Supplementary materials

Our code is available at `https://github.com/VlSomers/bpbreid` and is based on the **Torchreid**[2] framework. The clean and modular architecture of our framework with SOTA performance will hopefully attract researchers looking for a strong baseline to conduct further research on human-part based ReID. In the next section, we provide further details on the generation of our human parsing labels. We also provide further experiments on the number of body parts defined by the hyper-parameter K, and qualitative assessment for the ranking performance and the attention maps.

Human Parsing Labels Generation with PifPaf

Human parsing labels Y, required for training our part attention module, are generated using the 17 part confidence and 19 part affinity fields produced by the PifPaf [12] pose estimation model. These 36 part confidence and affinity fields are probability maps highlighting different human body region, i.e., 17 human keypoints and 19 joints between these keypoints. For further details about the encoder part of the PifPaf model, we refer readers to [12]. We split these 36 heatmaps into K groups and perform a pixel-wise max operation within each group to obtain K new maps highlighting K body regions. These K maps are then concatenated to produce a tensor $E \in R^{H \times W \times K}$. Each of the K groups correspond to a human semantic region (i.e. body part). These groups are defined manually for a given value of K. Choosing K and defining the right human semantic regions is therefore part of the model hyperparameter tuning process. For instance, with $K = 8$, we define the following semantic regions: {head, left/right arm, torso, left/right leg and left/right feet}. Each element (h, w, c) in E indicates to which degree the spatial location (h, w) belongs to body part c. We perform a final *argmax* operation on E to produce the human parsing label map:

$$Y(h,w) = \begin{cases} 0 & \text{if } \max\limits_{c}(E(h,w,c)) < \lambda_t \\ 1 + \underset{c}{\text{argmax}}(E(h,w,c)) & \text{otherwise ,} \end{cases} \tag{10}$$

where pixels with none of the K channel values above a threshold $\lambda_t = 0.5$ are considered background. An illustration of these coarse human semantic parsing labels is given in Figure 1 for $K = 5$. If multiple persons are detected within a sample, we assume the ReID target is the pedestrian with its head closer to the top center part of the bounding box and remove labels from other persons. We refer readers to our GitHub for more details about the human parsing labels generation strategy.

Instead of using PifPaf, we also tried using some popular human parsing models (**Densepose**[3] and **SCHP**[4]) to generate our human parsing labels, but obtained poor performance because of domain transfer and low image quality in the ReID datasets we target. Human parsing labels obtained with PifPaf gave the best results because it provides consistent predictions with few false negative on a wide range of image resolutions.

In experiment ***"BPBreID without learnable attention"*** from Table 2, the K body part probability maps $\{M_1, ..., M_K\}$ predicted by the body part attention module are replaced by the fixed tensor E described above, on which a channel wise softmax is applied to produce fixed body part classification scores, used as attention weights.

Study on K, the number of body parts

In this Section, we study the impact of the number K of body parts predicted by the body part attention module. The body part attention module is trained using some pre-generated human parsing labels: different labels should therefore be used depending on the value K. The human parsing labels are 2D human semantic segmentation maps, where each pixel is assigned an integer value from 0 to K, 0 being the background label and values from 1 to K being labels for the K body regions. These maps are therefore used to indicate to which body part each pixel of the input image belongs to. Human parsing labels for a few samples are illustrated in Figure 1. In Table 5, we report ranking performance on the Occluded-Duke dataset for various values of K and the corresponding grouping strategy. As demonstrated in this table, best performance is reached with $K = 8$. Other values of K provide too low/high granularity and lead to reduced performance.

Qualitative comparison of ranking performance

We compare ranking performance of our model to other works in Figure 4.

Qualitative comparison of attention maps

We compare the attentions maps of our model to other works in Figure 5.

[2] https://github.com/KaiyangZhou/deep-person-reid

[3] https://github.com/facebookresearch/DensePose

[4] https://github.com/GoGoDuck912/Self-Correction-Human-Parsing

K	R-1	mAP	Grouping strategy for defining human parsing training labels
2	58.3	49.0	{upper body (torso + arms + head), lower body (legs + feet)}
3	63.0	52.0	{head, middle body (torso + arms), lower body (legs + feet)}
4	64.3	52.9	{head, torso, arms, lower body (legs + feet)}
5	65.0	53.3	{head, torso, arms, legs, feet}
6	66.1	52.5	{head, torso, right arm, left arm, legs, feet}
8	**66.7**	**54.1**	{head, torso, right arm, left arm, right leg, left leg, right foot, left foot}
11	66.5	52.9	{head, upper torso, lower torso, upper right arm, lower right arm, upper left arm, lower right arm, right leg, left leg, right foot, left foot}

Table 5. Comparison on Occluded-Duke for different values of K, i.e., the number of body parts embeddings generated by our model, together with the grouping strategy used to generate the corresponding target human parsing labels. These labels are used to train the body part attention module and indicate to which human body region (or background) each pixel in the input image belongs to. The last column details the semantic meaning of each of the K body parts.

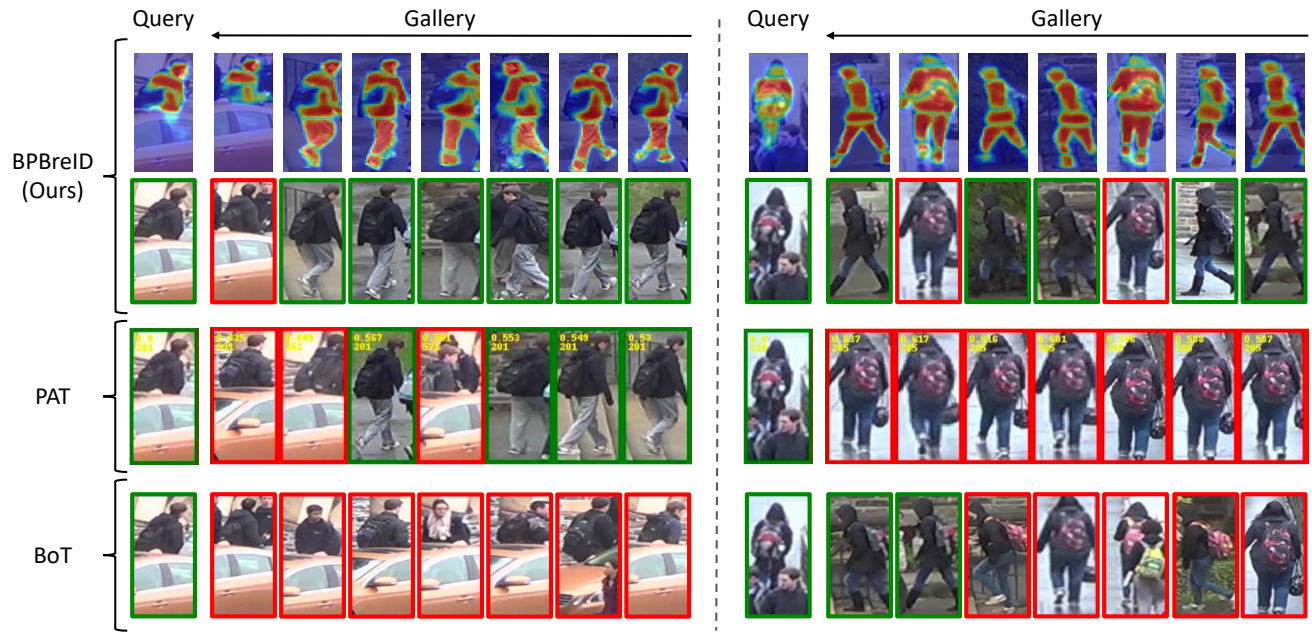

Figure 4. We compare the ranking performance of our model BPBreID with other methods: the part-based transformer method with part discovery PAT [13] and our baseline, the global method BoT [14]. As illustrated in this figure, BoT cannot handle occlusions and PAT is inferior in terms of detecting and aligning fine-grained local appearance features.

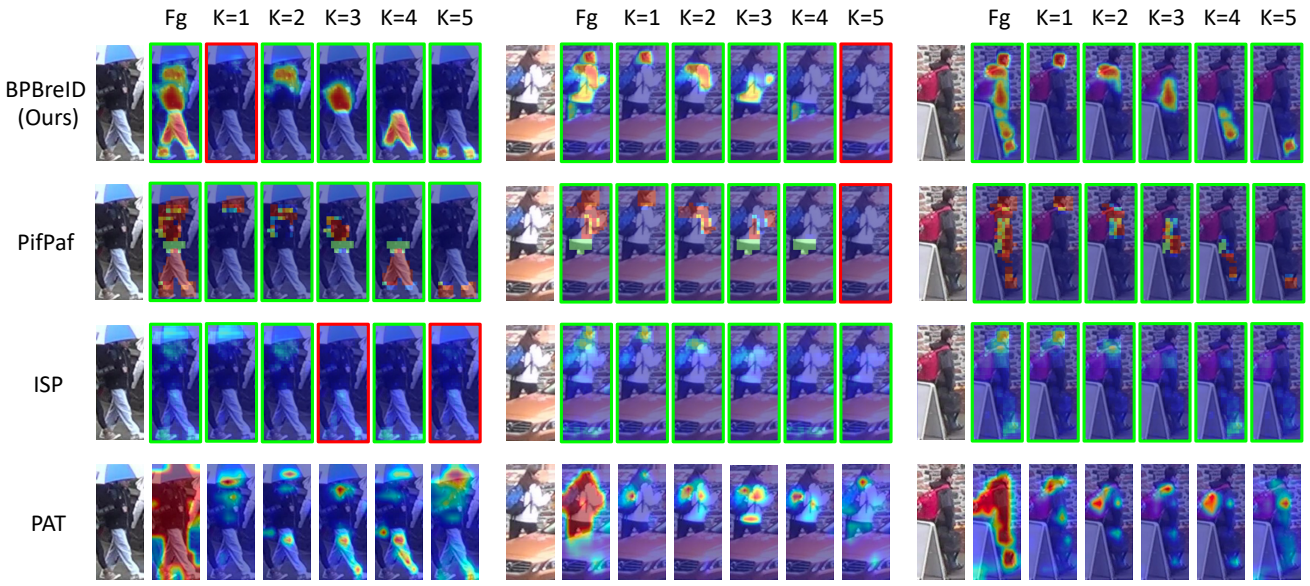

Figure 5. We compare the attentions maps produced by our model BPBreID (on test images unseen at training) with the attention maps from other state-of-the-art part-based methods: ISP [46] and PAT [13]. "Fg" refers to the foreground attention maps, which is obtained by fusing maps from all parts together. Green/red borders illustrate visible/unvisible parts and no color is displayed for PAT because this method is not designed with a visibility score mechanism. Both ISP and PAT use part-discovery to define the human semantic regions, which can lead to missed part, background clutter or feature misalignment. As illustrated in this figure, our attention maps doesn't suffer from these issues. However, unlike these methods, our method only detects body parts and no belongings, such as bags or umbrellas. Moreover, most part-based methods (such as PAT [13], ISP [46], HOReID [5], ...) tries to make each part-based embedding discriminative on its own. This is performed by either incorporating global information into each local embedding [5], or by having each part attending to multiple regions of target person body [13], or by mining discriminative local features [46], as illustrated in this Figure. Different from these methods, we learn part-based embeddings that well represent their associated body-part, without the requirement of being discriminative on their own, but with the requirement of being discriminative when used as a whole. The PifPaf row illustrate the coarse PifPaf part confidence and affinity fields described in the first section of these supplementary materials (tensor E for $K = 5$), from which we derive our human parsing labels used at training.