Supplementary Materials of Online Pseudo Label Generation by Hierarchical Cluster Dynamics for Adaptive Person Re-identification

1. Notation Clarification

To distinguish from the main text, we use S-Fig, S-Tab, and S-Eq to denote figures, tables and equations presented in the supplementary material.

2. Pseudo-code of the proposed algorithm

Algorithm 1 summarizes the pseudo label generation of our framework. Training procedure of our approach on domain adaptive person ReID and unsupervised person ReID are presented in Algorithm 2 and Algorithm 3, respectively. Both tasks adopt the same pseudo label generation strategy. The only difference between the training process of domain adaptive person ReID and unsupervised person ReID is whether $\mathbf{w}_k$ is included in Eq. 1. For adaptive ReID, there are prototypes of the source domain, *i.e.,* $\mathbf{w}_k$, but there are no prototypes of the source domain in unsupervised person ReID.

3. Comparison with states of the art in unsupervised image classification

Recent years have witnessed the great progress in unsupervised learning on image classification [8, 10, 7, 3, 2, 9]. We applied several recent methods, such as BYOL [6], MoCo [7] and ODC [9], on unsupervised person ReID tasks. Specifically, for ODC, we change the number of prediction classes to 500 following MMT [4] and adopt the same sampler strategy used in our approach, *i.e.,* each mini-batch contains 64 target domain images of at least 16 pseudo classes (4 images for each cluster or 1 image for each outlier). Moreover, the backbone for the encoder, data augmentation and network optimization method are replaced by strategies mentioned in implementation details (Sec 5.1). The results presented in S-Table 1 indicate that methods for unsupervised image classification fail to achieve competitive results on unsupervised person ReID tasks. The failure of BYOL and MoCo is due to the fact that they are all based on the paradigm of instance discrimination and no cluster base information is involved. Since person ReID struggles to explore the intra-class and inter-class relations, these instance-based methods fail in unsupervised person

Algorithm 1 Pseudo label generation of the proposed algorithm in one iteration

Require: target domain mini-batch B_t;
Require: a hierarchical label bank $\mathcal{H} = \{y_i^1, y_i^2, ..., y_i^H\}_{i=1}^{N_t}$;
Require: hyper-parameter threshold σ and K label anchors for cluster dynamics;
  for j in B_t **do**
    Find the nearest neighbor p of sample j;
    Update labels $\{y_j^l\}_{l=1}^H \leftarrow \{y_p^l\}_{l=1}^H$;
    for h in [1, H] **do**
      # cluster split
      Choose up to K label anchors in cluster y_j^h;
      Construct the normalized affinity matrix $\mathbf{S}_s$ by Eq. 3;
      Compute the closed-form solution $(\mathbf{P}_s^h)^*$ with $\mathbf{S}_s$ by Eq. 2;
      Split cluster y_j^h by Eq. 4;
      # cluster merge
      Construct the cluster set $\mathcal{O}_i^h$ containing clusters to be merged by Cluster Merge in Sec. 4.2.
      Compute the center feature collection $\mathbf{o}_i^h$ of $\mathcal{O}_i^h$.
      Construct the normalized affinity matrix $\mathbf{S}_m$ by Eq. 3;
      Compute the closed-form solution for cluster merge $(\mathbf{P}_m^h)^* = (p_1^h, p_2^h, ..., p_n^h)$ with $\mathbf{S}_m$ by Eq. 2;
      Merge y_j^h with clusters $\{y_q^h | p_{j,q}^h > \sigma\}$;
    end for
  end for

Algorithm 2 Training procedure of the proposed method on domain adaptive person ReID

Require: Source domain data $\mathcal{D}_s$ and target domain data $\mathcal{D}_t$;
Require: momentum m for updating feature bank $\mathcal{B}$;
  Initialize the backbone encoder f_θ with ImageNet-pretrained ResNet-50;
  Initialize feature bank $\mathcal{B}$ with features extracted by f_θ;
  Initialize the hierarchical label bank $\mathcal{H}$ by DBSCAN;
  for i in [1, $num\_iteration$] **do**
    Get mini-batch $B_s \subset \mathcal{D}_s$ and $B_t \subset \mathcal{D}_t$;
    Encode features F_s, F_t for B_s,B_t with f_θ;
    Compute the contrastive loss with F_s, F_t by Eq. 1;
    Update $\mathcal{B}$ with m in a momentum way as [7];
    Update hierarchical label banks $\mathcal{H}$ for samples in B_t following Algorithm 1;
  end for

Algorithm 3 Training procedure of the proposed method on unsupervised person ReID

Require: Unlabeled data $\mathcal{D}_t$;
Require: momentum m for updating feature bank $\mathcal{B}$;
Initialize the backbone encoder f_θ with ImageNet-pretrained ResNet-50;
Initialize feature bank $\mathcal{B}$ with features extracted by f_θ;
Initialize the hierarchical label bank $\mathcal{H}$ by DBSCAN;
for i in $[1, num\_iteration]$ **do**
Get mini-batch $B_t \subset \mathcal{D}_t$;
Encode features F_t for B_t with f_θ;
Compute the contrastive loss with F_t by Eq. 1;
Update $\mathcal{B}$ with m in a momentum way as [7];
Update hierarchical label banks $\mathcal{H}$ for samples in B_t following Algorithm 1;
end for

S-Table 1. Comparison with state-of-the-art methods in unsupervised images classification on unsupervised person ReID tasks. Implementation of all the methods are based on authors' code.

Methods	Market-1501			DukeMTMC-reID		
	mAP	R1	R5	mAP	R1	R5
BYOL [6]	4.9	11.8	21.8	2.7	5.3	10.3
MoCo [7]	6.1	12.8	27.1	5.6	10.7	22.0
ODC [9]	20.0	38.8	54.9	15.7	24.7	39.1
Ours	78.1	91.1	96.4	65.6	79.8	88.6

ReID. The conclusion is similar to that in [5]. Furthermore, we compare our method with a clustering-based method, ODC [9]. ODC is also an online clustering algorithm with the advantage of a deep clustering framework [1], and therefore it outperforms instance-based methods [6, 7] by approximate 15%. However, the clustering in ODC is actually K-Means algorithms and it heavily relies on a hyperparameter, *i.e.,* the number of clusters. Considering the difficulty of determining the number of people in the ReID dataset and the unchangeable number of clusters in K-Means, ODC achieves much lower performances than our method.

4. More Sensitivity Analysis of α_s, α_m

We explore the influence of α_s, α_m and present the performance in terms of mAP in ablation study (Sec 6.3). In this section, more detailed results, *i.e.,* mAP, rank-1 and rank-5, are shown in S-Table 2 and S-Table 3. The results of mAP and Rank 1 show that the performance of adaptive ReID increases when α_s and α_m increase. According to Eq. 2, larger α_s and α_m indicate considering more neighborhood information in label propagation for cluster split and cluster merge, respectively. Specifically, we find the performance of our proposed method is more sensitive to α_m in cluster merge than α_s in cluster split. The cluster merge often tackles more visually different images than

S-Table 2. Performance comparison with different α_M. D→M denotes adapting DukeMTMC-reid to Market-1501. M→D denotes adapting Market-1501 to DukeMTMC-reid.

α_M	D → M			M → D		
	mAP	R1	R5	mAP	R1	R5
0.1	66.9	84.7	93.1	56.4	72.6	82.2
0.3	68.0	83.8	94.1	58.5	72.8	83.8
0.5	69.2	84.2	94.1	63.3	76.8	87.2
0.7	70.7	84.9	94.6	64.6	78.8	87.9
0.9	78.9	91.1	96.6	69.0	82.6	89.9
0.99	80.0	91.5	96.3	70.1	82.2	89.7

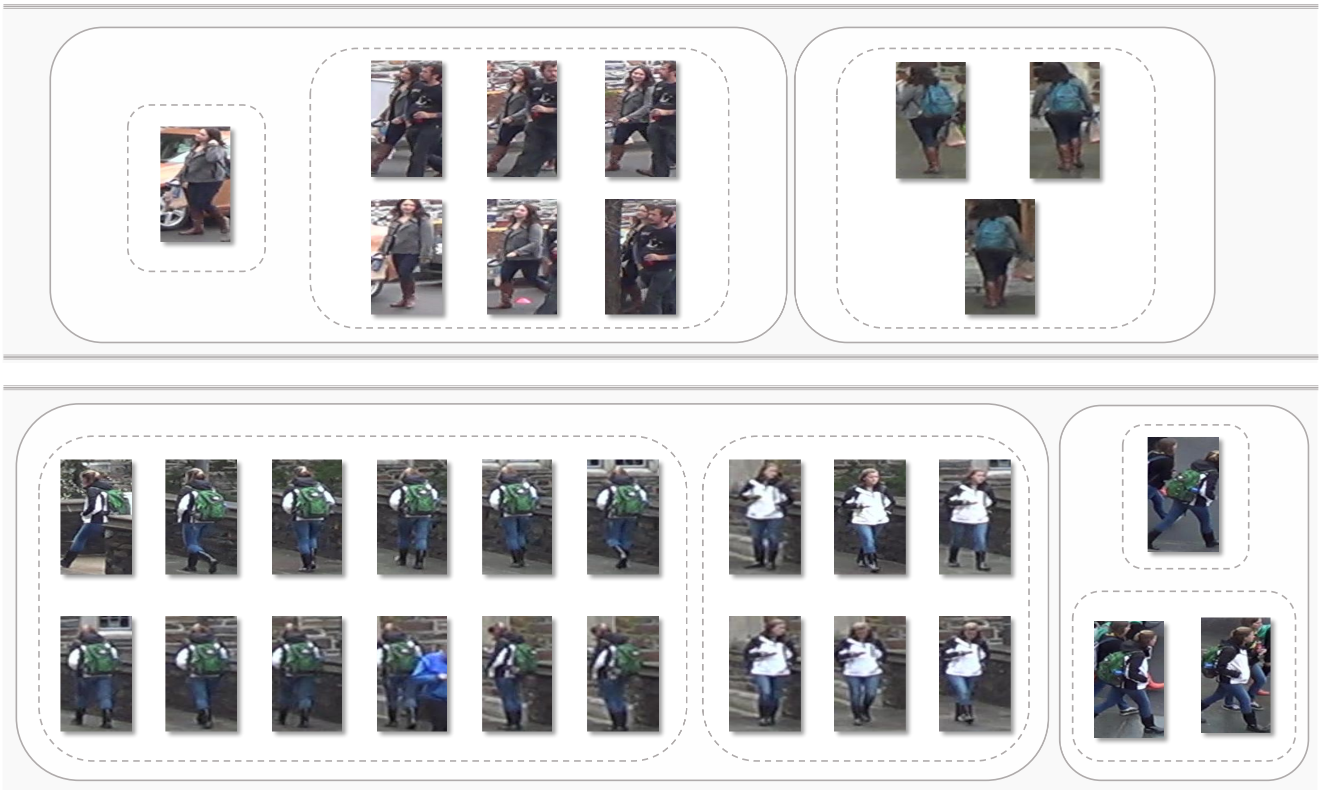
S-Table 3. Performance comparison with different α_S. D→M denotes adapting DukeMTMC-reid to Market-1501. M→D denotes adapting Market-1501 to DukeMTMC-reid.

α_S	D → M			M → D		
	mAP	R1	R5	mAP	R1	R5
0.3	76.1	89.6	95.2	68.8	81.5	90.0
0.5	76.3	89.8	95.4	69.0	82.4	90.4
0.7	76.4	90.2	95.7	69.2	82.1	90.0
0.9	77.1	90.0	96.0	69.2	82.0	89.5
0.95	77.9	90.7	96.0	69.5	82.3	90.4
0.99	80.0	91.5	96.3	70.1	82.2	89.7

cluster split, which makes neighborhood affinities are more important when propagating labels by Eq. 2.

5. Visualization of hierarchical structure of online label generation

In the ablation study (Sec. 6.2), we visualize hierarchical clustering results on DukeMTMC-reID dataset. In this section, more visualization examples are shown in S-Figure 1. We set the total number of levers $H = 3$ in all experiments. S-Figure 1(a) illustrates the hierarchical clustering results in DukeMTMC-reID and S-Figure 1(b) illustrates the hierarchical clustering results in Market-1501. The visualization in S-Figure 1 indicates hierarchical clustering results share similar patterns in both datasets and therefore empirically justifies the generality and effectiveness of our hierarchical online pseudo label generation method. At the first level $h = 1$, images tend to share high similarities within the same cluster, such as the same human posture or the same background. As the level increases to 2, samples with the same background or the same human posture are gathered together. With regard to the highest level $h = 3$, images of the same identity but with different backgrounds and human postures are clustered since they are semantically similar.



(a) Visualization of hierarchical clustering results on DukeMTMC-reID when setting total level $H = 3$.

(b) Visualization of hierarchical clustering results on Market-1501 when setting total level $H = 3$.

h=1 h=2 h=3

S-Figure 1. Visualization examples on DukeMTMC-ReID dataset and Market-1501 dataset. Different types of lines stand for clustering results at different levels.

References

[1] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

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

[5] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11309–11321. Curran Associates, Inc., 2020. 2

[6] J. B. Grill, F. Strub, F Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, and M. G. Azar. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1, 2

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 1, 2

[8] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 1

[9] X. Zhan, J. Xie, Z. Liu, Y. S. Ong, and C. C. Loy. Online deep clustering for unsupervised representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6687–6696, 2020. 1, 2

[10] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 1