Supplementary Materials for PromptPAR

Xiao Wang, *Member, IEEE*, Jiandong Jin, Chenglong Li*, Jin Tang, Cheng Zhang, Wei Wang

***Index Terms*—Pedestrian Attribute Recognition, Pre-trained Big Models, Prompt Learning, Multi-Modal Fusion, Vision-Language**

A. Dataset and Evaluation Metric

Extensive experiments are conducted on five publicly available pedestrian attribute recognition datasets, including **PETA** [1], **PA100K** [2], **RAPv1** [3], **RAPv2** [4], **WIDER** [5]. A brief introduction to these datasets is given below.

• **PETA** [1] contains 19,000 outdoor or indoor pedestrian images and 61 binary attributes. These images are split into training, validation, and testing subset, which contains 9500, 1900, and 7600. Following the work [1], we select 35 pedestrian attributes for our experiments.

• **RAPv1** [3] contains 41,585 pedestrian images and 69 binary attributes, where 33,268 images are used for training. Usually, 51 attributes are selected for training and evaluation.

• **RAPv2** [4] has 84,928 pedestrian images and 69 binary attributes, where 67,943 were used for training. We select 54 attributes for the training and evaluation of our model.

• **PA100K** [2] is the largest pedestrian attribute recognition dataset which contains 100,000 pedestrian images, and 26 binary attributes. Note that, 90,000 images are used for training and validation, and the rest 10,000 images are utilized for testing.

• **WIDER** [5] contains 13,789 images with 57,524 annotated pedestrians and 14 attributes. The authors divide 6871 images for the training and validation, and 6918 images for testing.

In addition to the standard setting as aforementioned above, we also validate our model based on the zero-shot setting proposed by Jia et al. [6]. The zero-shot PAR splits the training and testing subset based on personal identity and no overlaps are shared between them. Two datasets are adopted for this experiment and the detailed information is given below:

• **PETA-ZS** is proposed by Jia et al. based on PETA [1] dataset by following the zero-shot protocol. The training, validation, and testing subset contains 11241, 3826, and 3933 samples. 35 common attributes are adopted for our experiments by following Jia et al. [6].

Xiao Wang, Jin Tang, and Cheng Zhang are with the School of Computer Science and Technology, Anhui University, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230601, China. (email: xiaowang@ahu.edu.cn, tangjin@ahu.edu.cn, cheng.zhang@ahu.edu.cn)

Jiandong Jin, Chenglong Li are with Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, the School of Artificial Intelligence, Anhui University, Hefei 230601, China. (email: jdjinahu@foxmail.com, lcl1314@foxmail.com)

Wei Wang is with Video Investigation Detachment of Hefei Public Security Bureau, Hefei 230031, China. (email: yzyzww@sohu.com)

* Corresponding Author: Chenglong Li

• **RAP-ZS** is developed based on the RAPv2 dataset and contains 17062, 4628, and 4928 pedestrian images for training, validation, and testing, respectively. No shared personal identity between the training and inference data. Following Jia et al. [6], we select 53 attributes for the evaluation.

In our experiments, five widely used metrics are adopted for the evaluation of PAR models, including **mA**, **Acc**, **Precision**, **Recall**, and **F1-score**. To be specific, the label-based evaluation metric **mean Accuracy (mA)** is defined as:

$$mA = \frac{1}{2N}\sum_{i=1}^{N}\left(\frac{TP_i}{TP_i + FN_i} + \frac{TN_i}{TN_i + FP_i}\right) \tag{1}$$

where N is the number of attributes, TP_i and TN_i are the number of correctly classified positive and negative samples of the *i*-th attribute, and FN_i and FP_i are the misclassified numbers of positive and negative samples of the *i*-th attribute. The instance-based evaluation metric **Accuracy (Acc)** can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

The formulation of **Precision, Recall** and **F1-score (F1)** can be expressed as:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP is predicting the correct positive sample, TN is predicting the correct negative sample, FP is a negative sample of prediction errors, and FN is a positive sample of prediction errors.

REFERENCES

[1] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792.

[2] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 350–359.

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

[4] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575–1590, 2019.

[5] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*, 2016.

[6] J. Jia, H. Huang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting," *arXiv preprint arXiv:2107.03576*, 2021.