Supplementary Material: Part-Aligned Bilinear Representations for Person Re-Identification

Yumin Suh[1], Jingdong Wang[2], Siyu Tang[3,4], Tao Mei[5], and Kyoung Mu Lee[1]

[1] ASRI, Seoul National University, Seoul, Korea
[2] Microsoft Research Asia, Beijing, China
[3] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[4] University of Tübingen, Tübingen, Germany
[5] JD AI Research, Beijing, China
{n12345,kyoungmu}@snu.ac.kr, jingdw@microsoft.com,
stang@tuebingen.mpg.de, tmei@jd.com

A Details of the Visualization

A.1 Figure 3

In Figure 3 (a) of the main manuscript, we visualize the appearance descriptors by mapping the normalized local appearance descriptors $\tilde{\mathbf{a}}_{xy}$ into 2D space by t-SNE [3]. Similarly, the normalized local part descriptors $\tilde{\mathbf{p}}_{xy}$ are visualized in Figure 3 (b).

A.2 Figure 4 and 5

In Figure 4 and 5, the normalized feature maps are visualized following SIFTFlow [2]. In particular, we project the c_A(or c_P)-dimensional normalized local descriptor vector $\tilde{\mathbf{a}}_{xy}$(or $\tilde{\mathbf{p}}_{xy}$) onto the 3D RGB space, by mapping the top three principal components of descriptor to the RGB.

B Additional Visualization Examples of Feature Maps

We show the additional visualization examples of feature maps on the MARS dataset in Figure 7. For a given input image (left), appearance maps (center) and part maps (right) encode the appearance and body parts, respectively. It shows how the appearance maps differentiate different persons while being invariant for each person. By contrast, the part maps encode the body parts independently from their appearance.

C Additional Analysis

C.1 Effect of non-negative part descriptors

We test one variation of the proposed model, i.e., a model with non-negative part descriptors. In this model, we restrict the part descriptors $\mathbf{p}_{xy}$ to be element-wise non-negative by adding a ReLU layer after the part map extractor $\mathcal{P}$. It makes the local

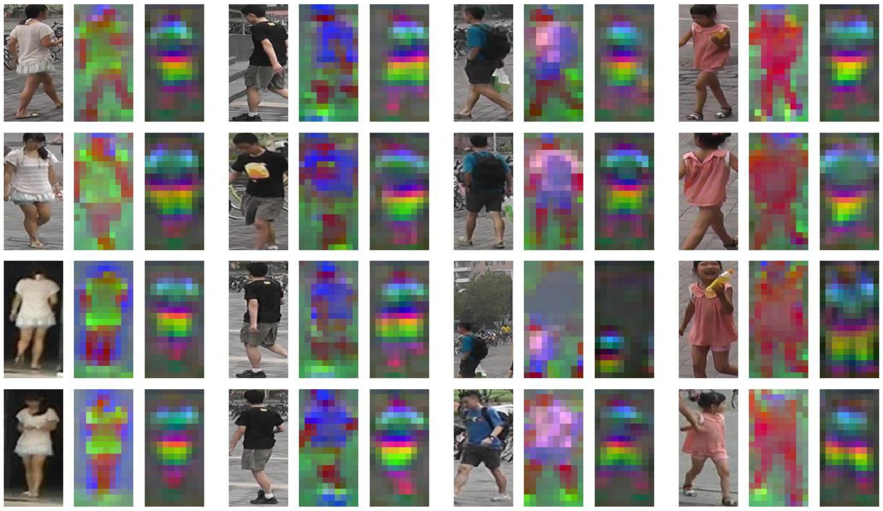

Fig. 7. Visualization of the appearance maps $\mathbf{A}$ and part maps $\mathbf{P}$ obtained from the proposed method on the MARS dataset. For a given input image (left), appearance maps (center) and part maps (right) encode the appearance and body parts, respectively. (Best viewed in color)

part similarity to be always non-negative, and therefore the sign of the local similarity (Eq.9) depends only on the sign of the local appearance similarity. The results on the Market-1501, MARS, and Duke dataset are shown in Table 7. We use the same baseline used in Figure 6 (a,b). Overall, the proposed method and the non-negative variant show similar accuracy in terms of rank@1 accuracy. The non-negative variant shows slightly improved accuracy in terms of mAP.

C.2 Effect of Pose sub-network $\mathcal{P}_{pose}$

Table 8 compares the accuracy when different pose sub-networks $\mathcal{P}_{pose}$ are used. *joint_only, limb_only, internal only*, and *joint_limb_internal* (proposed) denotes a network that generates the joint-based, limb-based, and internal confidence maps with 19, 38, 128, and

Table 7. Accuracy comparison of the baseline, proposed method, and its variation

	Rank	1	5	10	20	mAP
Market-1501	Baseline	81.6	92.0	95.0	96.9	63.6
	Proposed (original)	90.2	96.1	97.4	98.4	76.0
	Proposed (*non-negative*)	89.5	95.6	97.3	98.1	76.1
MARS	Baseline	76.8	89.8	92.3	94.6	63.1
	Proposed (original)	83.0	92.8	95.0	96.8	72.2
	Proposed (*non-negative*)	83.8	94.3	96.1	97.2	74.1
Duke	Baseline	70.6	83.8	87.8	91.2	50.6
	Proposed (original)	82.1	90.2	92.7	95.0	64.2
	Proposed (*non-negative*)	82.0	90.6	93.2	95.2	65.1

Table 8. Accuracy comparison on various pose sub-networks $\mathcal{P}_{pose}$ on Market-1501

Rank	1	5	10	20	mAP
joint_only	88.9	96.0	97.3	98.3	75.6
limb_only	**90.5**	95.9	97.3	98.0	75.5
internal_only	88.2	95.4	97.1	98.1	74.3
joint_limb_internal (proposed)	90.2	**96.1**	**97.4**	**98.4**	**76.0**

Table 9. The joints and limbs used in OpenPose. A limb refers to a connection of two joints.

joints	nose, reye, leye, rear, lear, neck, rsho, lsho, relb, lelb, rwri, lwri, rheap, lheap, rkne, lkne, rank, lank, background
limbs	neck-lsho, neck-rsho, neck-lheap, neck-rheap, lsho-lelb, lelb-lwri, rsho-relb, relb-rwri, lheap-lkne, lkne-lank, rheap-rkne, rkne-rank, nose-neck, nose-leye, leye-lear, nose-reye, reye-rear, lear-lsho, rear-rsho

185 channels, respectively, from OpenPose [1]. It shows that the proposed method achieves similar accuracy for joints and limbs. It implies that the proposed method performs insensitively to the initial pose information given by the pre-trained weights. As the internal feature map provides complementary information to the joints/limbs, we use their concatenation in the final model. The experiments are done without using the dilations.

D Body Joints and Limbs

Our model adopts the subnetwork of the pose estimation network (OpenPose [1] $\mathcal{P}_{pose}$) to form the part map extractor $\mathcal{P}$, i.e., from the image input to the output of the stage2 (*concat_stage3*). It generates a 185-dimensional feature map which consists of 19-dimensional joint confidence map, 38-dimensional limb confidence map, and 128-dimensional internal feature map. The body joints and limbs used are listed in Table 9. For more detailed representation, please refer to [1].

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
2. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE TPAMI **5**(33), 978–994 (2011)
3. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR **9**(Nov), 2579–2605 (2008)