Multi-Level Factorisation Net for Person Re-Identification (Supplementary Material)

Xiaobin Chang[1], Timothy M. Hospedales[2], Tao Xiang[1]
Queen Mary University of London[1], The University of Edinburgh[2]
x.chang@qmul.ac.uk t.hospedales@ed.ac.uk t.xiang@qmul.ac.uk

1. MLFN Architecture Parameter Selection

The number of blocks (N) in MLFN is set to 16 follows the ResNeXt-50 [3] architecture. The FS dimension K depends on N and the number of FMs at each MLFN block. We set these, without tuning, so that the model is of a comparable overall size to ResNeXt-50 [3] for direct comparison. On our GTX1080 GPU, the runtime is similar: MLFN (0.81s/batch) and ResNeXt (0.78s/batch), and so is the GPU memory consumption. The final feature dimension d of MLFN is set to 1024 since it is the widely used feature dimension for Person ReID such as [2]. The impacts of different d values on the re-id performance are illustrated as in Figure 1. It can be seen that the performance is consistently good when $d > 512$.

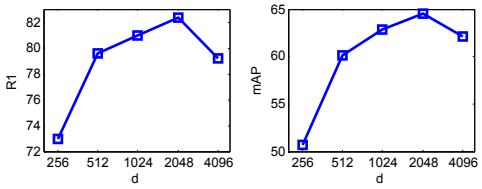


Figure 1: Sensitivity to dimemsion d. Duke [4] is used.

2. Examples of FS Predicted Attributes

In Sec. 4.4.2 of the main paper, we have shown that the attribute prediction accuracy obtained with the factor signature (FS, $\hat{\boldsymbol{S}}$) alone in the proposed MLFN is already better than a supervised attribute prediction model APR [1] (e.g., 82.30% vs 80.12% on DukeMTMC-reID). Here, we show some qualitative results.

Figure 2 shows three examples where the predicted attributes using our FS feature and the human labelled attributes are compared. For each person image, 35 binary attributes are annotated by human annotators on the identity level, that is, different images of the same person would have the identical attribute vectors regardless whether those attributes are visually observable in the images. These attributes form different groups and within each group, they are mutually exclusive. For example, female and male form one group, and young, teen, adult, old form another. Some attributes are thus subjective, e.g., no ground-truth age is known and there is no clear definition of what 'young' entails.

Figure 2(a) shows an example where our FS feature can be used to correctly predict all the attributes with SVM classifiers. In this example, although the big hat occludes the face and part of the hair of the person, the colour of the top and the shoe style give away the fact that this a female. A harder example is shown in Figure 2(c). This time the image is a bit blurred and the viewpoint is from the back. However, our FS feature can still predict all the attributes correctly. Our FS feature based prediction makes two mistakes for the person image shown in Figure 2(e). Specifically, the backpack attribute is missed and the lower-body garment colour is predicted to be black rather than blue. Both mistakes are understandable. For the backpack, since the frontal view is shown and the backpack has very thin straps, this attribute can be easily missed even by human (the human annotator labelled this because s/he had access to multiple views of this person including a back view where the backpack is clearly visible). As for the blue vs black for the lower-body cloth, it seems to be a close call even for humans.

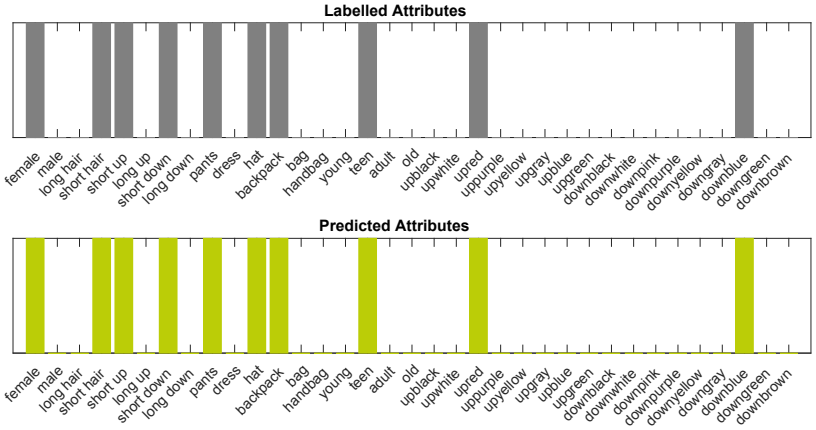
References

[1] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv:1703.07220*, 2017. 1

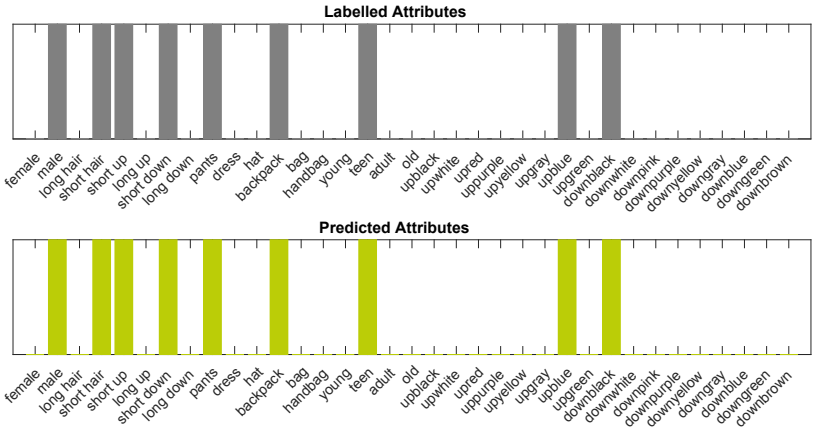
[2] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. *ICCV*, 2017. 1

[3] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CVPR*, 2016. 1

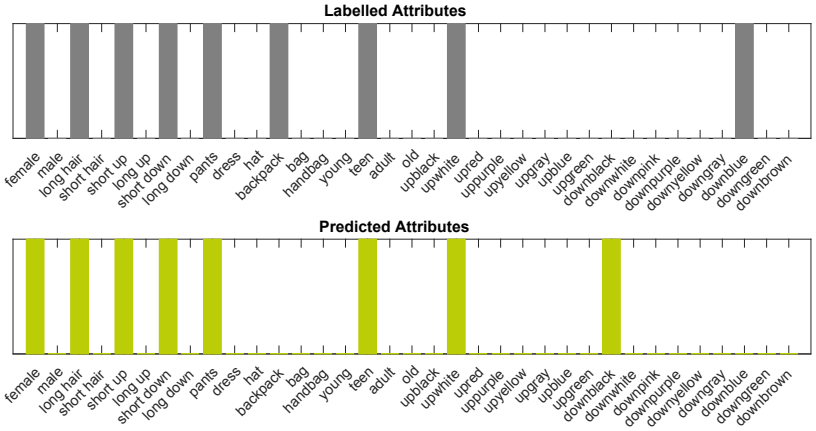
[4] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 1



(a) (b)



(c) (d)



(e) (f)

Figure 2: Examples of attribute prediction using our factor signature (FS) feature. Market-1501 dataset is used. Best viewed in colour.