall	F1
PAA + PA	81.23	78.15	86.49	87.15	86.82
PAA + LA	81.32	78.10	86.35	87.13	86.74
PAA + SA	81.27	77.98	86.19	87.15	86.87
PAA + 3×SA	81.29	78.18	86.12	87.46	86.78
JLLS-PAA	81.50	78.67	86.51	87.71	87.11
JLPLS-PAA (ours)	**81.61**	**78.89**	**86.83**	**87.73**	**87.27**

4) Joint Learning of Three Attention Mechanisms: In this section, we analyze the joint learning mechanism step-by-step. We start from the network with single attention mechanism. Then, two attention mechanisms are jointly learned, where the label attention and spatial attention are used for experiments. Finally, we test the model with a joint learning of three attention mechanisms. The results are shown in Table VI. The JLLS-PAA represents the Joint Learning of Label attention and Spatial attention for pedestrian attribute analysis, and it achieves the better performance compared with the model with single attention mechanism. When three attention mechanisms are employed, the performance can be further improved where the correlated complementary features are captured. Furthermore, an additional model incorporated with three spatial attention branches (denoted as PAA + 3×SA) is employed for comparisons. PAA + 3×SA can only improve the performance slightly although it involves more parameters compared with PAA + SA. This demonstrates that the performance improvements of JLPLS-PAA mainly come from the joint learning of different attention mechanisms rather than involving more parameters.

TABLE VII
THE COMPARISONS ON PA-100K DATASET

Method	Backbone	mA	Accu	Prec	Recall	F1
DeepMar[36]	CaffeNet	72.70	70.39	82.24	80.42	81.32
M-net[6]	Inception v2	72.30	70.44	81.70	81.05	81.38
HP-net[6]	Inception_v2	74.21	72.19	82.97	82.09	82.53
Fusion[61]	CaffeNet	74.95	73.08	84.36	82.24	83.29
PAA	SE-BN-Incep.	79.99	77.34	86.10	86.54	86.32
PAA + PA	SE-BN-Incep.	81.23	78.15	86.49	87.15	86.82
PAA + LA	SE-BN-Incep.	81.32	78.10	86.35	87.13	86.74
PAA + SA	SE-BN-Incep.	81.27	77.98	86.19	87.15	86.87
JLPLS-PAA (ours)	SE-BN-Incep.	**81.61**	**78.89**	**86.83**	**87.73**	**87.27**

TABLE VIII
THE COMPARISONS ON RAP DATASET

Method	Backbone	mA	Accu	Prec	Recall	F1
CNN+SVM[54]	CaffeNet	72.28	31.72	35.75	71.78	47.73
ACN[38]	CaffeNet	69.66	62.61	**80.12**	72.26	75.98
DeepMar[36]	CaffeNet	73.79	62.02	74.92	76.21	75.56
Fusion[61]	CaffeNet	74.31	64.57	78.86	75.90	77.35
VeSPA [4]	GoogleNet	77.70	67.35	79.51	79.67	79.59
JRL[3]	AlexNet	77.81	–	78.11	78.98	78.58
PAA	SE-BN-Incep.	81.19	66.63	77.25	81.14	79.15
PAA + PA	SE-BN-Incep.	80.63	67.05	78.03	80.75	79.37
PAA + LA	SE-BN-Incep.	80.82	67.39	78.07	81.20	79.61
PAA + SA	SE-BN-Incep.	81.69	66.95	77.20	81.74	79.41
JLPLS-PAA (ours)	SE-BN-Incep.	**81.25**	**67.91**	78.56	**81.45**	**79.98**

TABLE IX
THE COMPARISONS ON PETA DATASET

Method	Backbone	mA	Accu	Prec	Recall	F1
CNN+SVM[54]	CaffeNet	76.65	45.41	51.33	75.14	61.00
ACN[38]	CaffeNet	81.15	73.66	84.06	81.26	82.64
DeepMar[36]	CaffeNet	82.89	75.07	83.68	83.14	83.41
HP-net[6]	Inception_v2	81.77	76.13	84.92	83.24	84.07
Fusion[61]	CaffeNet	82.97	78.08	86.86	84.68	85.76
VeSPA [4]	GoogleNet	83.45	77.73	86.18	84.81	85.49
JRL[3]	AlexNet	**85.67**	–	86.03	85.34	85.42
VAA[62]	Densenet-201	84.59	78.56	86.79	86.12	86.46
PAA	SE-BN-Incep.	82.81	77.02	86.25	84.33	85.28
PAA + PA	SE-BN-Incep.	83.64	77.73	86.22	85.11	85.66
PAA + LA	SE-BN-Incep.	84.08	78.19	86.49	85.42	85.96
PAA + SA	SE-BN-Incep.	83.72	77.84	86.38	85.02	85.69
JLLS-PAA	SE-BN-Incep.	84.56	79.20	87.44	85.86	86.64
JLPLS-PAA (ours)	SE-BN-Incep.	84.88	**79.46**	**87.42**	**86.33**	**86.87**

F. Comparison to the Prior Arts

In this Section, we mainly compare the proposed methods with previous methods. The backbone of those methods, e.g., AlexNet [13], CaffeNet [57], Inception_v2 [58], GoogleNet [59], DenseNet-201 [60], ResNet-50 [15] and SE-BN-Inception [24], are also clarified.

1) Results on PA-100K Dataset: For PA-100K dataset, we compare our approach with DeepMar [36], M-net [6] and HP-net [6]. The comparisons are summarized in Table VII. All three attention mechanisms can improve the performance for pedestrian attribute analysis. In addition, the joint learning of three attention mechanisms can further improve the performance. The proposed JLPLS-PAA achieves the best accuracy on PA-100K dataset given all five evaluation metrics. To be specific, JLPLS-PAA outperforms the second best method HP-net [6] by 7.40%, 6.70%, 3.86%, 5.64% and 4.74% on mA, Accu, Prec, Recall and F1 respectively. It is a huge improvement, and it also shows clearly the benefits of the proposed JLPLS-PAA for pedestrian attribute recognition.

2) Results on RAP Dataset: Table VIII summarizes the comparisons on RAP dataset. The methods used for comparison include CNN+SVM [54], ACN [38], DeepMar [36], VeSPA [4] and JRL [3]. As is shown in Table VIII, different attention mechanisms enhance the performance for different criteria. For example, the parsing attention and label attention mainly improve the performance on Accu, Prec and F1, while the spatial attention mainly improves the performance on mA and recall. This also shows that the different attention mechanisms are complementary. The joint model JLPLS-PAA achieves the best accuracy by given four metrics except Prec. (Ours 78.56%, and the best 80.12% achieved by ACN [38]). However, ACN [38] achieves much lower performance on the other metrics compared with JLPLS-PAA , e.g., 81.25% by JLPLS-PAA vs. 69.66% by ACN on mA. For the other four metrics, our JLPLS-PAA improves the second best results by 3.44%, 0.56%, 1.78% and 0.39% on mA, Accu, Recall and F1 respectively.

3) Results on PETA Dataset: For PETA dataset, we compare our approach with CNN+SVM [54], ACN [38], DeepMar [36], VeSPA [4] and JRL [3]. The experimental results are shown in Table IX. The proposed three attention mechanisms also show their effectiveness on this dataset, with considerable performance promotions. When the joint learning strategy, the performance is further improved by a lot. Compared with the baseline method PAA (a plain SE-Net for pedestrian attribute analysis), our JLPLS-PAA improves the average performance among five criteria by 1.85%. The proposed JLPLS-PAA outperforms the prior arts on the given four metrics of Accu, Prec, Recall and F1. On the metrics of mA, the proposed JLPLS-PAA achieves the second best result and the best result are achieved by JRL. Note JRL is an ensemble model of multiple networks, where 10 RNN networks are included in total and each RNN network employs 6 Alexnet networks to extract features. Compared with JRL, our model is more slight and is constructed based on a single network. On the whole, our model achieves the comparable performance with JRL by using a single network on PETA dataset.

4) Results on Market-1501 Attribute Dataset: Market-1501 attribute dataset is a newly released dataset, hence only a small amount works have been evaluated on this dataset. For Market-1501 attribute dataset, we compare our approach with PedAttriNet [5], APR [5]. The works [63], [64] evaluate their model on the Market-1501 attribute dataset with different evaluation protocol. Therefore, those two works [63], [64] are not selected for comparisons. The comparisons are shown in Table X. All three attention mechanisms can yield the performance improvements on this dataset. The joint learning of three attention mechanism further improves the performance to 87.88%, which outperforms the previous best method by 2.55%.

5) Results on Duke Attribute Dataset: Similar to the comparisons on market-1501 attribute dataset, we also compare our

TABLE X
THE COMPARISONS ON MARKET-1501 AND DUKE ATTRIBUTE DATASETS

Dataset	Backbone	Market-1501 (%)	Duke (%)
PedAttriNet [5]	ResNet-50	84.64	80.07
APR [5]	ResNet-50	85.33	80.12
PAA	SE-BN-Incep.	86.34	83.95
PAA + PA	SE-BN-Incep.	87.69	84.82
PAA + LA	SE-BN-Incep.	86.56	84.22
PAA + SA	SE-BN-Incep.	86.94	84.35
JLPLS-PAA (ours)	SE-BN-Incep.	**87.88**	**85.24**

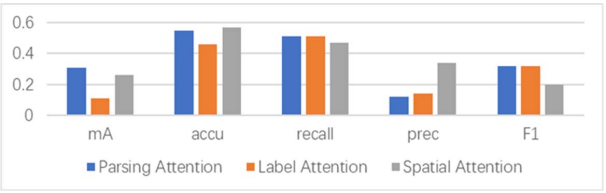


Fig. 7. The performance improvements for parsing attention, label attention and spatial attention with the joint learning framework on PA-100K dataset.

TABLE XI
THE COMPARISONS OF THE LEARNING A PARALLEL WAY OR SERIAL WAY ON PA-100K DATASET

Method	mA	Accu	Prec	Recall	F1
SLLS-PAA	80.59	76.08	84.57	86.34	85.46
JLLS-PAA	81.50	78.67	86.51	87.71	87.11

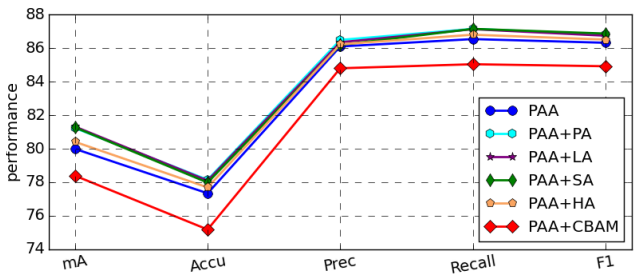


Fig. 8. The comparisons of different attention mechanisms on PA-100K dataset.

method with PedAttriNet [5], APR [5]. Each single attention mechanism and their joint learning form show their effectiveness on this dataset. Our proposed JLPLS-PAA achieves the state-of-the-art performance with an accuracy of 85.24%, which is 5.12% higher than the previous best method [5].

V. DISCUSSIONS

A. Analysis on the Joint Learning

We formulate a joint learning framework for parsing attention, label attention and spatial attention mechanisms, by considering the complementarity between different attention mechanisms. We evaluate the performance of each attention branch of the joint learning model, aiming at testing the effects on each branch taken by the joint learning. As is shown in Fig. 7, learning three attention branches concurrently and jointly can achieve better performance compared with learning each attention mechanism separately and individually. The improvements also validate the complementarity between three attention mechanisms, and learning them concurrently and jointly helps to extract the correlated and complementary features.

B. Learning in a Parallel Way or Serial Way?

In the proposed framework, three proposed attention mechanisms are learned in a parallel way, where each attention mechanism is incorporated into a separate branch and all branches are synergistically correlated and jointly learned. To further validate the effectiveness of the joint learning framework, a serial learning model is also formulated for comparisons. We take label and spatial attention mechanisms for analysis, and the experimental results are shown in Table XI. SLLS-PAA represents Serial Learning of Label attention and Spatial attention for Pedestrian Attribute Analysis, and JLLS-PAA denotes Joint/parallel learning of Label attention and Spatial attention for Pedestrian Attribute Analysis. As shown in Table XI, JLLS-PAA performs better than SLLS-PAA on all five metrics. Compared with parallel framework, serial framework may hardly be satisfied for all attention modules to extract effective features. For example, there may be cases where the information suppressed in a certain attention mechanism is actually useful in the next attention mechanism. What's more, the serial framework is more difficult to train than the parallel framework because the branch in the serial framework is deeper and contains more parameters than the branch in parallel framework.

C. Comparisons to Other Attention Mechanisms

We also compare the proposed attention mechanisms with other mechanisms, e.g., Harmonious Attention (HA) [28] and Convolutional Block Attention Module (CBAM) [65]. We implement those two attention modules by placing them behind the last convolutional layer (same to the proposed LA, SA and PA-II). The experimental results are shown in Fig. 8. Compared with HA and CBAM, the proposed LA, SA and PA modules perform better. It also validate the effectiveness of the proposed LA, SA and PA modules for pedestrian attributes analysis. In our experiments, the performance drops when incorporating the CBAM into the network. CBAM contains two components: channel and spatial attentions, and those components are formulated in a serial way to extract attentive features. The performance degradation may due to such serial components may hardly be learned with very limited data.

D. Efficiency Analysis

In this section, we mainly analyze the efficiency of the proposed method, including: parameters, memory complexity and speed. For comparison, the efficiency of some popular models like JRL [3] and VAA [62] would also be analyzed. The efficiency analysis is shown in Table XII. The most costly part is the Pedestrian Parsing Network (PPN), which is constructed based on ResNet-152. For the recognition networks (e.g., PAA, PAA+PA, PAA+SA, PAA+LA, JLLS-PAA and JLPLS-PAA), they are very efficient with fewer parameters, less memory complexity and faster speed compared with the previous methods, e.g., JRL and VAA. JLPLS-PAA needs PPN to generate the parsing maps, which is slight costly. Thus, JLLS-PAA, which is implemented without parsing attention

TABLE XII
THE COMPARISONS ON PARAMETERS, MEMORY COMPLEXITY AND SPEED

Model	#Param	#Memory Complexity	#GPU Speed
JRL*	58.3M×10	1051M×10	2.06ms×10
VAA†	20.2M	9287M	27.52ms
PPN	71.75M	12567M	39.45 ms
PAA	10.5M	1921M	8.67ms
PAA + PA	12.8M	2491M	10.67ms
PAA + LA	11.8M	1959M	9.15ms
PAA + SA	12.8M	2033M	9.91ms
JLLS-PAA	16.3M	2155M	11.98ms
JLPLS-PAA	21.0M	2803M	15.11ms

JRL* is a combined model based on 10 AlexNets, so its efficiency is estimated by using 10 AlexNets. VAA† is constructed based on the DenseNet-201, and its efficiency is estimated with a DenseNet-201. The Memory Complexity is calculated by using GPU memory footprint in the training process with a batch size of 10.

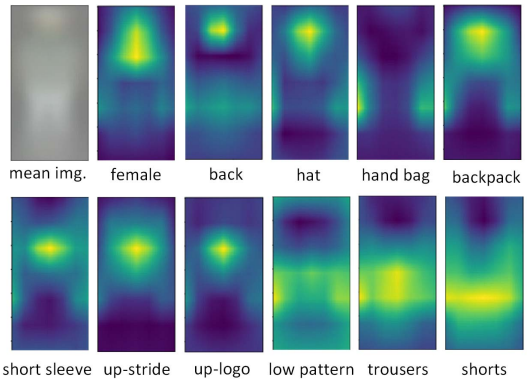


Fig. 9. Visualizations of the mean pedestrian image and 11 selected label attention maps on PA-100K dataset. Each attention map is drawn with averaging the activations among all images in the test set (10 thousand images in total).

modules, can be regarded as a trade off between memory loss cost and evaluation performance. JLLS-PAA also can achieve comparative performance but with drastically fewer cost compared with JRL and VAA as shown in Table IX and XII. Note that all the models are tested with Caffe [57] on GTX 1080Ti.

E. Visualizations of Label Attention

Fig. 9 shows the average label attention maps on the test set of the PA-100K dataset. We select 11 label attention maps in total, including gender, backpack, handbag and so on. The bright pixels represent the regions that the network focuses on for the corresponding label, and those regions are various for different labels. As is shown in Fig. 9, the network focuses head and upper body regions for gender attribute, and focuses head region for hat attribute. Moreover, for the attributes related to upper body e.g. short sleeve, up-stride and up-logo, the label attention mechanism also learns to select the regions of upper body. For the attributes related to lower body e.g. lower pattern, trousers and shorts, it also learns to focus the regions of lower body. The visualizations also qualitatively show the effectiveness of label attention mechanism.

F. Visualizations of Spatial Attention

Fig. 11 shows the average responses of the spatial attention masks on the test set of the PA-100K dataset. As is shown in Fig. 11, the attention models primarily focus on the foreground regions while suppressing the background regions. Each spatial attention mask aims to select the discriminative features from different human parts, e.g. head, upper body. In the spatial attention mechanism, all the selected features are gathered together with a concatenation operation, which can form more comprehensive features to achieve better recognition results.

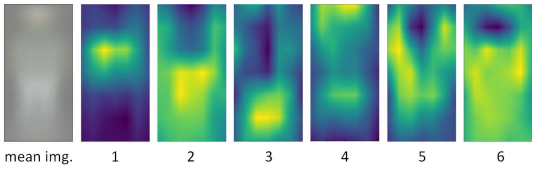


Fig. 10. Visualizations of the mean pedestrian image and six spatial attention maps on PA-100K dataset. The spatial attention map is drawn with averaging the activations among all images in the test set (10 thousand images in total). The 1^{th}, 2^{th}, 3^{th}, 4^{th}, 5^{th} and 6^{th} spatial attention map may mainly select the features from upper body, thigh, shank, head, arms and human body except for head.

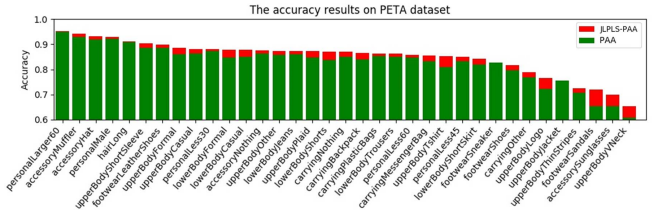


Fig. 11. The accuracy results (mA) of the proposed JLPLS-PAA and the baseline method PAA on PETA dataset.

G. Analysis on the Improvements

To analyze the improvement on each attribute, we draw the accuracy results on PETA dataset of the proposed JLPLS-PAA and PAA methods. Compared with the baseline PAA, the proposed JLPLS-PAA improve the performances of almost all attributes. Besides, the JLPLS-PAA outperforms PAA especially on these attributes with region-based saliency, e.g., "footwearSandals", "accessorySunglasses".

VI. CONCLUSION

The large variations of pedestrian images, e.g. large pose variations, complex backgrounds and various camera viewing angles, make the task of recognizing the pedestrian attributes a challenging work. In this paper, parsing attention, label attention and spatial attention have been developed to select the important and discriminative regions for pedestrian attribute analysis against the variations. Different attention mechanisms extract discriminative features by considering the problems from different perspectives, which are correlated and complementary. Moreover, a joint learning framework of the above three attention mechanisms has been formulated in a multi-task-like way to capture the correlated and complementary features. Various experiments conducted on PA-100K, PETA, RAP, Market-1501 and Duke attribute datasets further demonstrate the effectiveness of the proposed method.

ACKNOWLEDGMENT

The authors acknowledge the support of NVIDIA with the GPU donation for this research.

REFERENCES

[1] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *Proc. ICCVW*, Jun. 2013, pp. 331–338.

[2] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label CNN based pedestrian attribute learning for soft biometrics," in *Proc. ICB*, May 2015, pp. 535–540.

[3] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proc. ICCV*, vol. 2, 2017, pp. 531–540.

[4] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," in *Proc. BMVC*, 2017.

[5] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," 2017, *arXiv:1703.07220*. [Online]. Available: https://arxiv.org/abs/1703.07220

[6] X. Liu *et al.*, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. ICCV*, Oct. 2017, pp. 350–359.

[7] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, and Y. Chenggang, "Recurrent attention model for pedestrian attribute recognition," in *Proc. AAAI*, 2019.

[8] Q. Li, X. Zhao, R. He, and H. Kaiqi, "Visual-semantic graph reasoning for pedestrian attribute recognition," in *Proc. AAAI*, 2019.

[9] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, "Attribute-based people search: Lessons learnt from a practical surveillance system," in *Proc. CICMR*, 2014, p. 153.

[10] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. CVPR*, Jun. 2011, pp. 801–808.

[11] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. BMVC*, 2012, vol. 2, no. 3, p. 8.

[12] C. Shan, F. Porikli, T. Xiang, and S. Gong, *Video Analytics for Business Intelligence*. Berlin, Germany: Springer, 2012.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

[17] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cognit.*, vol. 7, nos. 1–3, pp. 17–42, 2000.

[18] J. Colombo, "The development of visual attention in infancy," *Annu. Rev. Psychol.*, vol. 52, no. 1, pp. 337–367, 2001.

[19] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural Comput.*, vol. 24, no. 8, pp. 2151–2184, Aug. 2012.

[20] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[21] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. CVPR*, Jul. 2017, pp. 4438–4446.

[22] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. CVPR*, Jul. 2017, pp. 3156–3164.

[23] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 369–378.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Jun. 2018, pp. 7132–7141.

[25] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. CVPR*, Jul. 2017, pp. 5513–5522.

[26] J. Si *et al.*, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 5363–5372.

[27] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. CVPR*, Jul. 2017, pp. 4747–4756.

[28] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 2285–2294.

[29] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. ICCV*, Dec. 2013, pp. 2648–2655.

[30] F. Xia, J. Zhu, P. Wang, and A. L. Yuille, "Pose-guided human parsing by an and/or graph using pose-context features," in *Proc. AAAI*, 2016, pp. 3632–3640.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 4, 2017, p. 12.

[32] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. ACM MM*, 2014, pp. 789–792.

[33] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014. pp. 93–117.

[34] Y. Zhao, Z. Jin, G.-J. Qi, H. Lu, and X.-S. Hua, "An adversarial approach to hard triplet generation," in *Proc. ECCV*, Sep. 2018, pp. 501–517.

[35] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 1062–1071.

[36] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. ACPR*, Nov. 2015, pp. 111–115.

[37] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *CoRR*, vol. abs/1611.05603, 2016. [Online]. Available: http://arxiv.org/abs/1611.05603

[38] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. ICCVW*, Dec. 2015, pp. 87–95.

[39] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *Proc. CVPR*, Jun. 2011, pp. 2265–2272.

[40] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. ICCV*, Sep./Oct. 2009, pp. 1365–1372.

[41] S. Eslami and C. Williams, "A generative model for parts-based object segmentation," in *Proc. NIPS*, 2012, pp. 100–107.

[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," 2016, *arXiv:1606.00915*. [Online]. Available: https://arxiv.org/abs/1606.00915

[44] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. CVPR*, Jul. 2017, pp. 2881–2890.

[46] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," 2017, *arXiv:1703.05446*. [Online]. Available: https://arxiv.org/abs/1703.05446

[47] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *Proc. CVPR*, Jun. 2018, pp. 2100–2108.

[48] C. Shen *et al.*, "Sharp attention network via adaptive sampling for person re-identification," 2018, *arXiv:1805.02336*. [Online]. Available: https://arxiv.org/abs/1805.02336

[49] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. ECCV*, 2014, pp. 94–108.

[50] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. NIPS*, 2007, pp. 41–48.

[51] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. CVPR*, Jun. 2018, pp. 6848–6856.

[52] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[53] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. CVPR*, Jun. 2018, pp. 9215–9223.

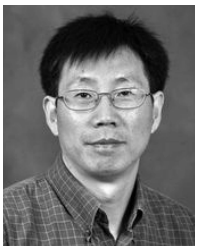
[54] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.
[55] P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," (2017), *arXiv:1703.01386*. [Online]. Available: https://arxiv.org/abs/1703.01386
[56] X. Liang *et al.*, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.
[57] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM*, 2014, pp. 675–678.
[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.
[59] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.
[60] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.
[61] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *Proc. ICME*, Jul. 2018, pp. 1–6.
[62] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. ECCV*, Sep. 2018, pp. 680–697.
[63] S. Huang, X. Li, Z.-Q. Cheng, Z. Zhang, and A. Hauptmann, "GNAS: A greedy neural architecture search method for multi-attribute learning," 2018, *arXiv:1804.06964*. [Online]. Available: https://arxiv.org/abs/1804.06964
[64] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proc. ACM MM*, 2017, pp. 1636–1644.
[65] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sep. 2018, pp. 3–19.

Zichang Tan received the B.E. degree from the Department of Automation, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences (CASIA). He was named an Outstanding Graduate of the college when he graduated. His main research interests include deep learning, face attribute analysis, pedestrian analysis, and face recognition.

Yang Yang received the B.E. and master's degree from Xidian University, China, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA). Since 2016, he has been an Assistant Professor with the National Laboratory of Pattern Recognition (NLPR), CASIA. He has published papers in top conferences and journals, such as AAAI, ECCV, and *Pattern Recognition*. His main research interests include person re-identification, pedestrian analysis, and face recognition.

Jun Wan received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, in 2015. Since 2015, he has been an Assistant Professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). He has published papers in top journals, such as JMLR, T-PAMI, and T-IP. His main research interests include computer vision and machine learning, especially for gesture and action recognition and facial attribution analysis (i.e., age estimation, facial expression, and gender and ethnicity classification). He has served as a Reviewer for several top journals and conferences, such as JMLR, T-PAMI, T-IP, T-MM, T-SMC, PR, CVPR, ICCV, ECCV, ICRA, BMVC, ICPR, and FG.

Hanyuan Hang received the B.S. degree in mathematics from Wuhan University, Wuhan, China, and the University of Stuttgart, Stuttgart, Germany, in 2007, the M.S. degree from the University of Stuttgart in 2010, and the Ph.D. degree from the University of Stuttgart in 2015. He is currently an Assistant Professor with the Institute of Statistics and Big Data (ISBD), Renmin University of China (RUC), Beijing, China. He has published papers in top journals, such as the *Journal of Machine Learning Research*, *Annals of Statistics*, *Neural Computation*, and the *Journal of Multivariate Analysis*. His main research interests include machine learning and deep learning, among which machine learning mainly includes algorithms, such as support vector machine and random forest.



Guodong Guo (M'07–SM'07) received the B.E. degree in automation from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of Wisconsin, Madison, WI, USA. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; and Microsoft Research, Beijing. He is currently the Deputy Head of the Institute of Deep Learning, Baidu Research, and also an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), USA. He has authored a book *Face, Expression, and Iris Recognition Using Learning-Based Approaches* (2008), has co-edited two books *Support Vector Machines Applications* (2014) and *Mobile Biometrics* (2017), and has published over 100 technical papers. His research interests include computer vision, biometrics, machine learning, and multimedia. He was a recipient of the North Carolina State Award for Excellence in Innovation in 2008, the Outstanding Researcher Award from 2013 to 2014 and from 2017 to 2018 at CEMR, WVU, and the New Researcher Award of the Year from 2010 to 2011 at CEMR, WVU. He was selected as the People's Hero of the Week by BSJB under the Minority Media and Telecommunications Council (MMTC) in 2013. Two of his papers were selected as The Best of FG'13 and The Best of FG'15, respectively. He is an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the *Journal of Visual Communication and Image Representation* and serves on the Editorial Board of *IET Biometrics*.

Stan Z. Li (F'09) received the B.Eng. degree from Hunan University, China, the M.Eng. degree from the National University of Defense Technology, China, and the Ph.D. degree from Surrey University, U.K. He was an Associate Professor with Nanyang Technological University, Singapore. He was with Microsoft Research Asia as a Researcher from 2000 to 2004. He is currently a Professor and the Director of the Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA). He has published more than 200 papers in international journals and conferences and has authored and edited eight books. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He was elevated to IEEE Fellow for his contributions to the fields of face recognition, pattern recognition, and computer vision, and he is a member of the IEEE Computer Society. He served as a Program Co-Chair for the International Conference on Biometrics 2007 and 2009 and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and is acting as the Editor-in-Chief of *Encyclopedia of Biometrics*.