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 \(w_k\) is included in Eq. 1. For adaptive ReID, there are prototypes of the source domain, i.e., \(w_k\), 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 3 Training procedure of the proposed method on unsupervised person ReID

Require: Unlabeled data \( \mathcal{D}_i \);

Require: momentum \( m \) for updating feature bank \( \mathcal{B} \);

1. Initialize the backbone encoder \( f_o \) with ImageNet-pretrained ResNet-50;
2. Initialize feature bank \( \mathcal{B} \) with features extracted by \( f_o \);  
3. Initialize the hierarchical label bank \( \mathcal{H} \) by DBSCAN;

for \( i \) in [1, num_iteration] do

1. Get mini-batch \( \mathcal{B}_t \subset \mathcal{D}_i \);
2. Encode features \( \mathcal{F}_t \) for \( \mathcal{B}_t \) with \( f_o \);
3. Compute the contrastive loss with \( \mathcal{F}_t \) by Eq. 1;
4. Update \( \mathcal{B} \) with \( m \) in a momentum way as [7];
5. Update hierarchical label banks \( \mathcal{H} \) for samples in \( \mathcal{B}_t \) 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Market-1501</th>
<th>DukeMTMC-reID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
<td>R1</td>
</tr>
<tr>
<td>BYOL [6]</td>
<td>4.9</td>
<td>11.8</td>
</tr>
<tr>
<td>MoCo [7]</td>
<td>6.1</td>
<td>12.8</td>
</tr>
<tr>
<td>ODC [9]</td>
<td>20.0</td>
<td>38.8</td>
</tr>
<tr>
<td>Ours</td>
<td>78.1</td>
<td>91.1</td>
</tr>
</tbody>
</table>

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, \( \alpha \), 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 \( \alpha_s, \alpha_m \)

We explore the influence of \( \alpha_s, \alpha_m \) and present the performance in terms of mAP in ablation study (Sec 6.3). In this section, more detailed results, \( \alpha_s \), 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 \( \alpha_s \) and \( \alpha_m \) increase. According to Eq. 2, larger \( \alpha_s \) and \( \alpha_m \) 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 \( \alpha_m \) in cluster merge than \( \alpha_s \) in cluster split. The cluster merge often tackles more visually different images than 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 $H = 3$.

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


