Hierarchical Reasoning Network for Pedestrian Attribute Recognition

Haoran An, Hai-Miao Hu, Yuanfang Guo, Qianli Zhou, and Bo Li

***Abstract*—Pedestrian attribute recognition, which can benefit other tasks such as person re-identification and pedestrian retrieval, is very important in video surveillance related tasks. In this paper, we observe that the existing methods tackle this problem from the perspective of multi-label classification without considering the hierarchical relationships among the attributes. In human cognition, the attributes can be categorized according to their semantic/abstraction levels. The high-level attributes can be predicted by reasoning from the low-level and medium-level attributes, while the recognition of the low-level and medium-level attributes can be guided by the high-level attributes. Based on this attribute categorization, we propose a novel Hierarchical Reasoning Network (HR-Net), which can hierarchically predict the attributes at different abstraction levels in different stages of the network. We also propose an attribute reasoning structure to exploit the relationships among the attributes at different semantic levels. Experimental results demonstrate that the proposed network gives superior performances compared to the state-of-the-art techniques.**

***Index Terms*—Pedestrian attribute recognition, video surveillance, abstraction levels, hierarchical, reason.**

I. Introduction

PEDESTRIAN attribute recognition in surveillance footage is highly desired due to its great potentials in several multimedia applications. For example, in pedestrian re-identification, the recognized attributes can serve as mid-level features to effectively improve the capabilities of ReID [1]–[4]. In pedestrian retrieval, the queried attributes can be utilized to acquire the potential targets from a large amount of videos efficiently [5]–[8].

Similar to the other attribute-based tasks [9]–[16], pedestrian attribute recognition, which is essentially a multi-label classification task, accurately predicts a set of attributes, given an pedestrian image as input. Since these attributes may possess different characteristics such as color, texture, hair, clothing, accessories, gender, age, etc., they tend to vary differently and possess complex relationships with each other, and thus induce many challenges. In this paper, we focus on two specific problems: (i) how to distinguish various attributes according to their characteristics? (ii) how to fully exploit the relationships among the attributes?

Early work addresses this problem via traditional features and Support Vector Machine (SVM) to classify single attribute in [1], [8], [17]–[19]. Meanwhile, recent techniques employ deep learning to boost the recognition accuracy. [20] firstly introduces a convolutional neural network (CNN) with the parameter sharing strategy to adaptively explore the semantic relationships among the attributes. This end-to-end network is trained with independent loss layers which are designed for each attribute.

The deep learning based approaches can be categorized based on their mechanisms. Some researchers focus on improving the feature representations. [21] and [22] capture multi-level features by the network to exploit both the global and local information. [23] combines the high-level CNN extracted features and the low-level handcrafted Local Maximal Occurrence (LOMO) features to address the problem of viewpoint variations. [24] infers the attribute labels from a set of middle-level attribute-relevant features instead of directly classifying the attributes with respect to the CNN extracted features. Unfortunately, since they have ignored the correlations among the different representations, they usually generate redundant features and high computational overheads.

Some researchers consider to explore the dependencies among the attributes. [25] carries out the prediction of an attribute by a weighted combination of the independent decision score and the prediction scores from other attributes to build interaction models among different attributes. [26] proposes to adopt a recurrent neural network (RNN), especially the Long Short-Term Memory (LSTM), to jointly modelling the intra-person, inter-person, and inter-attribute context information to better exploit the semantic relationships among the attributes. Although the employment of LSTM has achieved certain performance improvements, the dependencies among the attributes are less reasonable according to the human cognitions. [27] proposes the deep learning based single attribute recognition model (DeepSAR) to recognize each attribute one by one, and then introduces the unified multi-attribute jointly learning model (DeepMAR) to learn all the attributes at the same time in which one attribute can contribute to the representation of other attributes.

Manuscript received April 30, 2019; revised October 27, 2019; accepted February 3, 2020. Date of publication February 20, 2020; date of current version December 17, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61772058, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Elisa Ricci. *(Corresponding author: Hai-Miao Hu.)*

Haoran An, Hai-Miao Hu, Yuanfang Guo, and Bo Li are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: waterhran@qq.com; frank0139@163.com; eeandyguo@connect.ust.hk; boli@buaa.edu.cn).

Qianli Zhou is with the People's Public Security University of China, Beijing 100038, China (e-mail: 13331112522@189.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2975417

Some researchers consider to group the attributes, and then explore the dependencies among the attributes. [28] groups the attributes by the correlation and proposes an end-to-end Grouping Recurrent Learning (GRL) model to take advantage of the mutual exclusions within the groups and inter-group correlations to improve the performance of pedestrian attribute recognition. [29] groups the attributes according to the principle that attributes in the same group are encouraged to share more knowledge with each other. Although these methods claim to learn the dependencies among the attributes, there is no separate structure in the network to represent this relationship. They do not provide the relationship among the attributes learned by the network.

Other researchers intend to exploit the location information to predicte the attributes because some attributes can only be recognized at certain locations. [30] and [31] segment the input image into different body parts to predict the corresponding attributes. [32] trains a Convolutional Neural Network (CNN) to select the most attribute-descriptive human parts from all poselet detections, and combines them with the whole body as a pose-normalized deep representation. It also further improves by using deep hierarchical contexts ranging from human-centric level to scene level. [33] proposes the Spatial Regularization Network (SRN) to generate attention maps for all labels and captures the underlying relations between them via learnable convolutions. [34] analyzes the spatial and semantic relations that exist in the images and proposes an effective method that extracts and aggregates visual attention masks at different scales with high prediction variance accounts for the weak supervision of the attention mechanism. Although the obtained attributes are similar to the results of human cognition, they failed to utilize the correlations among the attributes and it is inappropriate to employ one network structure to predict these attributes in different body parts.

In general, early deep learning based human attribute recognition methods [1], [8]–[10], [17]–[24] usually predict the attribute independently. Thus, their recognition performances are limited, because various attributes tend to behave differently and usually prefer to extract specific information. For example, color and pattern attributes require local information such as colors and textures, while gender and age attributes require strong semantic information. Latter approaches [12]–[14], [25], [26], [35], [36] mine certain simple relationships among the attributes and [14], [30]–[34] mine the relationship between the attributes and the locations. However, these methods have not considered the hierarchical nature of the attributes and the dependencies, which can be obtained by reasoning between them. Some approaches [11], [15], [16], [27]–[29] explore to group the attributes, but they are mostly based on the corelation among the attributes and have nothing to do with the human cognitions. Besides, the existing networks have not exploited any prior knowledge. For example, when a person wears a short skirt and earrings, the probability of the person being a woman should be higher than the probability when these attributes do not exist. The existing drawbacks motivate us to develop a new method, which can specifically handle various attributes from different semantic levels according to their characteristics and utilize the relationships among the attributes to boost the accuracy of pedestrian attribute recognition.

Fig. 1. The attributes can be categorized into three levels based on their semantic/abstraction levels to achieve a better recognition result.

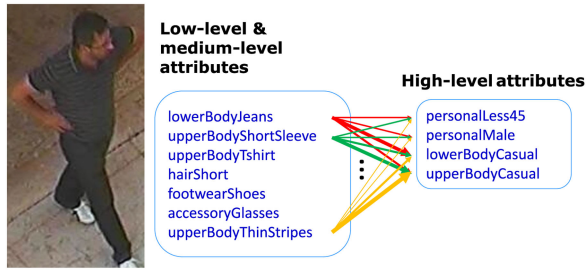


Fig. 2. The complex dependencies among the attributes at different semantic levels. Note that the size of the arrows demonstrates the magnitude of the correlations among different attributes.

Intuitively, we can categorize the attributes and recognize them with different features, as illustrated in Fig. 1, and effectively utilize some attributes to predict other attributes, as elaborated in Fig. 2. According to this intuition, we propose a deep network architecture, referred as Hierarchical Reasoning Network (HR-Net) as shown in Fig. 3, to extract the low-level, medium-level and high-level features for attribute predictions at different stages of the network. Besides, HR-Net also achieves the inferential predictions from the low-level attributes to the high-level ones with the help of the guidance, which is from the high-level predictions to the low-level ones.

The contributions of this work are summarized as below:

1) We propose a hierarchical categorization for the pedestrian attributes, which classifies the attributes into three levels according to their semantics and abstractions, to hierarchically tackle the pedestrian attribute recognition task.
2) We propose to enhance the high-level attribute predictions by reasoning from the low-level and medium-level predictions, and achieve this reasoning in our end-to-end deep neural network.
3) We further boost the accuracy of the low-level and medium-level attribute predictions with respect to the guidance from the high-level attributes.
4) We demonstrate the effectiveness of our HR-Net by various evaluations on the PETA and RAP datasets, and perform an ablation study to verify our contributions. Experimental results indicate that our HR-Net achieves the state-of-the-art performance in the pedestrian attribute recognition task.

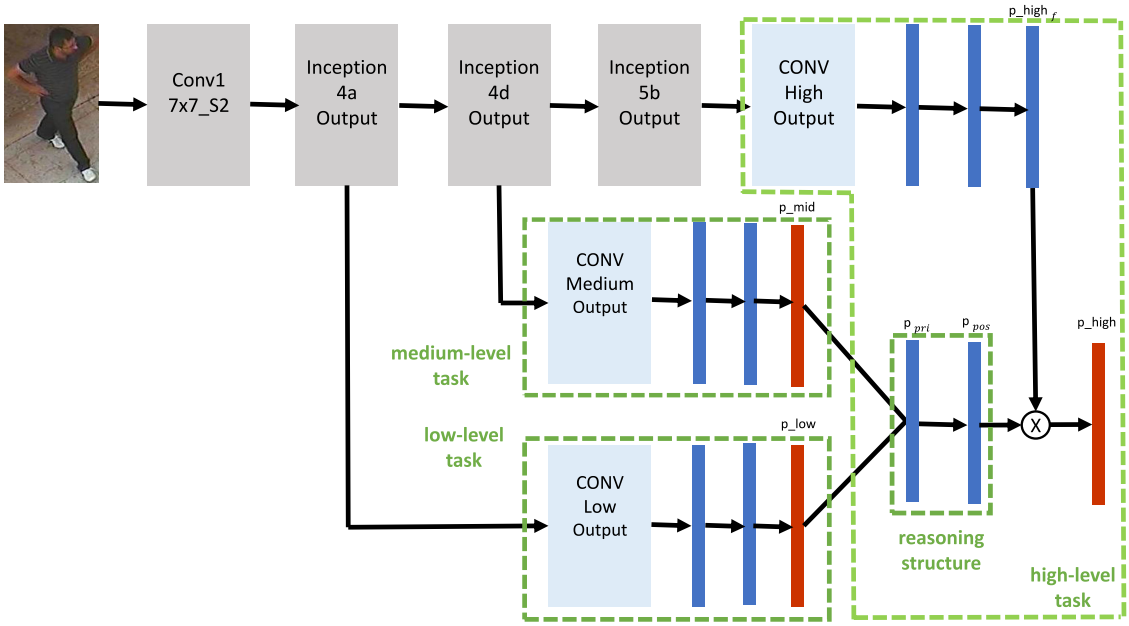


Fig. 3. The overall architecture of HR-Net. The input image goes through the backbone layers. Then, the three branches will predict the low-level, medium-level and high-level attributes, respectively. At last, the final prediction of the high-level attributes is obtained by reasoning from the priori information of the low-level and medium-level attributes.

The rest of the paper is organized as follows. Section II analyzes the existing attributes and introduces our attribute categorization. Section III provides the details of the proposed approach. Section IV presents the experimental results. Section V concludes the paper.

II. Observations

Features usually play an important role in pedestrian attribute recognition. The extracted features can be classified into two categories, handcrafted features and features extracted by deep neural networks. Traditionally, LBP [37] and HOG [38] features are commonly employed. LBP is an operator, which possesses significant advantages such as rotation invariance and intensity invariance, to describe the local textures of an image. HOG considers to describe the shape of the detected local object by gradients. By capturing the local shape information, HOG is invariant to both the geometrical and illuminant transformations. In deep learning, GoogLeNet [39] achieves very good classification performance with certain complexity and the amount of parameters. GoogLeNet can also be considered as a feature extractor and can automatically learn the representation of features from image.

We extract the LBP, HOG and GoogLeNet features from all the 19000 samples in the PETA database, and then employ the K-nearest neighbor classification algorithm to perform pedestrian attribute recognition. Table I presents the prediction accuracy of different attributes. As can be observed, LBP and HOG features tend to give higher recognition accuracy for some attributes, such as stripes, plaid, etc, while GoogLeNet features tend to outperform the LBP and HOG features for other attributes, such as gender, age, etc. Since the k-nearest neighbor algorithm simply uses the majority voting strategy for classification, the results reveal that different attributes may possess similarities in different aspects. This phenomenon is inevitable because deep networks usually extract more high-level semantic information and less low-level information as the number of layers increases [40]. When the attributes, such as stripes and plaid, are being predicted, the low-level information, such as edges and textures, which is less focused by the final output features of GoogLeNet, usually provides more meaningful information compared to the high-level features. On the contrary, when the attributes, such as gender and age, are desired to be recognized, the high-level semantic features extracted from a deep neural network are usually more effective. Besides of the low-level and high-level attributes, there also exists some other attributes like clothing and accessories, which require not only the semantic features but also certain local texture information. Since the k-nearest neighbor algorithm has no assumptions about the data distribution and is not sensitive to the outliers, the recognition accuracy of the rare categories is low when the samples are imbalanced. It explains why the recognition accuracy for some attributes, such as logo, are higher when the GoogLeNet features are employed. Based on the above results and observations, we propose to categorize the attributes into three levels, low-level attributes such as color and texture, medium-level attributes such as clothes, shoes, accessories and hair, and high-level attributes such as gender, age and style.

III. Methodology

According to our attribute categorization in Section II, we propose a hierarchical reasoning deep neural network for pedestrian attribute recognition, which is denoted as HR-Net. The overall design of our approach is shown in Fig. 3. Our network contains four components: 1) a shared feature learning module; 2) a low-level task module; 3) a medium-level task

TABLE I
RECOGNITION ACCURACY OF EACH ATTRIBUTES IN PETA

Attributes	LBP	HOG	GoogLeNet
accessoryHeadphone	87.42%	87.42%	**99.92%**
personalLess15	98.28%	98.28%	**98.63%**
personalLess30	91.99%	91.99%	**97.16%**
personalLess45	91.23%	91.23%	**96.98%**
personalLess60	89.75%	89.75%	**97.61%**
personalLarger60	94.21%	94.21%	**99.27%**
carryingBabyBuggy	95.76%	95.76%	**98.39%**
carryingBackpack	92.34%	92.34%	**95.55%**
hairBald	94.00%	94.00%	**99.35%**
footwearBoots	95.20%	95.20%	**98.07%**
lowerBodyCapri	91.64%	91.64%	**98.33%**
carryingOther	91.35%	91.35%	**94.97%**
carryingShoppingTro	89.99%	89.99%	**99.99%**
carryingUmbrella	90.60%	90.60%	**99.96%**
lowerBodyCasual	90.09%	90.09%	**97.74%**
upperBodyCasual	90.59%	90.59%	**98.13%**
personalFemale	90.66%	90.66%	**96.35%**
carryingFolder	91.43%	91.43%	**96.97%**
lowerBodyFormal	89.94%	89.94%	**97.73%**
upperBodyFormal	90.61%	90.61%	**98.10%**
accessoryHairBand	94.88%	94.88%	**95.83%**
accessoryHat	92.45%	92.45%	**96.39%**
lowerBodyHotPants	98.06%	98.06%	**99.24%**
upperBodyJacket	91.21%	91.21%	**97.57%**
lowerBodyJeans	90.86%	90.86%	**96.85%**
accessoryKerchief	96.68%	96.68%	**98.57%**
footwearLeatherShoes	91.27%	91.27%	**96.62%**
upperBodyLogo	88.09%	88.09%	**89.34%**
hairLong	91.27%	91.27%	**96.64%**
lowerBodyLongSkirt	92.68%	92.68%	**99.45%**
upperBodyLongSleeve	90.37%	90.37%	**97.03%**
lowerBodyPlaid	**99.97%**	**99.97%**	**99.97%**
lowerBodyThinStripes	**100.00%**	**100.00%**	99.99%
carryingLuggageCase	**98.35%**	**98.35%**	93.31%
personalMale	90.66%	90.66%	**96.35%**
carryingMessengerBag	91.65%	91.65%	**96.47%**
accessoryMuffler	94.75%	94.75%	**98.04%**
accessoryNothing	91.24%	91.24%	**95.36%**
carryingNothing	89.69%	89.69%	**95.35%**
upperBodyNoSleeve	94.46%	94.46%	**97.10%**
upperBodyPlaid	**95.98%**	**95.98%**	95.13%
carryingPlasticBags	91.64%	91.64%	**98.66%**
footwearSandals	86.29%	86.29%	**98.27%**
footwearShoes	91.07%	91.07%	**94.83%**
hairShort	91.36%	91.36%	**96.76%**
lowerBodyShorts	86.87%	86.87%	**97.09%**
upperBodyShortSleeve	89.83%	89.83%	**96.47%**
lowerBodyShortSkirt	89.81%	89.81%	**97.79%**
footwearSneakers	90.49%	90.49%	**94.74%**
footwearStocking	94.10%	94.10%	**98.08%**
upperBodyThinStripes	**93.60%**	**93.60%**	90.80%
upperBodySuit	87.76%	87.76%	**97.33%**
carryingSuitcase	83.68%	83.68%	**96.72%**
lowerBodySuits	87.22%	87.22%	**96.97%**
accessorySunglasses	89.14%	89.14%	**95.03%**
upperBodySweater	95.48%	95.48%	**96.97%**
upperBodyThickStripes	**99.85%**	**99.85%**	92.79%
lowerBodyTrousers	90.81%	90.81%	**96.20%**
upperBodyTshirt	88.71%	88.71%	**97.45%**
upperBodyOther	91.11%	91.11%	**96.08%**
upperBodyVNeck	**90.47%**	**90.47%**	88.45%

module; 4) a high-level task module. The shared feature learning module learns informative representations for the latter stages. Different task modules extract low-level, medium-level, and high-level features, and thus predict the desired attributes in the corresponding levels. As illustrated in Fig. 4, the high-level task incorporates the representations from the high-level features and the priori information obtained from the low-level

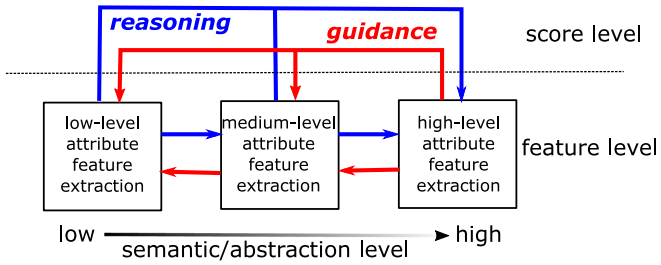


Fig. 4. Illustration of the reasoning and guidance among different levels. The blue lines represent the reasoning process which is achieved by inferences. The red lines represent the guidance process which is achieved by the back-propagations. In the reasoning process, the high-level attributes are predicted based on the priori information from the low-level and medium-level attributes. In the guidance process, the prediction errors of the high-level task is passed to the low-level and medium-level tasks to guide their recognition.

and medium-level tasks to give an overall prediction of the high-level attributes. Besides, the low-level and medium-level attribute predictions are guided by the high-level ones.

A. Shared Feature Learning

The GoogLeNet inception architecture is utilized as our progressive feature extractor, and different levels of tasks share this backbone network. The modules from the GoogLeNet inception architecture perform the convolution operation to the input feature maps to produce an output feature map defined by

$$f^j = \sum_i c^j * f^i + b^j, \tag{1}$$

where f^j and f^i are the j^{th} output and i^{th} input feature maps, respectively. c^j stands for the j^{th} convolution kernel and $*$ represents the convolution operation between the convolution kernels and the input feature map. b^j denotes the j^{th} bias corresponding to the j^{th} convolution kernel.

After the convolution operation, batch normalization (BN) is applied to all the output feature maps to accelerate the training process and reduce the internal covariate shift [41]. Rectified linear units (ReLUs) [42] are employed as the nonlinear activation functions.

This backbone network sequentially extracts features for the three tasks which will be performed in the three branches accordingly. Since the backbone network will be affected by these tasks in the training process via back propagations, it will learn to extract the task-independent features, i.e., it can extract the common information, which is useful to all the tasks, by exploiting the intrinsic similarities among these tasks.

B. Hierarchical Attribute Recognition

In Section II, we have concluded that the pedestrian attributes possess various characteristics and can be classified into hierarchical categories. The low-level attributes mainly focus on colors and textures. The recognition of these attributes can be improved when the local features are exploited. The high-level

attributes focus on the semantics. The recognition of these attributes requires strong semantic features. The medium-level attributes demand features containing both the semantic and local information.

According to these intuitions, we categorize the existing 35 attributes in the PETA database into three classes according to the level of abstraction:

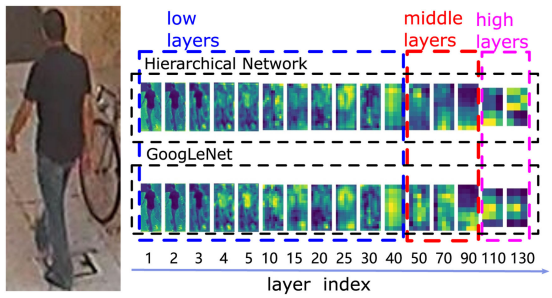
1) *Low-level attributes:* upperBodyLogo, lowerBodyThinStripes, upperBodyThinStripes, upperBodyThickStripes, lowerBodyPlaid, upperBodyPlaid;
2) *Medium-level attributes:* carryingOther, carryingBackpack, accessoryHat, upperBodyJacket, lowerBodyJeans, footwearLeatherShoes, accessoryMuffler, accessory-Nothing, carryingMessengerBag, lowerBodyLongSkirt, carryingNothing, carryingPlasticBags, footwearSandals, footwearShoes, lowerBodyShorts, upperBodyTshirt, lowerBodyShortSkirt, footwearSneakers, upperBody-ShortSleeve, lowerBodyTrousers, upperBodyOther, accessorySunglasses, upperBodyVNeck, hairLong;
3) *High-level attributes:* personalLess30, personalLess45, personalLess60, personalLarger60, lowerBodyCasual, upperBodyCasual, lowerBodyFormal, upperBodyFormal, personalMale.

We also categorize the existing 51 attributes in the RAP database into three classes with the same principle:

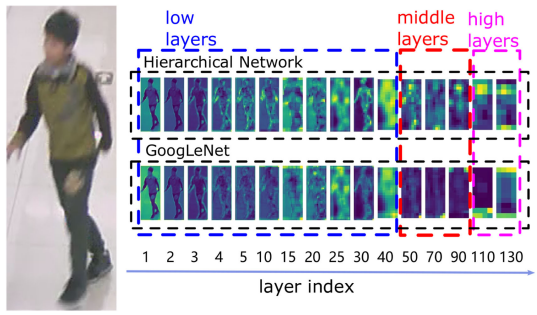
1) *Low-level attributes:* hs-BlackHair;
2) *Medium-level attributes:* BodyFat, BodyNormal, BodyThin, Customer, Clerk, hs-BaldHead, hs-LongHair, hs-Hat, hs-Glasses, hs-Muffler, ub-Shirt, ub-Sweater, ub-Vest, ub-TShirt, ub-Cotton, ub-Jacket, ub-SuitUp, ub-Tight, ub-ShortSleeve, lb-LongTrousers, lb-Skirt, lb-ShortSkirt, lb-Dress, lb-Jeans, lb-TightTrousers, shoes-Leather, shoes-Sport, shoes-Boots, shoes-Cloth, shoes-Casual, attach-Backpack, attach-SingleShoulderBag, attach-HandBag, attach-Box, attach-PlasticBag, attach-PaperBag, attach-HandTrunk, attach-Other, action-Calling, action-Talking, action-Gathering, action-Holding, action-Pusing, action-Pulling, action-CarrybyArm, action-CarrybyHand;
3) *High-level attributes:* Female, AgeLess16, Age17–30, Age31–45.

Then, the attribute predictions for each level can be considered as an individual attribute learning task. Based on the observations in Section II, since low-level attribute recognition requires more texture information than semantic information and vice versa, the selected layers, *Inception4a/output*, *Inception4d/output* and *Inception5b/output* (from shallow to deep), and three convolution layers are utilized to perform the attribute predictions (from the low-level to high-level attributes).

The hierarchical structure is adopted based on the following considerations. When recognizing the attributes in different levels, the network will generate features with different characteristics to provide specific information for each level. Besides, the low-level and medium-level attribute recognition will enable the backbone network to learn the filters which can extract more desired information. As can be observed from the visualization



(a) Feature maps of a sample in PETA.



(b) Feature maps of a sample in RAP.

Fig. 5. Visualization of the feature maps from different convolutional layers in the network. For each sample, the first row is the results of our hierarchical network, and the second row is the results of GoogLeNet.

in Fig. 5, our hierarchical network will extract more edge/texture information for the sake of the low-level attribute predictions, and extract the medium-level features, which contain balanced semantic and detailed information, from the middle layers. It is worth noting that our network gives the feature maps with more concentrated neural activations, compared to the sparse activations from the baseline method. These concentrated activations may correspond to different body parts.

C. Attribute Reasoning

Since our hierarchical network will generate more concentrated high-level features, we can utilize the predicted low-level and medium-level attributes to compensate this deficiency. According to [43], the semantic features can be characterized numerically by considering lower level image features, i.e., the predicted low-level and medium-level attributes can help to recognize the high-level attributes. Human cognition also reveals that lower level attributes usually give a certain boost to higher level attribute predictions. For example, when a person is wearing short skirt and earrings, the probability that this person is a woman is higher than the probability when these attributes do not exhibit. Therefore, we can utilize the dependencies among the attributes at different semantic levels to achieve the reasoning and guidance to improve the attribute recognition accuracy.

According to Markov Logic Network (MLN) [44], a simple state matrix operation can complete the transition from one state to another. In MLN, the transition probability between states can be expressed by the state transition matrix. It clearly describes the vast Markov Networks (MNs) and the flexibility to incorporate the modular knowledge domains into the networks. Inspired

by this intuition, the reasoning among the attributes can also be achieved by simple operations. To introduce the reasoning ability from MLN without inducing additional computational overheads, a fully-connected (*FC*) layer, which connects the attribute prediction layers, is designed to capture the correlations between all the low-level and medium-level attributes and the high-level attributes. The weights of the *FC* layer are equivalent to the elements in a state transition matrix in MLN, and reflect the reasoning process from the low-level and medium-level attributes to the high-level attributes.

The input to the reasoning layer is the predicted probability from the low-level and medium-level attributes, while the output is the high-level attribute probability. In the training process, the inference achieves the reasoning from the low-level and medium-level attribute predictions to the high-level attribute ones, i.e., the high-level prediction probability is calculated based on the prior information obtained from the low-level and medium-level attributes. Meanwhile, the back-propagations from the high-level attribute predictions serve as a guidance to the low-level and medium-level attribute predictions, i.e., their weight refinements, by passing the prediction error of the high-level task to the low-level and medium-level tasks.

After concatenating the low-level and medium-level attribute prediction scores, p^i_{pri}, where i represents the i^{th} sample, is obtained. Then, the *FC* layer converts the prior vector to the intermediate result of the high-level attribute prediction, named as $p\_high^i_{pos}$. The final score of the high-level attribute is defined by

$$\begin{aligned} p\_high^{i,j} &= p\_high_f^{i,j} * p\_high_{pos}^{i,j}, \\ p\_high_f^{i,j} &= \sigma(W_f^j \cdot f^i), \\ p\_high_{pos}^{i,j} &= W_r^j \cdot p_{pri}^i, \end{aligned} \tag{2}$$

where $p\_high^{i,j}$ stands for the final predicted score vector of the high-level attributes, $p\_high_f^{i,j}$ represents the score vector of the high-level attributes predicted by network features f^i, $\sigma(\cdot)$ is the sigmoid activation function, W_f^j and W_r^j are the parameters corresponding to the j^{th} attribute of the *FC* layers for the high-level attribute prediction and the reasoning, respectively.

D. Implementation Details

The specific architecture of our HR-Net is shown in Fig. 6. The standard GoogLeNet inception architecture possesses two auxiliary loss layers which intend to amplify the gradients and encourage discriminations at the primary stages of the network. In our design, these auxiliary loss layers are replaced by the cross-entropy loss function.

Typically, the logarithmic loss function, which is also known as the binary cross-entropy loss function, is exploited to estimate the degree of inconsistency between the predicted and true values as

$$\begin{aligned} \mathcal{L}_b(y, \hat{y_p}) = & -\sum_{c=1}^{C} y^c \cdot \log(\sigma(\hat{y_p}^c)) \\ & + (1 - y^c) \cdot \log(1 - \sigma(\hat{y_p}^c)), \end{aligned} \tag{3}$$

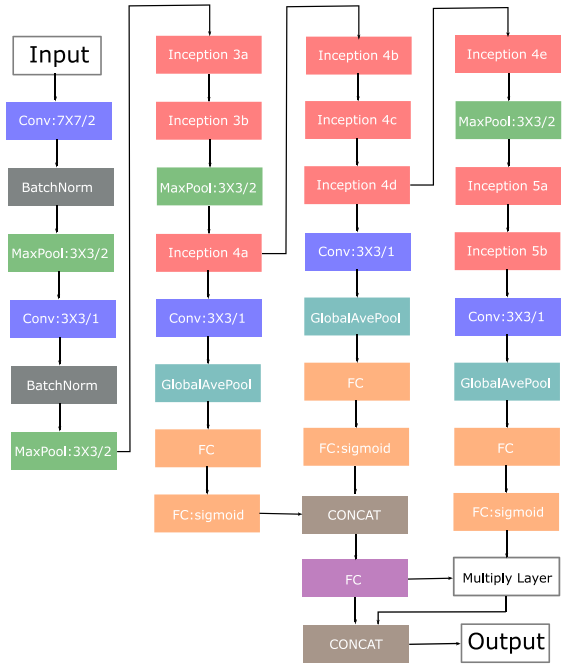


Fig. 6. System diagram of HR-Net.

where y and $\hat{y_p}$ represent the ground-truth and predicted labels for the attribute c, respectively.

However, in both the PETA and RAP databases, the distributions of the positive and negative labels in most attribute classes are usually imbalanced. Many attributes, such as wearing a muffler or not, are mostly labelled as negative in the training data. Directly employing the loss function in Eq. (4) may jeopardize the predictions of the rare attributes because of the class imbalance problem. To alleviate this problem, a weighted cross entropy objective function is introduced and defined as follows.

$$\begin{aligned} \mathcal{L}_b(y, \hat{y_p}) = & -\sum_{c=1}^{C} \frac{1}{2w_c} \cdot y^c \cdot \log(\sigma(\hat{y_p}^c)) \\ & + \frac{1}{2(1 - w_c)} \cdot (1 - y^c) \cdot \log(1 - \sigma(\hat{y_p}^c)), \end{aligned} \tag{4}$$

where w_c is a weight vector indicating the proportion of positive labels for the attribute c in the training set.

The low-level attribute predictor takes the 512 feature maps from the *inception 4a* layer as input. It comprises of a 3x3 conv block, a global max pooling operation and two *FC* layers, where the last *FC* layer generates a 4-dimensional feature which is transmitted to the sigmoid function to predict the low-level attributes. Similarly, the medium-level attribute predictor takes the 528 feature maps from the *inception 4d* layer as input. It comprises of a 3x3 conv block, a global max pooling opration and two *FC* layers, where the last *FC* layer produces a 43-dimensional feature which is transmitted to the sigmoid function to predict the medium-level attributes. The high-level predictor without reasoning takes the 1024 feature maps from the *inception 5b* layer as input. It comprises of a 3x3 conv block, a global max pooling oparetion and two *FC* layers, where the last *FC* layer outputs a 14-dimensional feature which is transmitted to the sigmoid

function to predict the high-level attributes. The final high-level attribute predictions (with reasoning) is computed via Eq. (2).

To avoid the overfitting problem and achieve a robust recognition result, image augmentation is applied during the training process. For batch creation, the input images are resized to 160 $\times$ 75, and random rotations, random resizings and random horizontal flippings are applied. The weights in the network are initialized by a pre-trained ImageNet model and fine-tuned with an initial learning rate of 0.001, the SGD optimizer and a batch size of 256.

IV. Experimental Results

A. Datasets

Our HR-Net is evaluated on two large-scale pedestrian attribute datasets, the PETA dataset [19] and the RAP dataset [8]. The PETA dataset is a collection of several surveillance datasets and consists of 19,000 cropped images. Each image is annotated with 61 binary and 5 multi-value attributes. According to the established protocol in [19], [20], [24], [27], [33], [34], [45], 35 attributes are selected by calculating and selecting the ratio of positive labels being higher than 5%. The PETA dataset is sampled to obtain 9,500 training images, 1,900 validation images and 7,600 test images. The RAP dataset is a larger pedestrian attribute dataset, including 41,585 images. Each image is annotated with 72 binary-valued attributes. Similar to [8], 51 attributes are used for evaluation. The RAP dataset is sampled to obtain 26,750 training images, 6,698 validation images and 8,137 test images.

B. Baseline Methods

We compare our HR-Net model with the state-of-the-art pedestrian attribute recognition techniques, including ACN [20], DeepMAR [27], SRN [33], VeSPA [45], WPAL [24] and VAA-ResNet [34]. The experimental results of these baseline methods are all provided by the authors or produced following their original settings and training protocols. Note that GoogLeNet is also employed as a baseline method in our evaluation to better demonstrate the gain of the proposed architecture.

C. Evaluation Protocols

In our evaluations, two metrics are employed. For the label-based evaluation, the mean accuracy (mA) is computed as the mean of the accuracy among the positive classification results and the accuracy among the negative classification results. This metric alleviate the class imbalance problem by equally penalizing the inaccurate predictions of the labels with different appearance frequencies. mA is formulated as

$$mA = \frac{1}{2L}\sum_{i=1}^{L}\left(\frac{|TP_i|}{|P_i|} + \frac{|TN_i|}{|N_i|}\right), \tag{5}$$

where L is the number of attributes, $|TP_i|$ and $|TN_i|$ respectively represent the number of correctly predicted positive and negative samples of the i^{th} attribute, and $|P_i|$ and $|N_i|$ are the number of ground-truth positive and negative samples of the i^{th} attribute, respectively.

Unfortunately, this metric does not account for the attribute relationships, e.g., the consistencies among the attributes for each person. To evaluate the attribute consistencies for each person, the popular example-based metrics, accuracy, precision, recall and F1 score, which are averaged across all the samples, are employed. The example-based evaluation criteria is defined as

$$ACC_{exam} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}, \tag{6}$$

$$Prec_{exam} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|}, \tag{7}$$

$$Rec_{exam} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i|}, \tag{8}$$

$$F1 = \frac{2 \cdot Prec_{exam} \cdot Rec_{exam}}{Prec_{exam} + Rec_{exam}}, \tag{9}$$

where N is the number of samples, Y_i is the set of ground-truth positive attribute labels of the i^{th} sample, $\hat{Y}_i$ is the set of predicted positive attribute labels of the i^{th} sample, and the $|\cdot|$ denotes the set cardinality.

D. Ablation Study

To explicitly demonstrate the effectiveness and contributions of our designed architecture, we perform a component-wise ablation study on the PETA dataset.

Hierarchical Structure: In this test, the backbone network is employed as the baseline. We compare the results of two networks in Table II: (i) the BACKBONE(ALL) network which predicts all the attributes without the hierarchical prediction structure and the reasoning and guidance design; (ii) the BACKBONE+HIER(ALL) network which add the hierarchical prediction structure to the backbone network. Regardless of the dependencies among the attributes, hierarchically performing the attribute recognitions, according to the abstraction levels at different stages of the network, can achieve notable improvements. The results in Table II have demonstrated that the improvements of the low-level and medium-level attribute predictions are particularly obvious, which mainly benefit from the fact that the feature maps obtained from the shallow layers contain more detailed appearances. As can be observed in Fig. 5, BACKBONE+HIER(ALL) can capture more low-level information at the shallow layers, which promotes both the recognition and learning process.

Reasoning Structure: To verify the effectiveness of our attribute reasoning structure, the performances of the high-level attribute predictions from the following network architectures are compared in Table III: (i) the backbone network, which is trained to predict the high-level attribute only, is denoted as BACKBONE-H(HIGH); (ii) the ordinary BACKBONE(HIGH) network, which is trained to predict all the attributes; (iii) the backbone network with the hierarchical structure only, which

TABLE II
MEAN ACCURACY OF THE BACKBONE WITH/WITHOUT THE HIERARCHICAL PREDICTION STRUCTURE ON PETA

Attributes		BACKBONE(ALL)	BACKBONE+HIER(ALL)
Low-level		63.26	**80.64**
Medium-level	Carrying	75.64	**79.83**
	Accessory	78.33	**91.18**
	Uppper-dody	80.76	**90.87**
	Lower-body	85.01	**88.88**
	Foot-wear	80.14	**89.08**
High-level	Age	85.72	**86.05**
	Gender	85.62	**86.42**
	Cloth-style	75.43	**87.10**
Average		79.67	**91.18**

TABLE III
PERFORMANCES OF THE PROPOSED AND INTERMEDIATE FOUR MODELS ON PETA

Methods	mA	Acc	Prec	Recall	F1
BACKBONE-H(HIGH)	84.03	87.84	92.24	91.44	91.84
BACKBONE(HIGH)	81.13	84.70	89.27	90.32	95.81
BACKBONE+HIER(HIGH)	90.70	93.76	95.92	95.69	95.81
HR-Net(HIGH)	**95.81**	**95.12**	**96.87**	**96.40**	**96.64**

is denoted as BACKBONE+HIER(HIGH), is trained to predict all the attributes; (iv) the complete network architecture of our proposed HR-Net(HIGH) which is trained to predict all the attributes. For better comparisons, we only measure the prediction accuracies of the high-level attributes. Comparing BACKBONE-H(HIGH) with BACKBONE(HIGH), it gives superior performances because BACKBONE(HIGH) is trained to predict all the attributes, which certainly affects the learning of high-level attribute predictions. However, if the attributes are predicted hierarchically by BACKBONE+HIER(HIGH), the low-level and medium-level attributes tend to induce positive influences to the high-level attribute predictions, according to the results of BACKBONE+HIER(HIGH). If the attribute reasoning structure, which emphasizes the relationships among the attributes at different semantic levels, is also added, the proposed method HR-Net(HIGH) can further boost the prediction accuracy. Therefore, by properly exploiting the dependencies among the attributes, the attribute recognition can be improved.

E. Performance Evaluations

Table IV shows the results on PETA and RAP. According to the experimental results in Table IV, the proposed approach outperforms the baseline methods for all the metrics on PETA. Particularly, our HR-Net has improved the mA score by at least 10% compared to GoogLeNet, which demonstrates the importance of the hierarchical reasoning structure. Besides, our approach yields significant improvements when the F1 score is employed as the measurement on PETA. It indicates a good precision-recall tradeoff of our approach. Due to the large number of complex samples in the RAP dataset, such as incomplete or occluded pedestrian images, the example-based metrics are not good as the results on PETA.

We also compare our HR-Net with MRFr2 [19], DeepSAR [27], DeepMar [27] and GoogLeNet in terms of the recognition accuracies for specific attributes on PETA. The results are given in Fig. 7 and Table V. As can be observed in Fig. 7, our HR-Net can correctly recognize the attributes, especially the high-level attributes, which are missing in the predictions of GoogLeNet. According to Table V, where the prediction accuracies of specific attributes are given, our method can achieve a better performance for these high-level attributes. These results can support our intuitions that the proposed method can effectively learn and exploit the relationships among the attributes in the attribute reasoning structure. It is worth noting that DeepMar gives better performances on certain attributes such as personalLess60, accessoryHat and accessoryMuffler, which may be induced by the small number of positive corresponding samples in the dataset. Thus, the network cannot be trained to learn effective features for attribute recognition.

We also compare our HR-Net with ELF [8], FC7-mm [8], FC6-mm [8], ACN [20], DeepMar [27], WPAL [24] and GoogLeNet in terms of the recognition accuracies for specific attributes on RAP. The results are given in Table VI. Our HR-Net gives better performances for some attributes such as long hair, sweater, short skirt and handbag compared to the baseline methods.

V. DISCUSSION

A. Parameter Analysis

To better understand our designed architecture, we analyze the number of parameters of the network and the parameters of the reasonning structure, accordingly.

Number of Parameters: We compare the total number of the parameters in HR-Net with some baseline approaches. As shown in Table VII, HR-Net can give superior performances with the fewest parameter increases compared to the baseline methods.

Weights of the FC layer in the Reasoning Structure: In Fig. 8, the weights of the FC layer in the reasoning structure are visualized. Large weights indicate that the correlations among the attributes are large. For example, the model trained on the PETA dataset indicates that the attribute *personalFemale* relies on the attributes such as *hairLong*, *accessoryKerchief* and *lowerBodyShortSkirt*. Meanwhile, the model trained on the RAP dataset shows that *hs-LongHair* has contributed significantly in

TABLE IV
COMPARISONS BETWEEN THE PROPOSED HR-Net AND THE BASELINE APPROACHES

Methods	PETA					RAP				
	mA	Acc	Prec	Recall	F1	mA	Acc	Prec	Recall	F1
ACN [20]	81.15	73.66	84.06	81.26	82.64	69.66	62.61	**80.12**	72.26	75.98
DeepMar [27]	82.89	75.07	83.68	83.14	83.41	73.79	62.02	74.92	76.21	75.56
SRN($/L_b$) [33]	80.55	74.24	84.04	82.48	83.25	-	-	-	-	-
SRN($/L_w$) [33]	82.36	75.69	85.25	84.59	84.92	-	-	-	-	-
VeSPA [45]	83.45	77.73	86.18	84.81	85.49	77.70	**67.35**	79.51	**79.67**	**79.59**
WPAL-GMP [24]	**85.50**	76.98	84.07	85.78	84.90	**81.25**	50.30	57.17	78.39	66.12
WPAL-FSPP [24]	84.16	74.62	82.66	85.16	83.40	79.48	53.30	60.82	78.80	68.65
VAA-ResNet [34]	84.59	**78.56**	**86.79**	**86.12**	**86.48**	-	-	-	-	-
GoogLeNet baseline	79.67	75.63	85.32	82.65	83.97	69.11	58.82	74.77	70.96	72.80
HR-Net-GoogleNet	**94.42**	**94.68**	**96.47**	**96.11**	**96.29**	**81.10**	45.70	51.48	78.56	62.20

Fig. 7. Examples of the pedestrian attribute recognition results on the PETA dataset. Correct attribute predictions are marked in bold. The attributes marked in green color are recognized by both the GoogLeNet baseline and our HR-Net. The attributes marked in red are recognized by our model only.

the prediction of the attribute *gender*. These observations are consistent with the prior knowledge of human cognitions and verify the attribute reasoning structure proposed in Section III-C.

B. Failure Cases

When will the proposed method fail, i.e., Which sample will be wrongly predicted? What are the characteristics of these failure cases? Here, we analyze the failures caused by data errors and inappropriate attribute classifications.

Data Errors: The limitations of typical pedestrian recognition methods, which cannot be ignored, is that that the input images are usually resized to a fixed resolution (e.g., 160 $\times$ 75) to be processed in the deep nerual network, which may induce certain information distortions. For example, the resized image may contain color artifacts, which tend to affect the attribute recognitions, induced by the downsamplings. Besides, the input data, including the image and label data, may contain noises. The images with poor recognition accuracies are shown in Fig. 9. As shown in Fig. 9, there may be multiple pedestrian targets in the same image or the pedestrian in the image is incomplete, which make it hard even for human eye to identify the correct attributes. Besides the errors produced by the images, the provided annotations contain a third unspecified class that is unrecognizable, which is used as negative during training, and the third unspecified class may cause confusion during training.

Errors by Modeling: Our attribute categorization, which is performed based on the human cognition, may not be optimal for the pedestrian attribute recognition task. For example, *BodyFat*, *BodyNormal* and *BodyThin* can be regarded as middle-level attributes or high-level attributes, which will induce different prediction results. Note that inappropriate attribute categorization may jeopardize network from learning the accurate representations.

C. Baseline Analysis

Comparing with WPAL [24]: Our HR-Net and WPAL networks both employ the GoogLeNet as our backbone network, so the network architecture is similar. Although some special kinds

TABLE V
RECOGNITION ACCURACIES ON PETA

Attributes	MRFr2 [19]	DeepSAR [27]	DeepMar [27]	GoogLeNet	**HR-Net**
upperBodyLogo	52.70	76.10	68.40	57.18	**87.23**
upperBodyThinStripes	51.90	72.80	66.50	71.79	**92.09**
carryingBackpack	70.50	78.80	82.60	79.52	**90.84**
accessoryHat	90.40	89.20	**91.80**	79.58	90.20
carryingOther	73.00	73.00	77.30	82.47	**89.45**
upperBodyJacket	72.20	77.50	79.20	81.92	**87.61**
lowerBodyJeans	81.00	80.20	85.70	91.45	**93.94**
footwearLeatherShoes	87.20	84.20	87.30	83.46	**92.60**
hairLong	80.10	83.20	88.90	81.94	**90.85**
carryingMessengerBag	78.30	77.40	82.00	85.00	**90.70**
accessoryMuffler	93.70	94.40	**96.10**	85.20	93.20
accessoryNothing	82.70	81.50	85.80	80.25	**90.93**
carryingNothing	76.50	78.80	83.10	81.60	**88.34**
carryingPlasticBag	81.30	82.90	87.00	73.51	**91.71**
footwearShoes	78.40	75.80	80.00	81.50	**90.23**
footwearSneakers	75.00	77.30	78.70	84.22	**89.74**
upperBodyShortSleeve	75.80	84.60	87.50	88.59	**88.77**
lowerBodyShortSkirt	69.90	83.20	82.20	85.43	**88.92**
accessorySunglasses	53.50	79.10	69.90	77.13	**87.83**
lowerBodyTrousers	82.20	78.40	84.30	86.72	**91.77**
upperBodyTshirt	87.30	83.40	86.10	87.19	**88.28**
upperBodyOther	87.30	83.40	86.10	83.64	**91.18**
upperBodyVNeck	53.30	75.40	69.80	74.98	**83.80**
lowerBodyCasual	78.20	81.60	84.90	79.93	**91.84**
upperBodyCasual	78.10	81.10	84.40	80.76	**91.48**
lowerBodyFormal	79.00	81.90	85.20	80.31	**91.39**
upperBodyFormal	78.70	81.60	85.10	81.64	**91.23**
personalLess30	86.80	82.90	85.80	88.05	**92.04**
personalLess45	83.10	79.40	81.80	86.12	**91.04**
personalLess60	80.10	83.30	**86.30**	79.26	84.62
personalLarger60	93.80	92.00	94.80	87.72	**90.37**
personFemale	-	-	-	85.63	**91.02**
personMale	86.50	85.10	89.90	85.61	**91.34**
Average	75.60	81.30	82.60	79.67	**94.42**

TABLE VI
RECOGNITION ACCURACIES ON RAP

Attributes	ELF [8]	FC7-mm [8]	FC6-mm [8]	ACN [20]	DeepMar [27]	WPAL [24]	GoogLeNet	**HR-Net**
Black Hair	-	-	-	-	-	-	64.33	**81.63**
Bald Head	71.90	69.69	72.85	60.78	62.39	**84.29**	63.86	81.80
Long Hair	78.00	79.99	81.50	88.66	**90.27**	52.52	85.45	88.15
Hat	69.50	73.31	72.92	57.33	62.35	84.55	59.37	**84.93**
Sweater	59.90	62.20	63.04	58.58	66.73	72.71	62.90	**76.81**
Tights	65.30	68.08	69.71	61.64	66.49	**76.45**	64.73	71.44
Skirt	-	-	-	-	-	-	79.60	**90.16**
Short Skirt	76.20	78.03	78.63	70.80	76.88	88.79	66.94	**89.75**
Boots	-	-	-	-	-	-	83.81	**92.86**
Single-Shoulder Bag	66.70	71.33	72.66	64.78	73.58	**81.28**	68.56	75.69
HandBag	66.40	71.85	72.30	63.13	72.46	85.27	64.34	**86.00**
Box(Attachment)	67.80	70.63	71.46	64.95	72.07	**78.65**	65.97	76.81
Plastic Bag	60.90	70.51	70.64	58.30	66.99	**82.39**	58.84	80.15
Paper Bag	63.90	66.71	68.51	53.84	60.47	**78.08**	58.18	74.74
Calling	65.90	68.74	70.15	69.54	76.88	**89.17**	62.80	80.57
AgeLess16	-	-	-	-	-	-	70.67	**85.22**
Clerk	-	-	-	-	-	-	87.76	**91.04**
Average	69.94	72.28	73.32	69.66	73.79	79.48	69.11	**81.10**

TABLE VII
COMPARISONS OF THE NUMBER OF MODEL PARAMETERS

Methods	Total Number	Increased Number
GoogLeNet	7.1×10^6	0
DeepMar [27]	138.0×10^6	130.9×10^6
VeSPA [45]	27.1×10^6	20.0×10^6
WPAL-FSPP [24]	30.8×10^6	23.7×10^6
VAA-ResNet [34]	60.4×10^6	53.3×10^6
HR-Net	21.8×10^6	14.7×10^6

of mid-level semantic features may be obtained from some layers at different levels in WPAL-FSPP and WPAL-GMP, they both ignore the hierarchical nature of the attributes and predict all the attributes with the same features. Unfortunately, they cannot learn the dependencies among the attributes.

Experiments with ResNet-50 [46]: There are both some shortcut connections in our HR-Net and ResNet-50. The shortcut connections in ResNet-50 are feature-based which act on the feature maps. On the other hand, our HR-Net learns

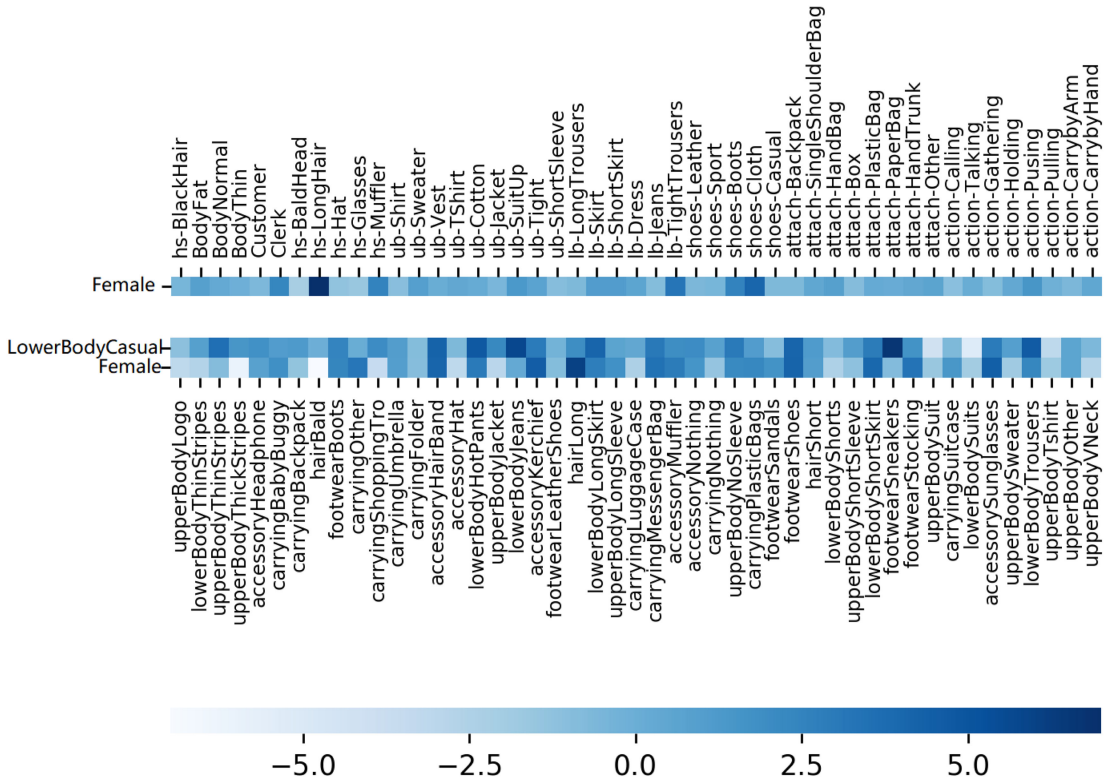


Fig. 8. Weights of the FC layer in the reasoning structure on the RAP (the top) and PETA (the bottom) datasets, which reveals the dependencies of high-level attributes on the low-level and middle-level attributes.

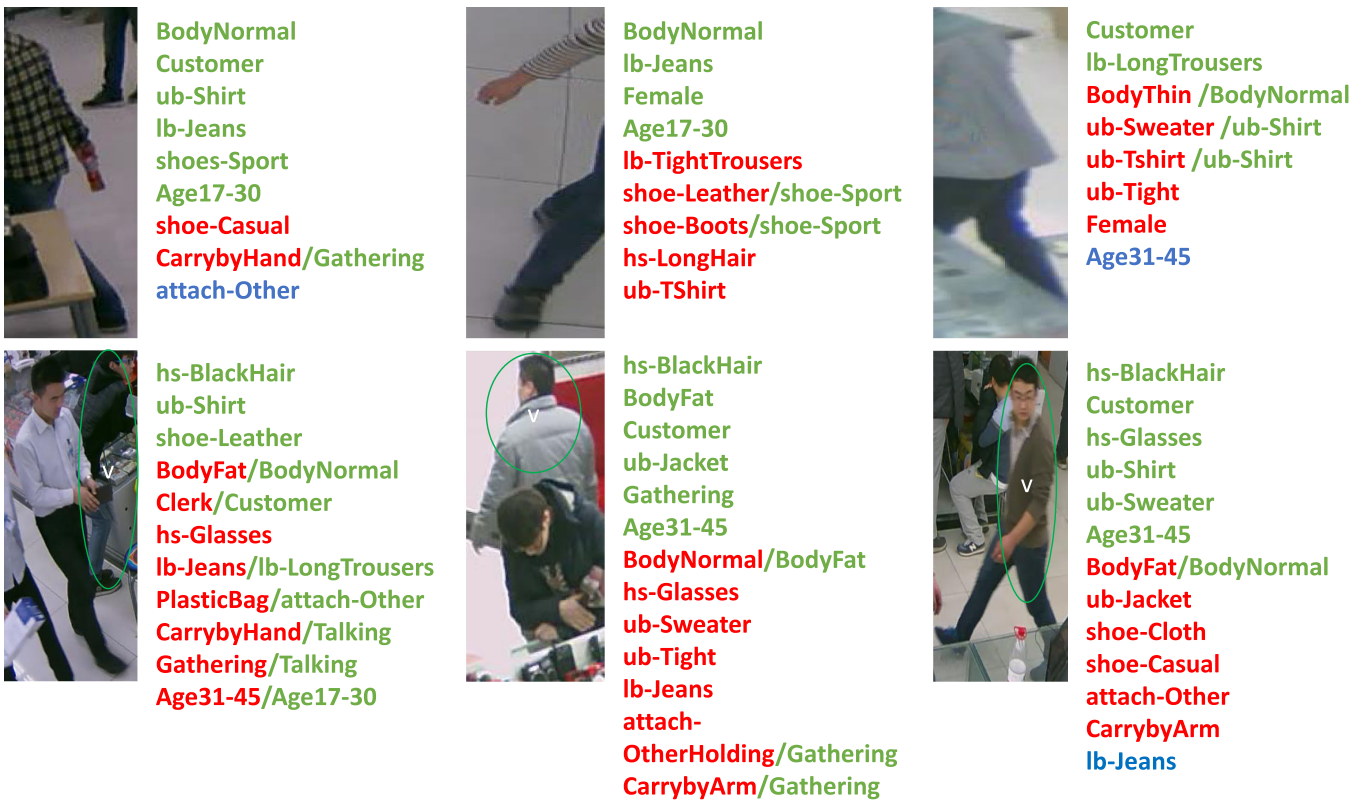


Fig. 9. The hard samples with poor attribute recognition accuracies in RAP dataset. The correct attribute predictions are marked in green, the wrong ones are marked in red which is followed by the corresponding groud truth label marked in green, and the missing ones are marked in blue.

TABLE VIII
COMPARISONS BETWEEN THE PROPOSED HR-NET-RESNET AND THE RESNET-50 BASELINE APPROACH

Methods	PETA					RAP				
	mA	Acc	Prec	Recall	F1	mA	Acc	Prec	Recall	F1
ResNet baseline	81.89	67.00	69.12	**94.22**	79.74	75.53	36.27	40.69	**75.64**	52.91
HR-Net-ResNet	**91.46**	**79.85**	**84.91**	90.43	**87.59**	**82.18**	**51.32**	**67.74**	65.94	**66.84**

the dependencies of high-level attributes on the low-level and medium-level attributes via a FC layer, thus the shortcut connections in our method are designed based on our specific task. We have also conducted extra experiments by employing ResNet-50 as the backbone network to validate the proposed hierarchical reasoning architecture. The results are shown in VIII and it obviously reveals that our hierarchical reasoning structure is very flexible and architecture-agnostic. It can be easily adapted to any kind of backbone networks such as GoogleNet and ResNet.

D. Extension

Since the proposed heuristic attribute categorization is based on the human cognitive process, it incorporates a local-to-global information integration mechanism in the human cognition. In general, the primary and secondary visual cortical neurons, which possess smaller spatial and temporal receptive fields, receive the local visual information. On the contrary, the neurons in higher visual brain regions generally have larger perceptive receptive fields and receive the visual information within a larger fields. Then, with the observations from Table 1, we manually classify the pedestrian attributes into three categories at three semantic levels. In the future, we will develop an automatic attribute categorization approach to jointly optimize the attribute categorizations and predictions.

VI. CONCLUSION

In this work, we tackle the pedestrian attribute recognition problem by proposing a hierarchical reasoning network. Instead of directly predicting the attributes, we firstly categorize the attributes into three categories according to their semantic levels. The low-level attributes focus on colors and textures. The medium-level attributes focus on hair, clothing, accessories and etc. The high-level attributes focus on gender, age and etc. Our HR-Net hierarchically extracts the low-level, medium-level and high-level features for the attribute predictions at the corresponding levels. Besides, HR-Net also achieves the inferential predictions from the low-level to high-level attributes by the guidance from the high-level predictions to the low-level ones. Extensive experiments have demonstrated the superiority of our method compared to the state-of-the-art baseline methods on the PETA and RAP datasets, while an ablation study has verified the effectiveness of our contributions.

REFERENCES

[1] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proc. BMVC*, 2012, pp. 8–8.

[2] S. Karanam *et al.*, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2019.

[3] A. Wu, W. S. Zheng, and J. H. Lai, "Robust depth-based person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, Jun. 2017.

[4] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.

[5] D. A. Vaqueroand *et al.*, "Attribute-based people search in surveillance environments," in *Proc. IEEE Workshop Appl. Comput. Vision*, 2009, pp. 1–8.

[6] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 65–81, Jan. 2007.

[7] B. Park, K. Lee, and S. U. Lee, "Face recognition using face-ARG matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1982–1988, Dec. 2005.

[8] 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.

[9] B. Gong, J. Liu, X. Wang, and X. Tang, "Learning semantic signatures for 3D object retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 369–377, Feb. 2013.

[10] Z. Chen, Z. Xu, Y. Zhang, and X. Gu, "Query-free clothing retrieval via implicit relevance feedback," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2126–2137, Aug. 2018.

[11] Y. Bai *et al.*, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[12] H. Mo, L. Liu, W. Zhu, S. Yin, and S. Wei, "Face alignment with expression-and pose-based adaptive initialization," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 943–956, Apr. 2019.

[13] Y. Chen, J. Wang, Y. Bai, G. Castañón, and V. Saligrama, "Probabilistic semantic retrieval for surveillance videos with activity graphs," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 704–716, Mar. 2019.

[14] J. Zhang, Q. Wu, C. Shen, J. Zhan, and J. Lu, "Multilabel image classification with regional latent semantic dependencies," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2801–2813, Oct. 2018.

[15] X. Zhang *et al.*, "Trip outfits advisor: Location-oriented clothing recommendation," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2533–2544, Nov. 2017.

[16] N. Pourian and B. S. Manjunath, "PixNet: A localized feature representation for classification and visual search," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 616–625, May 2015.

[17] B. Prosser, W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vision Conf.*, 2010, pp. 21.1–21.11.

[18] R. Layne, T. M. Hospedales, and S. Gong, *Person Re-identification*. Berlin, Germany: Springer, 2014.

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

[20] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, 2015, pp. 87–95.

[21] X. Liu *et al.*, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 350–359.

[22] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," 2018, *arXiv:1808.09102*.

[23] Y. Chen, S. Duffner, A. Stoian, J. Y. Dufour, and A. Baskurt, "Pedestrian attribute recognition with part-based CNN and combined feature representations," in *Proc. 15th Int. Conf. Comput. Vision Theory Appl.*, 2018, pp. 114–122.

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

[25] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Improve pedestrian attribute classification by weighted interactions from other attributes," in *Proc. Asian Conf. Comput. Vision*, 2014, pp. 545–557.

[26] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 531–540.
[27] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. IEEE Asian Conf. Pattern Recognit.*, 2015, pp. 111–115.
[28] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Li, "Grouping attribute recognition for pedestrian with joint recurrent learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 3177–3183.
[29] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.
[30] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2921–2929.
[31] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 1543–1550.
[32] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *Proc. IEEE Eur. Conf. Comput. Vision*, 2016, pp. 684–700.
[33] 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. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2027–2036.
[34] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. IEEE Eur. Conf. on Comput. Vision*, 2018, pp. 680–697.
[35] H. Han, A. K. Jain, F. Wang, S. Shan, and X. L. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2597–2609, Nov. 2018.
[36] S. Li, S. Shan, S. Yan, and X. L. Chen, "Relative forest for visual attribute prediction," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 3991–4003, Sep. 2016.
[37] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," *Pattern Recognit.*, vol. 1, pp. 582–585, 1994.
[38] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2005, pp. 886–893.
[39] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
[40] M. D. Zeiler, and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 818–833.
[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[42] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1097–1105.
[43] A. Kaya and A. B. Can, "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics," *J. Biomed. Informat.*, vol. 56, pp. 69–79, 2015.
[44] P. Domingos and M. Richardson, *Introduction to Statistical Relational Learning*. Cambridge, MA, USA: MIT Press, 2014.
[45] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen, "Deep view-sensitive pedestrian attribute inference in an end-to-end model," in *Proc. Brit. Mach. Vision Conf.*, 2017.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

Haoran An received the B.S. degree in software engineering from the China University of Petroleum, Qingdao, China, in 2017. He is currently working toward the M.S. degree in computer science and technology from Beihang University, Beijing, China.

Hai-Miao Hu received the B.S. degree in computer science from Central South University, Changsha, China, in 2005, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2012. He was a visiting student with the University of Washington from 2008 to 2009. He is currently an Associate Professor of Computer Science and Engineering with Beihang University. His research interests include video coding and networking, image/video processing, and video analysis and understanding.

Yuanfang Guo received the B.Eng. degree in computer engineering and the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2009 and 2015, respectively. Then he served as an assistant professor with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences for three years. He is currently an Assistant Professor with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include image/video security, compression and processing.

Qianli Zhou received the B.E. degree in communication technology from People's Public Security University of China, Beijing, China, in 2001, and the M.A. degree in international liaison from Westminster University, London, UK, in 2005. He is currently working toward the Ph.D. degree in engineering of security & protection system with the People's Public Security University of China, Beijing, China. His research interests include intelligent video analysis, vision and language cross-modal learning.

Bo Li received the B.S. degree in computer science from Chongqing University, Chongqing, China, in 1986, the M.S. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 1989, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 1993. In 1993, he joined the School of Computer Science and Engineering, Beihang University, where he has been a Full Professor since 1997. In 2002, he visited the University of Washington, Seattle, as a Senior Visiting Scholar for one year. He is currently the Director of the Digital Media Laboratory, School of Computer Science and Engineering, and the Vice-Director of the Professional Committee of Multimedia Technology of the China Computer Federation. He has authored/coauthored more than 100 conference proceedings and journal papers in diverse research fields, including digital video and image compression, video analysis and understanding, remote sensing image fusion, and embedded digital image processors.