Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization
Supplementary Material

Chufeng Tang[1] Lu Sheng[2] Zhaoxiang Zhang[3] Xiaolin Hu[1*]
[1]State Key Laboratory of Intelligent Technology and Systems,
Institute for Artificial Intelligence, Department of Computer Science and Technology,
Beijing National Research Center for Information Science and Technology, Tsinghua University
[2]College of Software, Beihang University [3]Institute of Automation, Chinese Academy of Sciences
{tcf18@mails,xlhu@mail}.tsinghua.edu.cn lsheng@buaa.edu.cn zhaoxiang.zhang@ia.ac.cn

1. Implementation Details

We adopt the BN-Inception model pretrained from ImageNet as the backbone network. The proposed framework is implemented with PyTorch framework and trained end-to-end with only image-level annotations. We adopt Adam optimizer since it converges faster than SGD in our experiments with momentum set to 0.9 and a weight decay equals to 0.0005. The initial learning rate equals to 0.0001 and the batch size is set to 32. For RAP and PA-100K dataset, we train the model for 30 epochs and the learning rate decays by 0.1 every 10 epochs. For the smaller PETA dataset, we double the training epochs. For data preprocessing, we resize the input pedestrian images to 256×128 and apply random horizontal mirroring and data shuffling for data augmentation.

2. Different Attribute-Specific Methods

In Section 4.4 of the main paper, we compare the proposed method against the other two attribute-specific localization methods, including visual attention and rigid parts. Different from most existing attribute-agnostic attention-based and part-based methods, we build two attribute-specific models based on these ideas for comparison. Here we show the details of the compared models.

2.1. Attention Masks Model

We replace the proposed ALM with a spatial attention module while keeping others unchanged for fair comparison. The spatial attention module is implemented by a tiny 3-layers sub-network, as shown in Figure S2, which is inspired by HA-CNN [2]. The input features $\mathbf{X}_i \in \mathbb{R}^{H \times W \times C}$ at the i-th level (a certain layer in the backbone network, totally three levels) are first fed into a cross-channel averaging layer. A 3×3 Conv-BatchNorm-ReLU block is followed to generate the expected attention mask $\mathbf{S}_i^m \in \mathbb{R}^{H \times W \times 1}$, which is used for localizing the m-th attribute at the i-th level. All channels share the identical spatial attention mask. Subsequently, the attentive features are obtained by channel-wise multiplying the attention mask with the input features, and the corresponding prediction is calculated as follows:

$$\hat{y}_i^m = f(\mathbf{S}_i^m \cdot \mathbf{X}_i), \tag{1}$$

where f denotes a fully-connected layer. Each spatial attention module only serves one attribute at a singe level, the same as Figure 2 in the main paper.

2.2. Rigid Parts Model

For attribute-specific part-based model, we replace ALM with a body-parts guided module, as shown in Figure S3. The key idea is to associate each attribute with a predefined body region, including *head, torso, legs*, and the whole image, *e.g.*, the *LongHair* attribute is associated with the head part. Since the body-part annotations are unavailable on most pedestrian attribute datasets, we adopt an external pose estimation model to localize the body parts, which is inspired by SpindleNet [3]. Specifically, we localize 14 human body keypoints for each pedestrian image using a pretrained pose estimation model [3]. The pedestrian image is then divided into three body-part regions based on these keypoints, as shown in Figure S1. In the body-parts guided module (Figure S3), the body-part-based local features are extracted from the input features $\mathbf{X}_i$ through an RoI pooling layer [1]. For attribute prediction, the most relevant features are selected according to the attribute-region correspondence, as listed in Table S1, *e.g.* recognizing *hat* using features only from the *head* part.

*Corresponding author.

Region	Attributes
Head	BaldHead, LongHair, BlackHair, Hat, Glasses, Muffler, Calling
Torso	Shirt, Sweater, Vest, TShirt, Cotton, Jacket, Suit-Up, Tight, ShortSleeve, LongTrousers, Skirt, ShortSkirt, Dress, Jeans, TightTrousers, CarryingbyArm, CarryingbyHand
Legs	LeatherShoes, SportShoes, Boots, ClothShoes, CasualShoes
Whole	Female, AgeLess16, Age17-30, Age31-45, BodyFat, BodyNormal, BodyThin, Customer, Clerk, Backpack, SSBag, HandBag, Box, PlasticBag, PaperBag, HandTrunk, OtherAttchment, Talking, Gathering, Holding, Pushing, Pulling

Table S1: Attribute-region correspondence in RAP dataset.

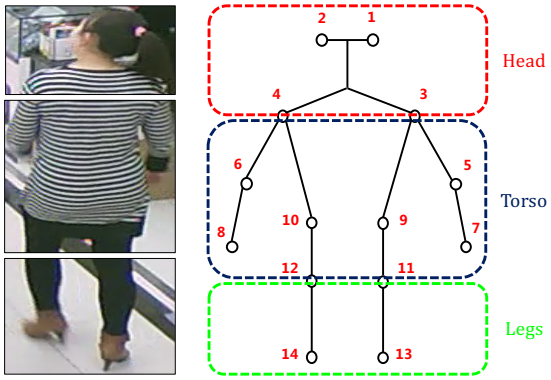

Figure S1: Illustration of body-parts generation. We divide a pedestrian image into three body-part regions (*head, torso, and legs*) based on 14 human body keypoints.

2.3. More Results

We provide more localization results belong to different attributes, as shown in Figure S4.

References

[1] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[2] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[3] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 907–915, 2017.

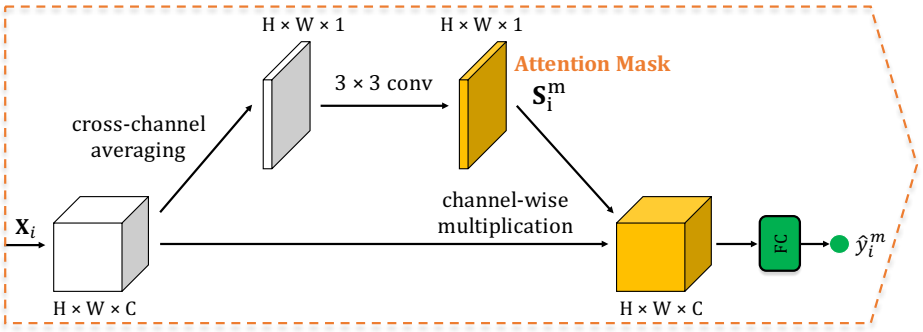

Figure S2: Details of the spatial attention module for one attribute at a singe level. The expected attention mask follows a cross-channel averaging layer and a 3×3 Conv-BN-ReLU block.

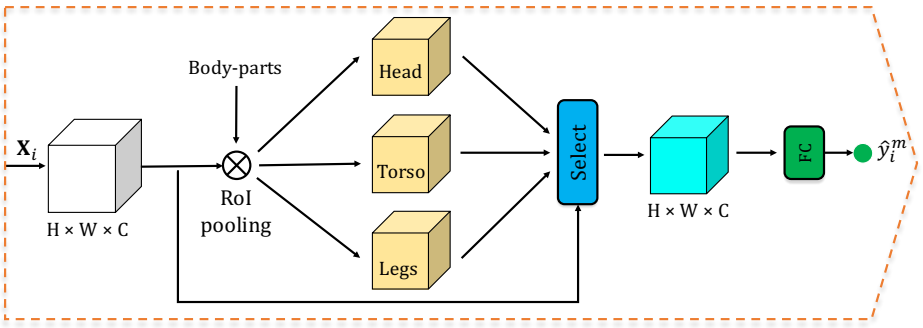

Figure S3: Details of the body-parts guided module for one attribute at a singe level. The three body-part regions are calculated based on several human body keypoints predicted by a pretrained pose estimation model. The local features belonging to different body-parts are extracted by an RoI pooling layer. The most relevant features are selected for attribute classification according to the predefined attribute-region correspondence (Table S1).

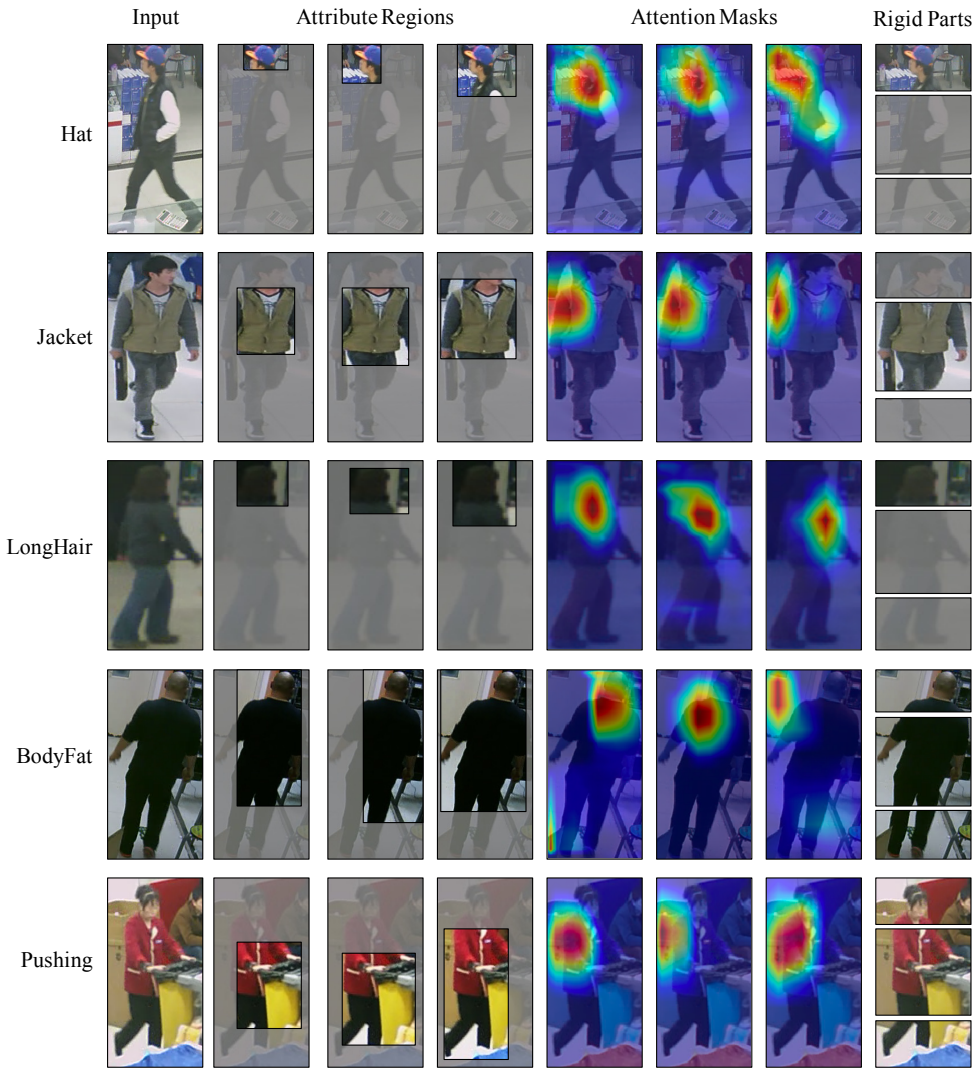

Figure S4: Case studies of different attribute-specific localization methods for five different attributes.