Visual Person Understanding through Multi-Task and Multi-Dataset Learning Supplementary Material

Kilian Pfeiffer, Alexander Hermans, István Sárándi,
Mark Weber, and Bastian Leibe

Visual Computing Institute, RWTH Aachen University

Abstract. We present several additional qualitative results for both our automatic annotation, as well as results obtained by our final model. Furthermore, we analyze the benefits provided by multi-task and multi-dataset learning over varying person ReID training-set sizes.

1 Automatic Annotations

Figure 1 shows some of our automatic annotations for pose estimation and part segmentation on the Market-1501 dataset. These annotations are predictions from our baseline models for the respective tasks. It can be seen that we achieve decent quality on several images, but failure cases are clearly present too. Please note that these are predictions using the full class set, we train our models on the merged set containing five classes.

2 Additional Qualitative Results

Figure 2 shows additional attribute classification results for binary attributes. Results are shown for both MOT16 sequences, as well as the Market-1501 Dataset.

Figure 3 shows several additional frames from MOT16 sequences [1]. Both pose estimation and part segmentations are shown, as well as the gender prediction visualized by the bounding box color. Please note that both additional attributes are classified and ReID embeddings are generated by the same model, but these are not visualized.

3 Result Video

This PDF is accompanied by videos showing results on several MOT16 sequences, visualized in the same way as Figure 3. Results can likely be improved by a tracking approach to reduce some temporal inconsistencies. Nevertheless, we generate the predictions with a single model, achieving qualitative results comparable to our baseline models and can thus need significantly fewer computational resources.

Fig. 1: Automatic annotations on the Market-1501 dataset using our baseline models for pose estimation (top) and part segmentation (bottom). Annotations quality is not always consistent, both very good (left) and rather bad (right) automatic annotations are created.

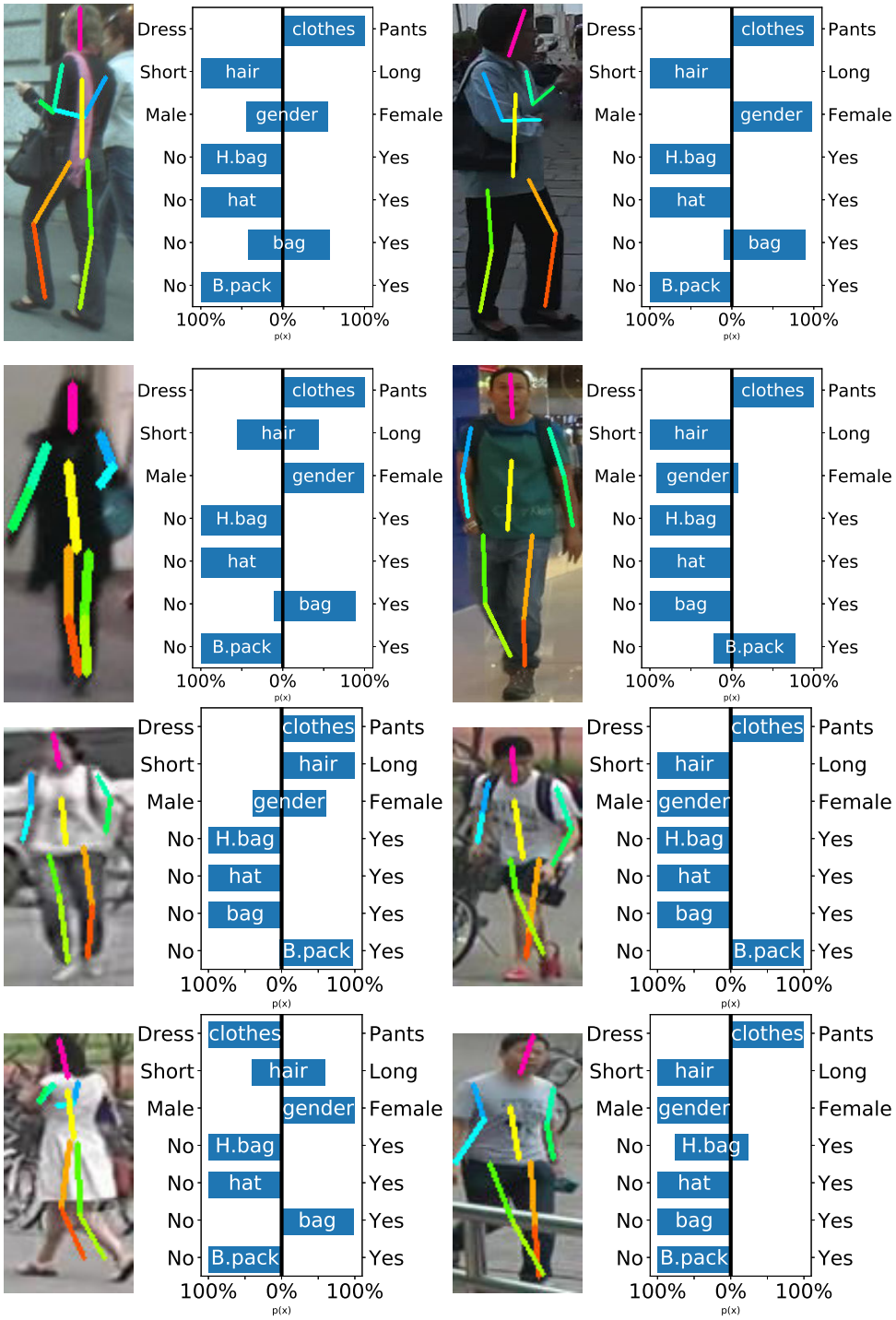

Fig. 2: Given person detections, we perform pose estimation and attribute classification jointly with person re-identification and part segmentation (not visualized) using a shared CNN backbone with small task-specific heads.

Fig. 3: Pose estimation and part segmentation results shown on MOT16 sequences [1]. We use ground truth bounding boxes for this visualization, but detection boxes can be used instead. The bounding box colors correspond to gender predictions (female, male).

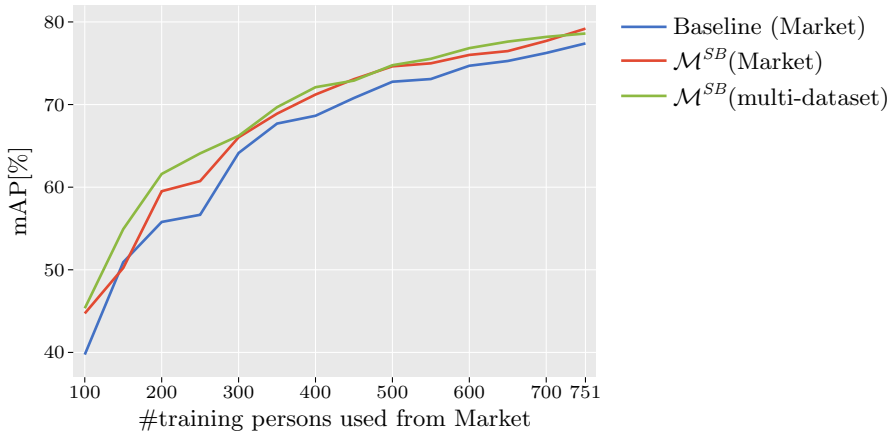

Fig. 4: Learning curves for the ReID performance. MTL (——) consistently outperforms the triplet ReID baseline (——). Joint multi-dataset learning with MPII and LIP (——) gives an additional, albeit smaller, boost.

4 Effect of the Training Set Size

In Figure 4 we analyze the benefits provided by multi-task and multi-dataset learning over the baseline as a function of the amount of ReID training data used. These learning curves show that the improvements hold across a wide range of training set sizes and suggests that the combination of diverse sources of supervision will remain a relevant topic even as available datasets grow in size.

References

1. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A Benchmark for Multi-Object Tracking. arXiv:1603.00831 (2016)