Let's Observe Them Over Time: An Improved Pedestrian Attribute Recognition Approach (Supplementary Document)

This supplementary document includes a brief discussion of key aspects of the paper, a few additional qualitative results, and implementation details of the baselines. In Sec. 1, we have discussed common questions regarding the proposed work. This section helps to understand our thought process and research directions that we have followed during the implementation of the present work. In Sec. 2, we present a few additional results and the architecture of the baseline.

1. FAQs and Key Aspect

Here, we have briefly attempted to discuss the essential aspects of the proposed work.

1.1. Existing Time-domain PAR Datasets

How many existing PAR datasets constituent multiple appearances of a single pedestrian? There are few. The popular and widely used datasets such as PR-100K [3] and RAP [1] consist of a single pedestrian appearance. However, other datasets like Market-1501 [5], Duke [3], and a few subsets of PETA like 3DPeS, Grid, i-LiD, and MIT constituent of multiple occurrences of the same pedestrian. Hence these datasets are suitable for multi-perspective approaches or spatio-temporal analysis. Fig. 3 depicts a pedestrian with four appearances that have been visualized for the *accessories* attribute.

1.2. Class-Activation Energy Estimation

How class-activation energy proportion has been estimated? Firstly, we obtain the class-activation map from the *target* layer of the baseline. Secondly, we divide the input image into two halves and estimate the energy corresponding to it using Eq.(1).

$$\frac{\sum L^{c}_{(i,j)\in bbox}}{\sum L^{c}_{(i,j)\in bbox} + \sum L^{c}_{(i,j)\notin bbox}} \tag{1}$$

In Eq.(1), $L^{c}_{(i,j)\in bbox}$ is the class-activation energy value of a pixel of coordinate (i, j) present in the bounding box. As suggested by the Score-CAM [4], this estimation is *energy-based pointing game* where we add the energy of each pixel present inside the bounding box (upper/lower body part). Fig. 1 depicts two examples of energy proportions.

Figure 1. **Class-activation Energy:** Class-activation energy in the lower and upper body part. The presented numbers are the energy *proportion* in the specific part. The first prediction is for *hand back*, and the second is for *boots*.

1.3. Body Parts Division

Why division of the body is necessary? We have calculated the energy proportion using the upper or lower body part as a reference box. This formulation is inspired by the observation that many pedestrian attributes tend to appear in specific regions of the image. For example, at the time of energy calculation of the *hat*, we used the upper body part as a reference box. In vice versa, if the energy proportion for *long hair* is higher in the upper part than the lower, we safely concluded that it is stronger evidence for *long hair*. In Fig.1, the second row represents the CAM-activation for *boots*, which is higher in the lower part, making it stronger evidence.

This approach assumes attributes are location-specific in nature. For example, *handbags, boots, pants* can always found in the *lower* body part of the pedestrian. However, this approach can be extended to multiple body parts such as

4×4 grid. However, such granular division not necessarily promise the location of each attribute. For example, *hat* attribute can appear either in *top-right* or *top-left* part of the image depending on the view. We believe this will make system vulnerable to view-specific inputs.

2. Additional Qualitative Results

In this section, we have presented a few additional qualitative results of the work. Fig. 2 shows attribute-specific localization of the proposed method for different attributes. Fig. 4 depicts two prediction results comparisons of the proposed method, DeepMAR and a variant of the DeepMAR. In Fig. 5, we have shown the CNN architecture proposed by the DeepMAR [2].

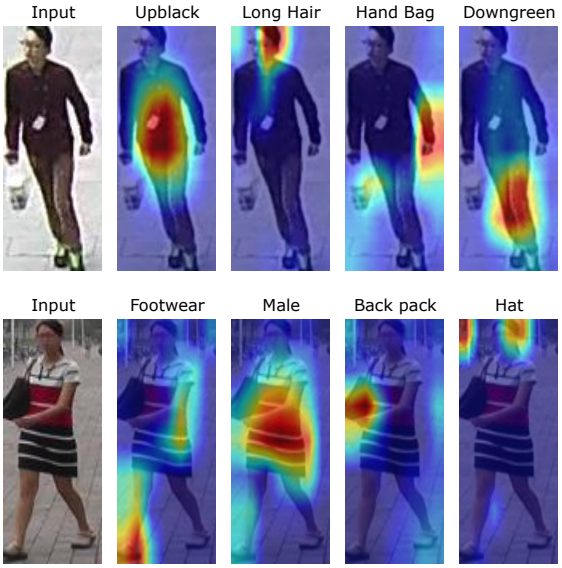

Figure 2. **Attribute-specific Localization:** Visualization of the prediction of specific attributes in the image generated by the proposed model.

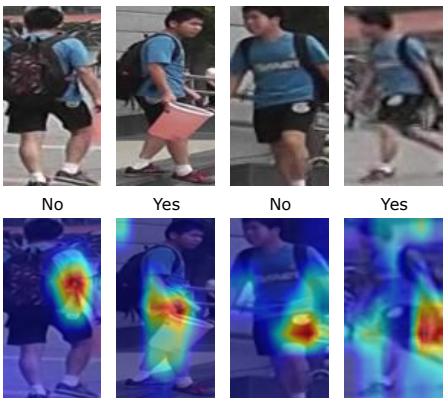

Figure 3. **Occlusions hurt the prediction: For *accessories* attribute.** Four perspectives of a person and class-activation energy of the proposed model for the *hand bag*.

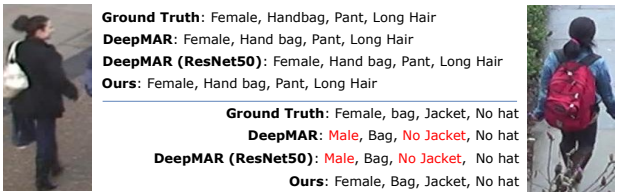

Figure 4. **Prediction Quality:** Qualitative analysis of the proposed method, DeepMAR [2], DeepMAR with ResNet50 as the visual encoder. Wrong predictions are in red; right predictions are in black.

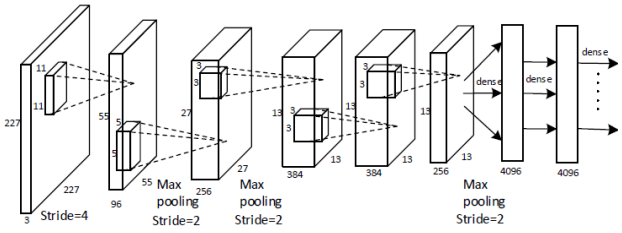

Figure 5. **DeepMAR Network:** A DeepMAR CNN architecture proposed in [2]. We have replaced this CNN network with ResNet50 due to its proven efficiency.

References

[1] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021. 1

[2] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, 2015. 2

[3] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 1

[4] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 1

[5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 1