Part-based Pseudo Label Refinement for Unsupervised Person Re-identification
SUPPLEMENTARY MATERIAL

Yoonki Cho Woo Jae Kim Seunghoon Hong Sung-Eui Yoon
KAIST
{yoonki, wkim97, seunghoon.hong, sungeui}@kaist.ac.kr

In this supplementary material, we provide more experimental results and details of our method. We also provide larger versions of Fig.2 and Fig.5 of the main manuscript.

A. More Experimental Results

A.1. Parameter Analysis

We analyze the impact of the number of parts N_p in our part-based framework, and the results are in Fig. 1. A larger value of N_p reduces the receptive fields of part features so that it contains limited cues for re-identifying a person. Thus, the baseline performance is decreased as the number of parts increases because the part features with smaller receptive fields are trained by hard pseudo-labels that do not consider the context of each part. Nevertheless, PPLR consistently improves the baseline performance with a significant margin throughout different values of N_p. Thanks to the cross agreement score, agreement-aware label smoothing adjusts the label distribution while considering the context of each part, leading to proper part feature learning.

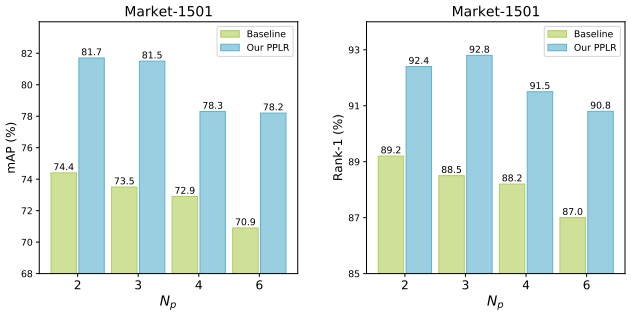

Figure 1. Parameter analysis of the number of parts N_p on Market-1501. 'Baseline' is the part-based unsupervised re-ID framework.

A.2. Training Computational Cost

We compare training computation costs of PPLR with other methods that utilize auxiliary teacher networks to refine pseudo-labels: MEB-Net[1] [4] and MMT[2] [2]. We analyze the number of training parameters, training stage time, and clustering stage time of each method, and the results are in Tab. 1. PPLR only uses features from a single backbone, and it is more efficient than other methods, even including the time to compute the cross agreement score. On the other hand, MMT and MEB-Net use an averaged feature for each sample from multiple backbones for clustering, which requires additional computational cost in the clustering stage. While other methods leverage multiple networks (*e.g.*, dual ResNet in MMT, and single DenseNet, ResNet, and Inception-v3 in MEB-Net), our PPLR is a self-teaching method and requires fewer parameters for training. Furthermore, other methods require multiple feedforwards to refine the pseudo-labels; *e.g.,* MMT feedforwards two current models and two mean-teacher models a total of four times. PPLR only requires a single feedforward for pseudo-label refinements and shows efficiency in the training stage.

Method	Parameters (M)	Clustering stage time (sec / epoch)	Training stage time (sec / iter)
MEB-Net [4]	56.867	120.731	0.998
MMT [2]	50.096	48.545	0.511
Baseline	29.668	36.103	0.194
PPLR	29.668	38.484	0.202

Table 1. Training cost comparison on Market-1501. The clustering stage time includes the time for feature extraction, clustering, and cross agreement score computation. Since the number of iterations per epoch is different for each method, we measure 'sec/iter' for a fair evaluation of training stage time.

A.3. Qualitative Results

To further analyze the cross agreement score, we visualize images that have low- and top-50 cross agreement scores on Market-1501. As shown in Fig. 2, the images with low cross agreement scores contain less discriminative information irrelevant to identifying a person in corresponding part (*e.g.,* occlusions, backgrounds, and presence of multiple people). In contrast, the images with high cross agreement scores are well-aligned with discriminative information. There are also some failure cases to overcome that we leave for future work. Some misaligned parts with

[1] https://github.com/YunpengZhai/MEB-Net
[2] https://github.com/yxgeee/MMT

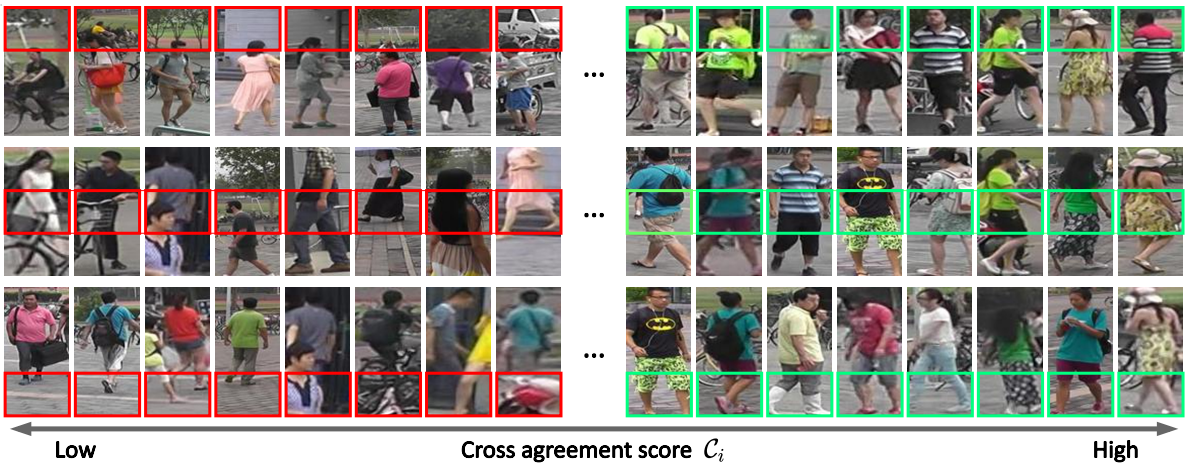

Figure 2. Visualization of images with low-50 and top-50 cross agreement scores on Market-1501. Very similar and duplicated images were excluded to show various cases.

discriminative information have low cross agreement scores because they capture different body features compared to the corresponding parts in the rest of the images. To overcome this limitation, a promising solution would be using auxiliary human semantic information by person attribute recognition or human parsing techniques to construct feature spaces that represent similar semantic parts. To overcome this limitation, a promising solution would be using auxiliary human semantic information by person attribute recognition and human parsing techniques to construct feature spaces that represent similar semantic parts.

B. Camera-aware Proxy Details

When the camera labels are available, PPLR can optionally leverage the inter-camera contrastive loss with camera-aware proxies [3]. Let y_i and c_i respectively denote the pseudo-label and the camera label of the image x_i. We compute the camera-aware proxy $\mathbf{c}_{(a,b)}$, which is the centroid of the features $\mathbf{f}_i$ that have the same camera label a and belong to the same cluster b, defined by:

$$\mathbf{c}_{(a,b)} = \frac{1}{|S_{(a,b)}|} \sum_{i \in S_{(a,b)}} \mathbf{f}_i, \tag{1}$$

where $S_{(a,b)} = \{i | c_i = a \wedge y_i = b\}$ is the index set for the proxy $\mathbf{c}_{(a,b)}$, and $|\cdot|$ is the cardinality of the set.

To compute the inter-camera contrastive loss, the index set $\mathcal{P}_i$ for the positive proxies of the feature $\mathbf{f}_i$ is defined as the proxy indices that have the same pseudo-label y_i but different camera labels with $\mathbf{f}_i$. The index set $\mathcal{Q}_i$ for the hard negative proxies of the feature $\mathbf{f}_i$ is defined as the indices of nearest proxies that have different pseudo-labels to y_i. We utilize the inter-camera contrastive loss with camera-aware proxies on each feature space to reduce the large intra-class variance by disjoint camera views.

C. More Implementation Details

We implement our framework based on PyTorch. Four NVIDIA TITAN RTX GPUs are used for training, and only a single GPU is used for testing. We compute the Jaccard distance based on the k-reciprocal encoding [5] for clutsering, where k is set to 30. For parameters of DBSCAN [1], we set the minimum number of neighbors for a core point to 4 and the distance threshold between samples to 0.7 for MSMT17 and VeRi-776 and 0.6 for Market-1501. With the inter-camera contrastive loss, we use smaller distance threshold between samples, *e.g.*, 0.6 for MSMT17. For stable training, we apply the agreement-aware label smoothing after the first five epochs.

References

[1] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 2

[2] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 1

[3] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, 2021. 2

[4] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *ECCV*, 2020. 1

[5] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 2

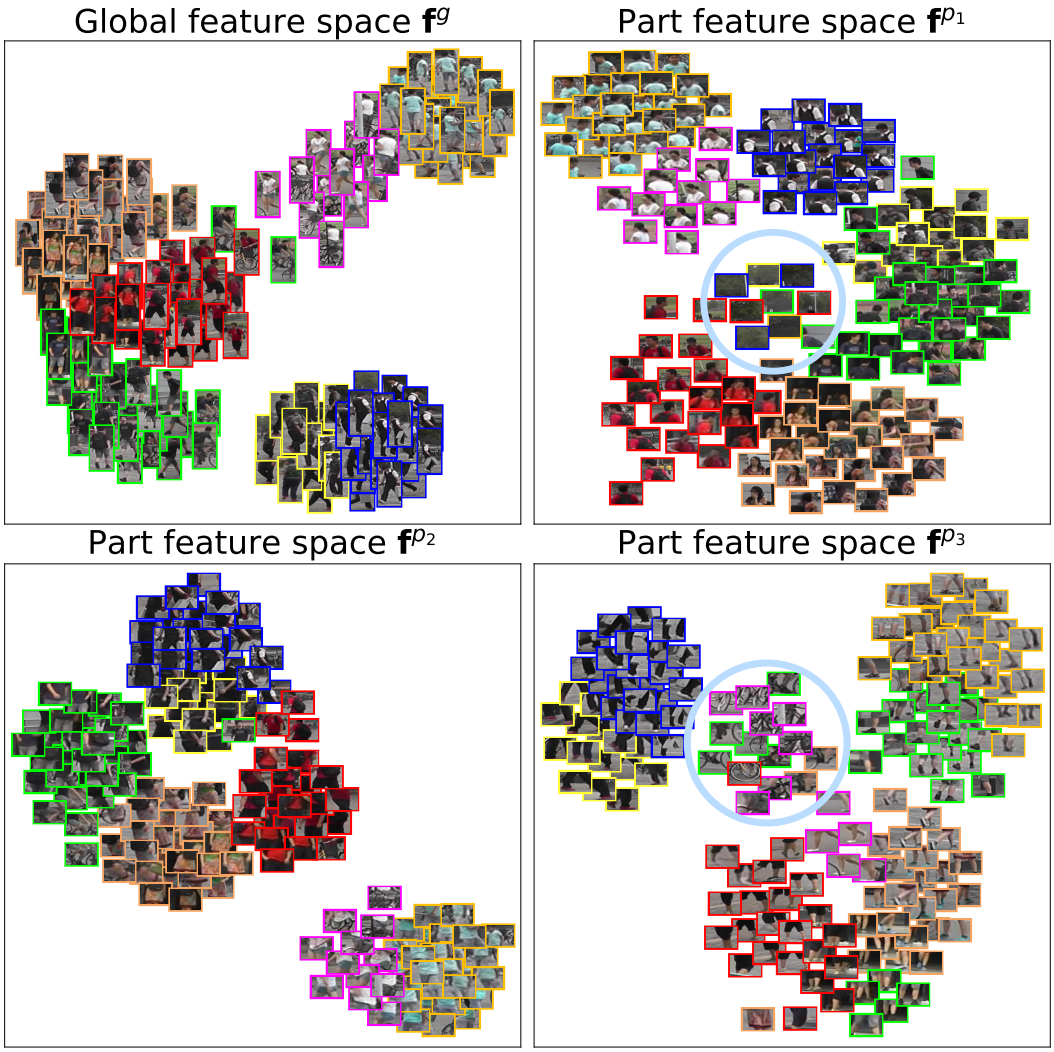

Figure 3. The large version of Fig.2 in the main manuscript.

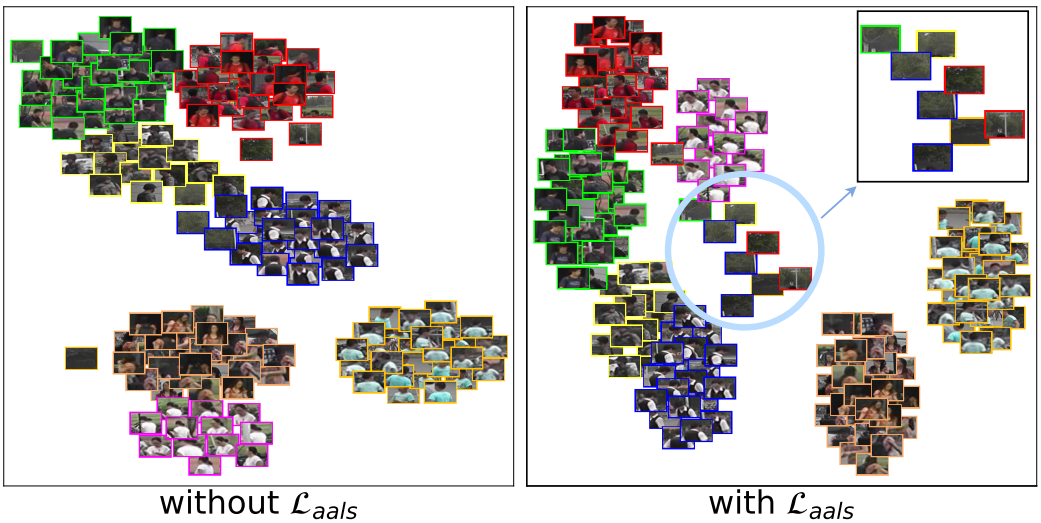

Figure 4. The large version of Fig.5 in the main manuscript.