Pedestrian Attribute Recognition by Joint Visual-semantic Reasoning and Knowledge Distillation

Qiaozhe Li[1,3], **Xin Zhao**[1,3], **Ran He**[2,3] and **Kaiqi Huang**[1,3]
[1]CRISE, CASIA
[2]CRIPAC & NLPR, CASIA
[3]University of Chinese Academy of Sciences
liqiaozhe2015@ia.ac.cn, {xzhao,rhe,kqhuang}@nlpr.ia.ac.cn

Abstract

Pedestrian attribute recognition in surveillance is a challenging task in computer vision due to significant pose variation, viewpoint change and poor image quality. To achieve effective recognition, this paper presents a graph-based global reasoning framework to jointly model potential visual-semantic relations of attributes and distill auxiliary human parsing knowledge to guide the relational learning. The reasoning framework models attribute groups on a graph and learns a projection function to adaptively assign local visual features to the nodes of the graph. After feature projection, graph convolution is utilized to perform global reasoning between the attribute groups to model their mutual dependencies. Then, the learned node features are projected back to visual space to facilitate knowledge transfer. An additional regularization term is proposed by distilling human parsing knowledge from a pre-trained teacher model to enhance feature representations. The proposed framework is verified on three large scale pedestrian attribute datasets including PETA, RAP, and PA-100k. Experiments show that our method achieves state-of-the-art results.

1 Introduction

Pedestrian attribute recognition aims to make prediction of a set of attributes, e.g., age, gender and clothing, as the semantic descriptions of a pedestrian image. It has recently drawn increasing attentions due to its great potential in real applications such as person retrieval [Siddiquie *et al.*, 2011] and person re-identification [Wang *et al.*, 2018]. Although great development has been made in recent years, there still exist a variety of challenges to address, such as significant pose variation, viewpoint change and poor image quality.

To boost the performance of attribute recognition, it's important to model both spatial and semantic relations of attributes. In terms of spatial distribution, some attributes may be correlated to different body parts, e.g., *Longhair* and *Boots*, while others may correspond to the same region, e.g., *Sweater* and *Shirt*. From view of semantics, some attributes are mutually exclusive, e.g., *Long-Sleeve* and *Short-Sleeve*, while others may co-appear with a high probability, e.g., *Dress* and *Female*. These relations provide important constraints for attribute recognition complementary to visual appearance features.

Previous methods [Zhu *et al.*, 2013; Deng *et al.*, 2014] solved the pedestrian attribute recognition by optimizing a separate classifier for each of the attributes. In this way, the relations between attributes are simply ignored. Some methods model the semantic relations or dependencies between attributes using weighted loss functions [Li *et al.*, 2015], probabilistic graphical models [Chen *et al.*, 2012], or Recurrent Neural Networks [Wang *et al.*, 2016; Wang *et al.*, 2017]. In these methods, pedestrian images are usually represented by a holistic model or a simple rigid structure. As each attribute may intrinsically be tied to different local regions, the spatial relations of attributes may not be captured.

To explore spatial context, some methods [Liu *et al.*, 2017; Liu *et al.*, 2018; Sarafianos and Kakadiaris, 2018] treated pedestrian attribute recognition as a weakly supervised localization problem, and proposed attention mechanisms to extract attribute-specific local features for image representation. Since accurate localization information is not available, these methods may lack the ability to describe human body structures. To overcome the above-mentioned problem, some methods [Li *et al.*, 2016c; Li *et al.*, 2018] utilize additional knowledge to guide the learning process. By extracting local features using pre-trained part detectors or around detected body key points, these methods can learn well aligned features of body parts. However, the bounding boxes are coarse annotations, thus may have limited capability to describe some fine-grained details. Besides, additional background noise may also be introduced since the rectangular bounding boxes may not always match the irregular body contours.

In this paper, a graph-based global reasoning framework is proposed to model both spatial and semantic relations of attributes. To exploit potential constraints between attributes, we first divide the attributes into multiple groups according to their semantics or their described body parts. A reasoning module is proposed to model attributes on a graph structure, with each vertex representing one particular group of attributes. To bridge the gap between visual features and semantic attributes, a projection function is learned to assign each local feature to the nodes of the graph. By aggregating

local visual features as semantic representations, the attribute groups can adaptively relate to their corresponding regions. To perform global reasoning, graph convolution [Kipf and Welling, 2017] is proposed to propagate information across the nodes. Compared with traditional methods [Wang *et al.*, 2016; Wang *et al.*, 2017] that employ RNNs to model long-range dependencies of attributes, semantic relations of attributes can be modeled in a more efficient way using the graph convolution. After performing graph-based reasoning, separate linear classifiers are applied on each node to separately predict the attributes.

There have been a variety of attempts to enhance feature representations, e.g., using reconstruction loss as a regularization [Sabour *et al.*, 2017]. However, these attempts may not be appropriate for pedestrian images as they usually contain much background noise. As an alternative, we project the node features back to visual space to predict human part segmentation maps, and utilize pixel-level classification loss as a regularization. This process can also be viewed as an exploration of human parsing knowledge to guide the visual-semantic reasoning. Compared to bounding box part detection, human parsing can precisely localize deformable body parts with more fine-grained details. Besides, the way of introducing auxiliary knowledge is different from previous methods [Li *et al.*, 2016c]. Instead of simply adopting a pre-trained detector for feature extraction, we jointly optimize semantic part localization and attribute recognition tasks, and thus can benefit from the cross-domain multi-task learning. To facilitate knowledge transfer and discovery, we perform knowledge distillation from a pre-trained human parsing model to align to its prediction distributions at each location.

The contributions of this paper are as follows:

- A graph-based reasoning module is proposed to adaptively bridge visual features and semantic attributes and to perform global reasoning between attribute groups to jointly model their spatial and semantic relations.
- A regularization term is proposed by distilling auxiliary human parsing knowledge to guide the visual-semantic reasoning and enhance feature representations.
- Experiments on three large scale pedestrian attribute datasets including PETA, RAP and PA-100k demonstrate the effectiveness of the proposed framework.

2 Related Work

2.1 Pedestrian Attribute Recognition

Semantic pedestrian attribute has been widely exploited in a variety of vision tasks [Siddiquie *et al.*, 2011; Wang *et al.*, 2018]. Earlier methods [Zhu *et al.*, 2013; Deng *et al.*, 2014] treated multiple attributes independently and trained a separate classifier for each of the attributes. Later, [Sudowe *et al.*, 2015] trained a holistic CNN model for joint multi-attribute classification. Based on [Sudowe *et al.*, 2015], [Li *et al.*, 2015] adopted weighted cross entropy loss to additionally model inter-attribute correlation. Although achieving great improvement in recognition performance, these methods fail to model potential relations between attributes. On the other hand, some methods studied semantic dependencies between attributes. [Chen *et al.*, 2012] employed a Conditional Random Field (CRF) to model mutual dependencies between cloth attributes. Inspired by [Wang *et al.*, 2016], [Wang *et al.*, 2017] proposed a RNN based recurrent sequential prediction model to capture high-order dependencies of attributes. By representing images with a holistic model or a rigid encoding scheme, these methods may not capture spatial relations of attributes.

Some methods formulate attribute recognition as a weakly supervised localization problem. [Liu *et al.*, 2017] proposed multi-directional attention modules to learn attention-strengthened features at multiple levels and scales. Based on a multi-scale attention model, [Sarafianos and Kakadiaris, 2018] added penalties on attention masks with high prediction variance to boost the recognition performance. [Liu *et al.*, 2018] extracted attribute-specific local features using a variant of class activation map to achieve attribute prediction. Without accurate localization information, these methods may have limited capability to describe human body structures.

Other methods depend on auxiliary knowledge to assist part-based models. [Zhang *et al.*, 2014] and [Li *et al.*, 2016c] utilized pretrained body-part detectors to extract multiple local features for image representation. In this way, background noise may also be included into the regions generated by coarse bounding boxes. [Li *et al.*, 2018] combined multiple local features extracted around body key points which are predicted by a pose estimation model. However, more fine-grained details may not be explored by only focusing on partial regions.

2.2 Graph-based Reasoning

Graph-based reasoning has been proved to be beneficial to a variety of vision tasks, e.g., object recognition [Chen *et al.*, 2018a] and video understanding [Ma *et al.*, 2018]. CRFs are utilized to model the dependencies between labels [Li *et al.*, 2016b] in multi-label image classification. Recently, Graph Convolutional Network (GCN) [Kipf and Welling, 2017] was proposed for semi-supervised classification in language processing. Further, [Wang and Gupta, 2018] employed GCN to perform relational learning between detected objects for video classification. [Li and Gupta, 2018] proposed to directly learn graph representations from 2D feature maps by the clustering process. For more generic context modeling, [Chen *et al.*, 2018b] proposed an end-to-end trainable reasoning module with simpler convolutional operations. [Li *et al.*, 2019] proposed a graph-based reasoning module to capture potential relations between pedestrian attributes.

2.3 Knowledge Distillation

To transfer knowledge between network models, [Hinton *et al.*, 2015] distilled knowledge from a pre-trained teacher model to improve the learning of a target net. By aligning to the teacher's prediction distributions, the representation power of the target model can be improved. For pedestrian attribute recognition, it's also desirable to explore auxiliary knowledge to achieve effective training. In this paper, we perform knowledge distillation from a pre-trained human pars-

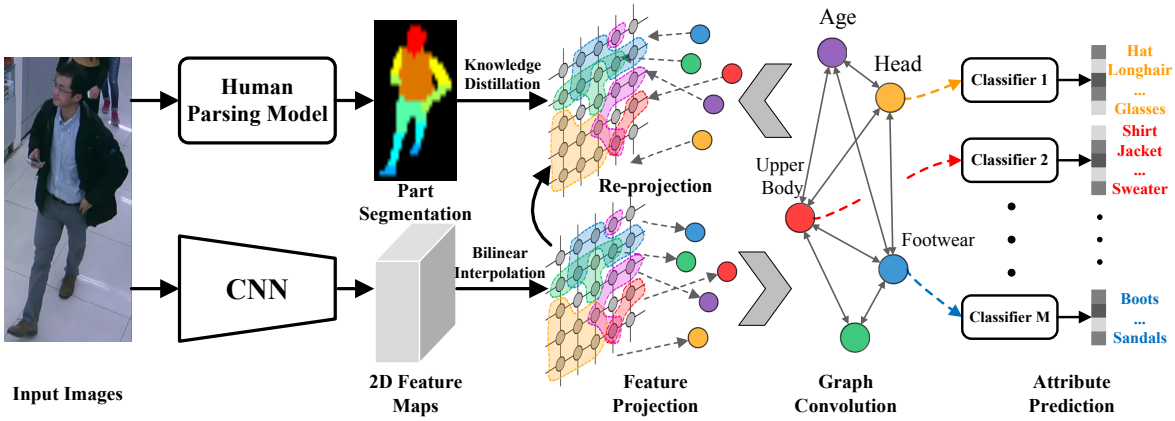

Figure 1: Overview of the proposed framework.

ing model to introduce human body knowledge to guide our visual-semantic reasoning.

3 Methodology

3.1 Framework Overview

In this paper, a graph-based reasoning framework is proposed to capture both spatial and semantic relations for attribute recognition. Given an input image, the reasoning module first projects its 2D feature maps into a graph by assigning local features to the nodes of the graph. In the graph, each node represents one specific group of attributes grouped by their semantics or their described body parts. To model mutual dependencies between the attribute groups, graph convolution is performed to propagate information along the edges and update node features. After that, separate linear classifiers are adopted on each node to classify corresponding attributes. Besides, the learned node features are also projected back to visual space to enhance feature representations. To equip the framework with human body knowledge, a residual block is adopted to utilize both inverse transformed features and the original features to predict human part segmentation maps. To achieve effective knowledge transfer, knowledge distillation is performed from a pre-trained human parsing model to align to its prediction distributions at each location. The whole framework is illustrated in Figure 1.

3.2 Visual-semantic Reasoning

To bridge local regions and semantic attributes, a projection function ϕ is learned to encode spatial features into representations of semantic nodes. In this way, different semantic nodes will adaptively relate to corresponding regions according to their characteristics. Let $\mathbf{X} \in \mathbb{R}^{N \times D^v}$ denote the visual features extracted from a convolutional layer, where $N = W \times H$ is the number of locations and D^v is the feature channel. The projection function can be formulated as:

$$\mathbf{B} = \phi(\mathbf{A}^{vs}, \mathbf{X}, \mathbf{W}^{vs}) \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{M \times D^s}$ denotes the feature matrix of semantic nodes, in which each node feature $\mathbf{b}_m \in \mathbb{R}^{D^s}$ is used to represent one specific group of attributes (e.g., gender, age or accessories). $\mathbf{W}^{vs} \in \mathbb{R}^{D^v \times D^s}$ denotes the trainable transformation matrix which projects each local visual feature $\mathbf{x}_n \in \mathbf{X}$ into the dimension D^s. $\mathbf{A}^{vs} \in \mathbb{R}^{M \times N}$ denotes the adjacency matrix which computes the assignment weights for local visual features to each semantic node. Specifically, the feature of each semantic node is computed by weighted summation of transformed local features via the assignment weights. The element $a_{m,n} \in \mathbf{A}^{vs}$, which represents the confidence of assigning local features $\mathbf{x}_n$ to the node m, is computed as:

$$a_{m,n} = \frac{\exp\left(\mathbf{w}_m^a \mathbf{x}_n\right)}{\sum_m \exp(\mathbf{w}_m^a \mathbf{x}_n)} \tag{2}$$

where $\mathbf{W}^a = [\mathbf{w}_1^a, ..., \mathbf{w}_M^a] \in \mathbb{R}^{D^v \times M}$ denotes the trainable weight matrix for computing the assignment weights. $\mathbf{A}^{vs}$ is normalized using the softmax function at each location, which means the contribution of each local feature to voting all semantic nodes sums to 1. Based on Eq.(2), the function ϕ is computed as:

$$\mathbf{B} = \mathbf{A}^{vs}\mathbf{X}\mathbf{W}^{vs} \tag{3}$$

In practice, Eq.(2) and Eq.(3) can be implemented by two convolutional layers with 1×1 kernel sizes, which is easy to implement and end-to-end trainable. Different from [Li *et al.*, 2016c; Zhang *et al.*, 2014] which represent part regions using rectangular bounding boxes, the soft assignment scheme provides a more generic solution to better describe deformable part regions.

Given the matrix $\mathbf{B}$, it's desirable to perform reasoning over the graph to capture the semantic relations between different groups of attributes. Therefore, graph convolution [Kipf and Welling, 2017] is utilized to propagate information across nodes, which is formulated as:

$$\mathbf{Z} = (\mathbf{I} - \mathbf{A}^s)\mathbf{B}\mathbf{W}^s \tag{4}$$

where $\mathbf{W}^s \in \mathbb{R}^{D^s \times D^s}$ denotes the learnable weight of the layer. $\mathbf{A}^s \in \mathbb{R}^{M \times M}$ denotes the adjacency matrix and $\mathbf{I}$ is the identity matrix. The identity matrix is adopted as a shortcut connection to facilitate optimization and $\mathbf{A}^s$ is learned from data during the training process.

After performing graph convolution, the output representations of the graph are employed for attribute prediction. It's achieved by applying separate linear classifiers for each of the semantic nodes:

$$\hat{\mathbf{p}}_{cls}^{m} = f_{cls}^{m}(\mathbf{z}_m; \theta_{cls}^{m}) \tag{5}$$

where $\mathbf{z}_m \in \mathbf{Z}$ denotes the output of the m-th node. $\hat{\mathbf{p}}_{cls}^{m}$ denotes the predicted attribute vector of group m. θ_{cls}^{m} denotes the linear weights for the m-th node. Then, the total attribute vector can be written as $\hat{\mathbf{p}}_{cls} = [\hat{\mathbf{p}}_{cls}^{1}, ..., \hat{\mathbf{p}}_{cls}^{M}] \in \mathbb{R}^{K}$. Note that it's also feasible to directly represent each attribute with one node. However, since some correlated attributes may relate to the same local region, reasoning on groups can exploit the potential constraints between attributes.

To facilitate knowledge transfer, the output features of the nodes are then projected back to visual space. Given the representations $\mathbf{Z}$, a mapping function $\widetilde{\mathbf{X}} = \varphi(\mathbf{A}^{sv}, \mathbf{Z}, \mathbf{W}^{sv})$ is learned to perform inverse feature transformation. Similar as Eq.(3), function φ is implemented as:

$$\widetilde{\mathbf{X}} = \mathbf{A}^{sv}\mathbf{Z}\mathbf{W}^{sv} \tag{6}$$

where $\mathbf{A}^{sv} \in \mathbb{R}^{N \times M}$ is the inverse assignment matrix and $\mathbf{W}^{sv} \in \mathbb{R}^{D^s \times D^v}$ is the learnable weight matrix. The inverse assignment matrix is set to $\mathbf{A}^{sv} = (\mathbf{A}^{vs})^{\top}$ for computational efficiency. Further, a residual connection is adopted to utilize both transformed features and the original features to train a human parsing classifier $f_{prs}((\mathbf{X} + \widetilde{\mathbf{X}}); \theta_{prs})$. By imposing an additional constraint for the reasoning module, the proposed framework can introduce auxiliary human body knowledge to improve its representation capability.

3.3 Loss Function

The whole network is end-to-end trained using an object function which is the sum of three losses. First, the cross entropy loss is employed to achieve multi-class attribute classification:

$$L_{cls} = -\frac{1}{K}\sum_{k=1}^{K} y_{cls}^{k}\log(\hat{p}_{cls}^{k}) + (1 - y_{cls}^{k})\log(1 - \hat{p}_{cls}^{k}) \tag{7}$$

where $\hat{p}_{cls}^{k} \in \hat{\mathbf{p}}_{cls}$ denotes the output probability of the k-th attribute. y_{cls}^{k} is the corresponding ground truth annotation.

Besides the attribute classification, our proposed network also predicts a set of segmentation maps for localizing human parts. The output label maps are 3D tensors with a shape of $H \times W \times C$, where C denotes the number of classes including the background. Let $\hat{z}_{prs}^{i,c}$ denote the logits for the i-th location predicted by our network where $c \in \{1, ..., C\}$ belongs to one of C classes, the normalized output probability $\hat{p}_{prs}^{i,c}$ can be computed as $\hat{p}_{prs}^{i,c} = \exp(\hat{z}_{prs}^{i,c})/\sum_{j=1}^{C}\exp(\hat{z}_{prs}^{i,j})$. Similarly, the teacher's output probability is computed as $p_{prs}^{i,c} = \exp(z_{prs}^{i,c})/\sum_{j=1}^{C}\exp(z_{prs}^{i,j})$ with the logits $z_{prs}^{i,j}$. Thus, the pixel-wise classification loss can be formulated as:

$$L_{prs} = -\frac{1}{H \times W}\sum_{i=1}^{H \times W}\sum_{c=1}^{C}\delta_{i,c}\log(\hat{p}_{prs}^{i,c}) \tag{8}$$

where $\delta_{i,c}$ is the Dirac delta function which returns 1 if $c = \arg\max_{c \in \{1,...,C\}}(p_{prs}^{i,c})$, and 0 otherwise.

With the pixel-wise classification loss, the proposed network is trained to predict the pseudo labels in principle of maximum likelihood. To further enhance knowledge discovery and transfer, knowledge distillation is performed by computing soft probability distributions at a temperature of T for both the teacher and our proposed network as:

$$p_{kl}^{i,c} = \frac{\exp(z_{prs}^{i,c}/T)}{\sum_{j=1}^{C}\exp(z_{prs}^{i,j}/T)}, \hat{p}_{kl}^{i,c} = \frac{\exp(\hat{z}_{prs}^{i,c}/T)}{\sum_{j=1}^{C}\exp(\hat{z}_{prs}^{i,j}/T)} \tag{9}$$

To measure prediction similarity between the proposed framework and the teacher at each pixel location, the Kullback Leibler divergence is employed as:

$$L_{kl} = \frac{1}{H \times W}\sum_{i=1}^{H \times W}\sum_{c=1}^{C} p_{kl}^{i,c}\frac{\log(p_{kl}^{i,c})}{\log(\hat{p}_{kl}^{i,c})} \tag{10}$$

Finally, the overall loss function can be obtained by:

$$L = L_{cls} + L_{prs} + T^2 * L_{kl} \tag{11}$$

where T^2 denotes the scaling factor for distillation loss to make sure the contributions of the second term and third term are comparable since the gradient magnitudes produced by the soft targets are scaled by $1/T^2$.

4 Experiments

Datasets. The proposed method is evaluated on three large-scale pedestrian attribute datasets: (1) The PEdesTrian Attribute (PETA) dataset [Deng *et al.*, 2014] consists of 19, 000 person images collected from 10 small-scale person datasets. The whole dataset is randomly divided into three non-overlapping partitions: 9500 for training, 1900 for verification, and 7600 for evaluation. In this dataset, 35 attributes whose positive ratios are higher than 5% are used for evaluation. (2) The Richly Annotated Pedestrian (RAP) attribute dataset [Li *et al.*, 2016a] contains 41,585 images drawn from 26 indoor surveillance cameras. Each image is labelled with 69 binary attributes and 3 multi-class attributes. Following the official protocol, the whole dataset is split into 33,268 training images and 8,317 test images. The recognition performance is evaluated on 51 binary attributes. (3) The PA-100k Dataset [Liu *et al.*, 2017] consists of 100,000 pedestrian images from 598 outdoor scenes. Each image is described with 26 commonly used attributes. The whole dataset is split into training, validation and test sets with a ratio of 8:1:1.

Implementation Details. For human semantic parsing, we adopt the architecture of [Kalayeh *et al.*, 2018] as the teacher model and use the Densepose [Alp Guler *et al.*, 2018] dataset for training. The Densepose dataset contains 14 part annotations. To reduce training difficulties, the left/right parts are fused and the hand regions are assigned to lower arm class, which lead to 7 parts eventually. The parsing net takes images of size 512×512 as inputs and outputs prediction maps of size 30×30. The network is trained for 20 epochs with a batch size of 8. We employ a ResNet-50 network for image

representation, and extract convolutional features of the last residual block ("Res_5c") as the input for our visual-semantic reasoning module. For data augmentation, the input images are randomly scaled from 384×192 to 256×128 for each mini batch. To match the output resolution of the parsing net, bilinear interpolation is employed to scale up the input feature maps of the reasoning module to size 30×30. D^v is 2048 and D^s is set to 512. The temperature T is set to 3. The attributes are divided into 7 groups for PETA and 10 groups for RAP following [Zhao *et al.*, 2018]. For PA-100k dataset, the attributes are divided into 8 groups including *gender, age, view angle, head, accessories, upper body, lower body* and *footwear*. Function $f_{prs}(\cdot)$ is implemented using an atrous spatial pyramid pooling followed by a 1×1 convolution layer for classification. The network is optimized by stochastic gradient descend algorithm with a batch size of 16, a momentum of 0.9 and a weight decay of 0.0005. The initial learning rate is set to 0.001 and is divided by 10 after every 30 epochs. The reasoning network is trained for 60 epochs.

Performance Metrics. Two kinds of metrics are adopted to evaluate attribute recognition performance. (1) Class-based: The mean Accuracy (mA) is usually utilized as the class-based measure. (2) Instance-based: The instance-based metrics include accuracy, precision, recall rate and F1-score [Li *et al.*, 2016a]. For accuracy, precision and recall, the scores of the predicted attributes against the groundtruth are first computed for each instance and then averaged over all test images. The F1-score is computed based on precision and recall.

Competitors. The proposed method is compared against 10 state-of-the-art models. (1) ELF-mm [Gray and Tao, 2008] employs SVM classifier with Ensemble of Localized Features (ELF) for attribute recognition; (2)-(3) FC7-mm and FC6-mm replace the hand-crafted ELF features with CNN features (FC7 and FC6 output of the AlexNet); (4) Attributes Convolutional Network (ACN) [Sudowe *et al.*, 2015] jointly trains a CNN model for all attributes, which allows to share weights and transfer knowledge among different attributes; (5) DeepMAR [Li *et al.*, 2015] additionally considers inter-attribute correlation by weighted cross entropy loss function; (6) HP-net [Liu *et al.*, 2017] is an attention based method that employs multi-directional attention modules to train multi-level and multi-scale attention-strengthened features; (7) MsVAA [Sarafianos and Kakadiaris, 2018] also aggregates visual attention on multi-scales, combined with additional penalties on attention masks and a weighted loss function. (8) JRL model [Wang *et al.*, 2017] employs RNN encoder-decoder to jointly learn image level context and attribute level sequential correlation for prediction; (9) VeSPA model [Sarfraz *et al.*, 2017] jointly learns a coarse view predictor and view-dependent image features for attribute inference; (10) PGDM [Li *et al.*, 2018] learns a pose-normalized feature representation for recognition by extracting and aligning local features around detected key points.

4.1 Experimental Results

Table 1 reports the evaluation results on three datasets. On PETA dataset, JRL achieves the best score in mA and our model reports the second best result (85.67% vs. 84.90%).

Dataset	Method	Metric				
		mA	Acc	Pre	Recall	F1
PETA	ELF-mm	75.21	43.68	49.45	74.24	59.36
	FC7-mm	76.65	45.41	51.33	75.14	61.00
	FC6-mm	77.69	48.31	54.06	76.49	63.35
	ACN	81.15	73.66	84.06	81.26	82.64
	Deep-Mar	82.89	75.07	83.68	83.14	83.41
	HP-net	81.77	76.13	84.92	83.24	84.07
	MsVAA	84.59	78.56	86.79	86.12	86.46
	JRL	**85.67**	-	86.03	85.34	85.42
	VeSPA	83.45	77.73	86.18	84.81	85.49
	PGDM	82.97	78.08	86.86	84.68	85.76
	Ours	84.90	**80.95**	**88.37**	**87.47**	**87.91**
RAP	ELF-mm	69.94	29.29	32.84	71.18	44.95
	FC7-mm	72.28	31.72	35.75	71.78	47.73
	FC6-mm	73.32	33.37	37.57	73.23	49.66
	ACN	69.66	62.61	80.12	72.26	75.98
	Deep-Mar	73.79	62.02	74.92	76.21	75.56
	HP-net	76.12	65.39	77.33	78.79	78.05
	MsVAA	-	-	-	-	-
	JRL	77.81	-	78.11	78.98	78.58
	VeSPA	77.70	67.35	79.51	79.67	79.59
	PGDM	74.31	64.57	78.86	75.90	77.35
	Ours	**78.30**	**69.79**	**82.13**	**80.35**	**81.23**
PA-100k	Deep-Mar	72.70	70.39	82.24	80.42	81.32
	HP-net	74.21	72.19	82.97	82.09	82.53
	PGDM	74.59	73.08	84.36	82.24	83.29
	Ours	**77.87**	**78.49**	**88.42**	**86.08**	**87.24**

Table 1: Evaluation results on three datasets. The 1st and 2nd best results are in bold fonts and underlined, respectively.

Despite that, the proposed method still outperforms the state-of-the-arts on all four instance-based metrics by 2.39%, 1.51%, 1.35%, and 1.45%, respectively. On RAP dataset, the proposed method has achieved the best performance on both class-based and instance-based metrics. ACN model presents the second best result in precision and VeSPA achieves the second best results in accuracy, recall rate and F1-score. PA-100k is a newly proposed dataset thus has fewer released results. On this dataset, PGDM has reported better results compared to Deep-Mar and HP-net due to its exploration of coarse pose information. However, its scores are lower than our proposed method, especially in accuracy (73.08% vs. 78.49%) and recall rate (82.24% vs. 86.08%). It indicates that PGDM tends to miss some attributes in recognition, which might be caused by its limited ability in capturing fine-grained details. In contrast, our method has significantly improved the results by all metrics due to its effectiveness of distilling human parsing knowledge as the guidance for reasoning.

4.2 Ablation Study

The improvement of the proposed method can be contributed to two aspects: visual-semantic graph reasoning and auxiliary human parsing knowledge distillation. In this section, we conduct experiments to show how these two aspects improve recognition performance.

Effect of Visual-semantic Graph Reasoning. For better comparison, a simple ResNet-50 model is adopted as the baseline. Without the visual-semantic reasoning module, another model is implemented by exploiting parsing results to

Dataset	Method	Metric				
		mA	Acc	Pre	Recall	F1
PETA	Baseline	81.27	76.69	87.33	82.76	84.99
	Model-F	82.09	77.40	87.84	83.58	85.75
	Ours	**84.90**	**80.95**	**88.37**	**87.47**	**87.91**
RAP	Baseline	75.12	66.67	81.16	76.52	79.00
	Model-F	76.08	67.45	81.48	77.33	79.82
	Ours	**78.30**	**69.79**	**82.13**	**80.35**	**81.23**
PA-100k	Baseline	76.31	76.76	**88.62**	83.22	85.84
	Model-F	76.91	77.37	88.30	83.72	85.95
	Ours	**77.87**	**78.49**	88.42	**86.08**	**87.24**

Table 2: Effect of visual-semantic graph reasoning. The best result is in bold.

Dataset	Method	Metric				
		mA	Acc	Pre	Recall	F1
PETA	Model-N	83.37	78.73	87.84	84.13	85.94
	Model-I	80.75	76.05	86.37	82.45	84.36
	Ours	**84.90**	**80.95**	**88.37**	**87.47**	**87.91**
RAP	Model-N	76.76	67.98	82.01	78.06	79.99
	Model-I	75.04	66.17	80.64	76.13	78.32
	Ours	**78.30**	**69.79**	**82.13**	**80.35**	**81.23**
PA-100k	Model-N	77.28	78.31	**88.51**	84.96	86.70
	Model-I	76.12	75.93	87.76	82.79	85.20
	Ours	**77.87**	**78.49**	88.42	**86.08**	**87.24**

Table 3: Effect of auxiliary human parsing knowledge distillation. The best result is in bold.

extract multiple part features and then concatenating local features as a global representation for multi-attribute classification. It's denoted as Model-F. Instead of acting as a semantic regularization, the human parsing knowledge is merely used for feature extraction and alignment. These two models are compared with the proposed method. As is shown in Table 2, by introducing human parsing results for feature extraction and alignment, the evaluation results can be slightly improved. However, there still exists a significant gap in performance between Model-F and our reasoning framework, especially in recall rate (83.58% vs. 87.47% on PETA and 77.33% vs. 80.35% on RAP). It demonstrates the effectiveness of our visual-semantic reasoning in modeling the spatial and semantic relations of attributes.

Effect of Auxiliary Human Parsing Knowledge Distillation. Two additional models are implemented to show the effectiveness of auxiliary knowledge distillation. The first model only performs visual-semantic reasoning and ignores the regularization term for knowledge transfer. The second model replaces human parsing knowledge distillation with reconstruction of input images and uses reconstruction loss as a regularization following [Sabour *et al.*, 2017]. They are respectively denoted as Model-N and Model-I. As is shown Table 3, simply adopting reconstruction loss as the regularization could lead to performance decrease on all three datasets. It might be caused by the difficulties to reconstruct input images as they usually contain much background noise. In comparison, the proposed model can benefit from knowledge distillation as this process introduces human body knowledge for the reasoning module.

Figure 2: Qualitative results of human parsing and attribute recognition. The correct and wrong attribute predictions are marked in green and red, respectively. The samples are from RAP and PA-100k.

Qualitative Evaluation. Figure 2 shows the human parsing and attribute recognition results of two pedestrian images from RAP and PA-100k. Results show that the pretrained human parsing model can accurately segment most body parts of pedestrian images, which is favorable for our visual-semantic reasoning framework. In recognition, the baseline ResNet-50 model makes some wrong predictions and misses some attributes. It might be caused by its limited capability to describe human body structures. In contrast, the proposed method can correctly predict *lb-LongTrousers* and *shoes-Casual* in the first image and recognize all attributes in the second image.

5 Conclusion

In this paper, a graph-based global reasoning framework is proposed to jointly model potential spatial and semantic relations of attributes and exploit auxiliary knowledge for attribute recognition. The reasoning module not only adaptively bridges local visual features and semantic attributes but also models the dependencies between attribute groups by performing graph-based reasoning. A regularization term is proposed by distilling human parsing knowledge to enhance feature presentations and guide the visual-semantic reasoning. Experiment results show superiority of the proposed method over state-of-the-arts and effectiveness of our reasoning module and auxiliary human parsing knowledge distillation.

Acknowledgements

This project is supported by the National Key Research and Development Program of China (Grant No. 2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61673375 and No.61602485), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006).

References

[Alp Guler *et al.*, 2018] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018.

[Chen *et al.*, 2012] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623, 2012.

[Chen *et al.*, 2018a] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, pages 7239–7248, 2018.

[Chen *et al.*, 2018b] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. *arXiv preprint arXiv:1811.12814*, 2018.

[Deng *et al.*, 2014] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, pages 789–792, 2014.

[Gray and Tao, 2008] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Kalayeh *et al.*, 2018] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[Li and Gupta, 2018] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *NIPS*, pages 9245–9255, 2018.

[Li *et al.*, 2015] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, 2015.

[Li *et al.*, 2016a] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. In *arXiv preprint arXiv:1603.07054*, 2016.

[Li *et al.*, 2016b] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In *CVPR*, pages 2977–2986, 2016.

[Li *et al.*, 2016c] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, pages 684–700, 2016.

[Li *et al.*, 2018] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *ICME*, pages 1–6, 2018.

[Li *et al.*, 2019] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Visual-semantic graph reasoning for pedestrian attribute recognition. In *AAAI*, 2019.

[Liu *et al.*, 2017] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, pages 350–359, 2017.

[Liu *et al.*, 2018] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. In *BMVC*, 2018.

[Ma *et al.*, 2018] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, pages 6790–6800, 2018.

[Sabour *et al.*, 2017] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NIPS*, pages 3856–3866, 2017.

[Sarafianos and Kakadiaris, 2018] Nikolaos Sarafianos and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, pages 680–697, 2018.

[Sarfraz *et al.*, 2017] M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *BMVC*, 2017.

[Siddiquie *et al.*, 2011] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, pages 801–808, 2011.

[Sudowe *et al.*, 2015] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV Workshops*, 2015.

[Wang and Gupta, 2018] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018.

[Wang *et al.*, 2016] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, pages 2285–2294, 2016.

[Wang *et al.*, 2017] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, pages 531–540, 2017.

[Wang *et al.*, 2018] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, pages 2275–2284, 2018.

[Zhang *et al.*, 2014] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, pages 1637–1644, 2014.

[Zhao *et al.*, 2018] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *IJCAI*, pages 3177–3183, 2018.

[Zhu *et al.*, 2013] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *ICCV Workshops*, pages 331–338, 2013.